Direct estimation of the index coefficients in a single-index model

Hristache, Marian

ENSAI, Campus Ker Lann, rue B. Pascal, 35170 Bruz, France

hristach@ensai.fr

Juditski, Anatoli*

INRIA Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot, France

anatoli.iouditski@inrialpes.fr

Spokoiny, Vladimir

Weierstrass-Institute, Mohrenstr. 39, 10117 Berlin, Germany

spokoiny@wias-berlin.de

May 29, 1998

Running title: Direct index estimation.

Keywords: single-index model, index coefficients, link function, direct estimation, iterations.

AMS 1995 Subject Classification. Primary 62G05; Secondary 62H40, 62G20.

*This is the author to whom the correspondence should be sent

Abstract

We propose a new method of estimating the index coefficients in a single index model which is based on iterative improvement of the average derivative estimator. The resulting estimate is $\sqrt{n}$–consistent under mild assumptions on the model.

1 Introduction

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

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

where Y_i a scalar response variables, $X_i \in [-1, 1]^d$ 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 specific structure:

$$f(x) = g(x^T \theta^*). \tag{1.2}$$

Here $g(\cdot)$ is an unknown 1-dimensional *link* function, e.g. $g(\cdot) : \mathbb{R} \to \mathbb{R}$ and θ is an unknown *index* vector. In the statistical literature the relations as in (1.1) and (1.2) are referred to as the *single-index regression* models. These models are often used in econometrics as a reasonable compromise between fully parametric and fully nonparametric modelling (see e.g. McCullagh and Nelder, 1989). For instance, they are extensively used in projection pursuit regression (cf. Friedman and Stuetzle, 1981 and Hall, 1989).

Two estimation problems for single-index models are intensively discussed in the literature. The first is to estimate the unknown function $f(x)$, the second is to recover the index-vector θ^*. In this paper we focus on the second one. A variety of methods to estimate 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 of θ^* is constructed by minimization of an M-functional with respect to θ^*, when replacing g by its nonparametric estimator. Typical examples are semiparametric maximum likelihood estimator (SMLE) and semiparametric least squares estimator (SLSE). 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. Then the problem of the choice of bandwidth for the nonparametric estimation of the link function has been considered in Härdle et al., 1993. In Delecroix and Hristache, 1998 a rather general type of M-estimators has been studied, and the asymptotic efficiency of the general semiparametric maximum-likelihood estimator has been proved in Bonneu et al., 1997 and Delecroix et al., 1997 for particular classes of single-index models.

In spite of nice theoretical properties of M-estimators they are rarely implemented in practice. The reason for this is twofold. First of all, the above mentioned results are valid only under quite restrictive model assumptions. In particular discrete regressors are not allowed. However, it is the second reason which is crucial: the computation of these estimators leads to a general 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 et al., 1989. The idea of this method is to estimate the expected value of the (weighted) gradient of the regression function which is obviously proportional to θ^*. This method leads to a $\sqrt{n}$-consistent estimator of the index vector (cf. also Härdle and Tsybakov, 1993). An advantage of this approach is that it allows to estimate the vector θ^* "directly" and does not require to solve a hard optimization problem. Unfortunately, the conditions required for this method to work are rather restrictive. For instance, the regressors X must possess a smooth density. A generalization of the ADE for the case where the components of X_i are continuous and/or discrete has been provided in Härdle and Horowitz, 1997. Meanwhile, an estimator for the subvector of θ^* which corresponds to the continuous part of X is prerequisite.

Another direct method of the index coefficient estimation has been proposed in Li and Duan, 1989, where the single-index model of the form

$$Y_i = g(\alpha + X_i\theta, \varepsilon_i)$$

is concerned (here ε_i is supposed to be independent of X_i). A major inconvenience of this approach is that it can be applied only if the regressors X_i have an elliptic distribution.

In the present paper we introduce a new type of direct estimate of the index coefficient θ^*. It can be regarded as an iterative improvement of the average derivative estimator. The underlying idea (as for the ADE method) is that the gradient of the function $f(x) = g(x^T\theta^*)$ is proportional to θ^*. We show that the proposed estimator is $\sqrt{n}$-consistent. The results are valid under rather mild conditions on the design (X_i), $i = 1, ..., n$. 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 can note that a $\sqrt{n}$-estimator can be used as a departure point for the so called "one-step efficient estimator" as discussed, e.g. in Delecroix et al., 1997.

The paper is organized as follows. In the next section we describe the estimation algorithm. Next the properties of the proposed algorithm are studied in Section 3. In Section 4 we consider details of implementation of the proposed estimate and present some simulation results. The proof are gathered in Section 5.1.

2 Algorithm description

We start with the informal description of the proposed estimate.

2.1 The idea

Let us suppose for a moment that $d = 2$ and the observations X_i are scattered uniformly over the square $[0,1]^2$.

The idea of the construction is as follows. Assume that we are interested in estimating ∇f at one of the points X_i and that we know in advance that ∇f is Lipschitz continuous at X_i. Then the natural way to estimate $\beta_i = \nabla f(X_i)$ is to use a "local" least-squares estimate

$$\widehat{\beta}_i = \arg\min_{\beta} \sum_{j=1}^{n} [Y_j - Y_i - \beta^T (X_j - X_i)]^2 K\left(\frac{|X_j - X_i|^2}{h^2}\right), \tag{2.1}$$

where *the 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. Then the averaged gradient

$$\beta^* = \frac{1}{n}\sum_{i=1}^{n} \nabla f(X_i)$$

can be estimated with

$$\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta}_i.$$

Recall that

$$\beta^* = \frac{1}{n}\sum_{i=1}^{n} \nabla f(X_i) = \theta^* \left(\frac{1}{n}\sum_{i=1}^{n} g'(X_i^T \theta^*)\right).$$

This implies that if the quantity

$$\frac{1}{n}\sum_{i=1}^{n} g'(X_i^T \theta^*)$$

is separated away from zero, one can construct an estimate $\widehat{\theta}$ of θ^* as

$$\widehat{\theta} = \frac{\widehat{\beta}}{|\widehat{\beta}|}. \tag{2.2}$$

Note that one can obtain the following upper bound for the error of the estimate $\widehat{\beta}$:

$$|\widehat{\beta} - \beta^*| \leq C_1 h + C_2 \frac{|\xi|}{\sqrt{n}h}, \tag{2.3}$$

where ξ is a normal Gaussian random variable with zero mean. The right hand side of (2.3) is comprised of two terms. The first term is the deterministic error (the bias), which is due to the error in the local approximation of f by a linear function. This error is proportional to h. The second term is the stochastic error $C_2\frac{\xi}{\sqrt{n}h}$ which is independent of f, this term is typically of order $(\sqrt{n}h)^{-1}$. The balance of the two terms gives $h \sim n^{-1/4}$ and the error

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

and

$$|\widehat{\theta} - \theta^*| = O(n^{-1/4}),$$

for the estimate $\widehat{\theta}$ in (2.2).

This rate of convergence ($n^{-1/4}$) is, of course, much worse than $n^{-1/2}$ that can be attained for this problem. However, the simple estimate (2.2) can be significantly improved. First we note that the bias of the estimate (the first term in the right hand side of (2.3)) is in fact proportional to the width of the projection of the spheric window $U_h(X_i)$, defined by the kernel $K(\cdot/h)$ on the direction θ^*. On the other hand, the function f is constant in the direction orthogonal to θ^*, so we can stretch the window $U_h(X_i)$ along this direction without increasing the bias term. Though the true index θ^* is not known, we have already a rather good estimate of it due to (2.2). Now we proceed as follows: at any X_i we define an elliptic window $U_{h,\rho}$ centered at X_i with the small axis of size $O(h\rho)$, oriented along $\widehat{\theta}$, and the large axis of size $O(h)$ orthogonal to $\widehat{\theta}$. We can expect that if ρ is small enough and $\widehat{\theta}$ is close to θ^* then the error of approximation of $f(x) = g(x^T\theta^*)$ by a linear function in the neighborhood $U_{h,\rho}$ of X_i would be rather small. In order to define such an elliptic window we substitute the weights $K(h^{-2}|X_j - X_i|^2)$ in (2.1) with $K(h^{-2}|\Lambda_{\rho,\theta}(X_j - X_i)|^2)$, where the positive symmetric matrix

$$\Lambda_{\rho,\widehat{\theta}} = I + \rho^{-1}\widehat{\theta}\widehat{\theta}^T$$

defines the "elliptic" geometry of the window. (Here I denotes the unit $d \times d$-matrix.) Then we continue as above: we compute the estimates

$$\widehat{\beta}_i^{(1)} = \arg\min_{\beta} \sum_{j=1}^{n} [Y_j - Y_i - \beta^T(X_j - X_i)]^2 K\left(\frac{|\Lambda_{\rho,\widehat{\theta}}(X_j - X_i)|^2}{h^2}\right), \tag{2.4}$$

and their average

$$\widehat{\beta}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta}_i^{(1)}.$$

Finally, we come up with the estimate

$$\widehat{\theta}^{(1)} = \frac{\widehat{\beta}^{(1)}}{|\widehat{\beta}^{(1)}|}.$$

After some tedious computations we obtain that if for some $\gamma > 0$, $|\widehat{\theta} - \theta^*| \le \gamma$ and $\rho \ge \gamma$, the estimate $\widehat{\theta}^{(1)}$ satisfies

$$|\widehat{\theta}^{(1)} - \theta^*| \sim C_3 h\rho^2 + C_4 \frac{|\xi|}{\sqrt{nh}}.$$

Since $\gamma = O(n^{-1/4})$, the choice $h = O(1)$ and $\rho = n^{-1/4}$ gives

$$|\widehat{\theta}^{(1)} - \theta^*| = O(n^{-1/2}),$$

so that the estimate $\widehat{\theta}^{(1)}$ is $\sqrt{n}$-consistent.

This simple method of improvement of the simple estimate (2.2) constitutes the basis of the algorithm described below. However, there two problems which should be addressed:

1. when the model dimension $d > 4$ one cannot take the "optimal" initial window $h = O(n^{-1/4})$ in (2.1) which balances the terms in the right-hand side of (2.3). The reason for it is that there will not be enough (i.e. $d+1$) observations points in the neighborhood $U_h(X_i)$ needed to compute a d-dimensional vector $\widehat{\beta}_i$. One has to take $h = O(n^{1/d})$ in this case. Therefore, for $d > 4$, a $\sqrt{n}$-consistent estimate of θ^* cannot be obtained as a result of a single iteration, the iteration is to be repeated several times in order to attain the rate of convergence $n^{-1/2}$.

2. There is also another reason to make several iterations even in the case when $d \le 4$: the bias term of the estimation error rapidly becomes negligible with respect to the main stochastic term. On the other hand, the stochastic term does not degrade noticeably during the iterations. Therefore, one can flatten the window slowly, e.g. by the factor of 2, in the direction of the concurrent estimate $\widehat{\theta}$ (and stretch it slowly in the orthogonal subspace). This way we obtain the algorithm which possesses good asymptotic properties and is quite robust at the same time.

2.2 Estimation procedure

Let $K : I\!R \to I\!R$ be a function which is positive on $[0,1)$ and vanishes elsewhere. We consider the following

Algorithm 1.

1 `Initialization:` `set` $k = 0$,

$$h_k = C_0 \left(\frac{\ln n}{n}\right)^{-\frac{1}{4}\wedge\frac{1}{d}}, \quad \Lambda_k = I, \quad \rho_k = 1. \tag{2.5}$$

% Iteration description:

```
While $\rho_k > (\ln n/n)^{1/3}$

    2 Compute the local to $X_i$ solution $\widehat{\beta}_k(X_i)$ of the least-squares
      problem (2.4)

      $\widehat{\beta}_k(X_i) = V_{h_k,\Lambda_k}^{-1}(X_i) \sum_{j=1}^{n} (Y_j - Y_i)(X_j - X_i) K\left( \frac{|\Lambda_k(X_j - X_i)|^2}{h_k^2} \right)$

      with

      $V_{h_k,\Lambda_k}(x) = \sum_{j=1}^{n} (X_j - x)(X_j - x)^T K\left( \frac{|\Lambda_k(X_j - X_i)|^2}{h_k^2} \right);$

    3 Compute the mean $\widehat{\beta}_k$ of $\widehat{\beta}_k(X_i)$

      $\widehat{\beta}_k = \frac{1}{n} \sum_{i=1}^{n} \widehat{\beta}_k(X_i).$        (2.6)

      Set $\widehat{\theta}_k = \widehat{\beta}_k / |\widehat{\beta}_k|$.

    4 Set $k = k + 1$, $h_k = 2h_{k-1} \wedge 1$, $\rho_k = \rho_{k-1}/2$, $\Lambda_k = I + \rho^{-1}\widehat{\theta}_k\widehat{\theta}_k^T$.
      Continue with Step 2;

End While

5 Set $\widehat{\theta} = \widehat{\theta}_k$. Terminate;
```

3 Main result

We consider the following assumptions:

Assumption 1. The kernel function $K(\cdot)$ satisfies the following conditions:

1. K is nonnegative and bounded by 1, i.e. $0 \le K(x) \le 1$;
2. K is positive on $[0, 1)$ and vanishes outside, i.e. $K(x) > 0$ for $0 \le x < 1$ and $K(x) = 0$ for all $|x| \ge 1$;
3. K is continuously differentiable on $[0, 1]$.

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

Assumption 3. The function g is two times differentiable with a bounded second derivative,

$$|g''(u)| \le C_g \ \text{ for all } \ u \in I\!R;$$

We put

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

where $\nabla f(x) = g'(x^T\theta^*)\theta^*$ is the gradient of the regression function $f(x) = g(x^T\theta^*)$. Obviously, β^* is proportional to θ^*. We have the following identifiability

Assumption 4. The value

$$|\beta^*| = \frac{1}{n}\sum_{i=1}^{n} g'(X_i^T\theta^*).$$

is separated away from zero, i.e. $|\beta^*| \ge G_0 > 0$ for all n large enough.

In order Algorithm 1 to work, we have to suppose that the design points (X_i) are “well diffused” and, as a consequence, all the matrices $V_{h_k,\Lambda_k}(X_i)$ are well defined. Given a $d\times d$-matrix Λ, we define the normalization $\overline{V}_{h,\Lambda}(x)$ of $V_{h,\Lambda}(x)$, as follows: let

$$N_{h,\Lambda}(x) = \sum_{j=1}^{n} K\left(\frac{|\Lambda(X_j - x)|^2}{h^2}\right),$$

then

$$\begin{aligned}\overline{V}_{h,\Lambda}(x) &= \frac{1}{N_{h,\Lambda}(x)h^2}\Lambda\, V_{h,\Lambda}(x)\,\Lambda\\ &= \frac{1}{N_{h,\Lambda}(x)}\sum_{j=1}^{n}\left(\frac{\Lambda(X_j - x)}{h}\right)\left(\frac{\Lambda(X_j - x)}{h}\right)^T K\left(\frac{|\Lambda(X_j - x)|^2}{h^2}\right).\end{aligned}$$

Obviously $\overline{V}_{h,\Lambda}(x)^{-1} = N_{h,\Lambda}\, h^2\, \Pi\, V_{h,\Lambda}^{-1}\, \Pi$ with $\Pi = \Lambda^{-1}$.

Assumption 5. There exist constants C_V, C_N and C_w such that for all values h_k and Λ_k involved and for every X_i,

1. the inverse matrices $\overline{V}_{h_k,\Lambda_k}(X_i)^{-1}$ are well defined and uniformly bounded i.e.

$$\left|\overline{V}_{h_k,\Lambda_k}(X_i)^{-1}\right| = h_k^2\left|\Pi_k V_{h_k,\Lambda_k}^{-1}(X_i)\Pi_k\right|\sum_{j=1}^{n} K\left(\frac{|\Lambda_k(X_j - x)|^2}{h^2}\right) \le C_V$$

(here $\Pi_k = \Lambda_k^{-1}$ and $|A|$ stands for the Euclidean norm of A);

2. $$\sum_{j=1}^{n}\frac{K\left(\frac{|\Lambda_k(X_j - X_i)|^2}{h_k^2}\right)}{\sum_{\ell=1}^{n} K\left(\frac{|\Lambda_k(X_\ell - X_j)|^2}{h_k^2}\right)} \le C_N;$$

3. $$\frac{\sum_{j=1}^{n} \left| K'\left(\frac{|\Lambda_k(X_j - X_i)|^2}{h_k^2}\right) \right|}{\sum_{j=1}^{n} K\left(\frac{|\Lambda_k(X_j - X_i)|^2}{h_k^2}\right)} \le C_w.$$

Here K' means the derivative of the kernel K.

We can now state the main result of the paper.

Theorem 1 *Let assumptions 2–5 hold. Then there is a constant κ such that for every $z \ge 1$ and n large enough*

$$\boldsymbol{P}\left(|\widehat{\theta} - \theta^*| \ge \frac{C_1}{\sqrt{n}} z + \kappa \left(\frac{2\log n}{n} \right)^{2/3} \right) \le \exp(-z^2/2) + \frac{3\log n}{n}.$$

where

$$C_1 = \frac{C_V(1 + C_N)\sigma}{|\beta^*|}, \qquad (3.1)$$

3.1 Comments

In Algorithm 1 and assumptions above we have not considered the effects which may occur at the boundary of the cube $[-1, 1]$. Indeed, for certain values of θ and the points X_i which are close to a vertex of the cube the size of the "effective window", i.e. the diameter of the intersection of the set $U_{h,\rho,\theta}(X_i) = \{x : |\Lambda_{\rho,\theta}(x - X_i)| \le h\}$ with the cube can be much less than h. Clearly, for such points X_i Assumption 5 does not hold. One of the solutions to this problem could be to include in the expression (2.6) for $\widehat{\beta}$ only the estimates $\widehat{\beta}(X_i)$ which were computed over "effective windows" of diameter larger than, say, $\frac{h}{4}$. Of course, in order such an estimate to work the identifiability Assumption 4 should hold for the restricted set of design points. We do not consider this situation rigorously here.

In Assumption 5 we assess certain properties of the design (X_i), $i = 1, ..., n$. For instance, 5 does not hold, at least for $d > 4$, in the case when X_i form a regular grid, θ coincides with one of the axes of the grid and $\rho < n^{-1/d}$. In such a case, the set $\{X_j : |X_j - X_i|_{\Lambda_{\rho,\theta}} \le 1\}$ is a grid hyperplane of dimension $d - 1$ and the matrix $\overline{V}_{h,\rho,\theta}(x)$ is degenerate. On the other hand, one can verify that Assumptions 4 – 5 hold true in a rather general situation of random design.

The values of the constants C_V, C_N which define the rate of convergence in Theorem 1 depend heavily on the design (X_i) and on the particular kernel K. Note that if the kernel K satisfies Assumption 1, then a simple bound for C_N can be easily obtained (cf. Lemma 1 in Section 5). However, this bound is rather pessimistic and maybe significantly improved for a particular design (for instance, when (X_i) are uniformly distributed, C_N is close to 1 with overwhelming probability).

By inspecting the proof of the theorem 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 $\sup_{1 \le i \le n} \boldsymbol{E}\varepsilon_i^2$.

Similarly, the results apply for non-Gaussian errors under the condition

$$\sup_{1 \le i \le n} \boldsymbol{E} \exp(\lambda \varepsilon_i) \le \kappa_\lambda$$

fore some positive constants λ and κ_λ. Of course, in this situation the constant C_1 in (3.1) is to be modified.

One natural question that arises when Theorem 1 is concerned is what happens if this model assumption is inadequate, i.e. if the regression function $f(x)$ does not possess a single-index structure. It is known that the average derivative method gives a $\sqrt{n}$-consistent estimate 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 et al., 1989) A similar result holds for the first step estimate $\widehat{\theta}_0$, however, now the rate of convergence is $n^{-1/4}$ for $d \le 4$ and $n^{-1/d}$ for $d > 4$. Unfortunately, if the model structure does not correspond to (1.1) further iterations do not lead to the improvement of this initial estimate but may even deteriorate the accuracy of estimation. The reason is that the choice of the specific form of nonparametric neighborhood allows to reduce the bias of the estimate $\widehat{\theta}$ only for the special structure (1.1) of the regression function. Therefore any application of the proposed procedure should be combined with a careful justification of the model assumption.

4 Implementation and simulation results

In order to implement Algorithm 1 one has to choose the constant C_0 in the definition (2.5) of the initial bandwidth h_0. This should be done to guarantee the matrices $\overline{V}_{h_0,I}(X_i)$ to be non-degenerate for $i = 1, ..., n$ (cf. Assumption 5). Moreover, one can include in the sum in the expression (2.6) only those i's for which the estimate $\widehat{\beta}_k(X_i)$ is well defined (i.e. the matrix $\overline{V}_{h_0,I}^{-1}(X_i)$ is well conditioned). Clearly, the bandwidth h_0 is to be selected in such a way that the total number of such terms is $O(n)$.

4.1 Modified algorithm

Another method to ensure that the matrices $V_{h_k,\Lambda_k}(X_i)$ are well conditioned is to select at each iteration the bandwidth $h_k(X_i)$ which is proper to each point X_i. Let $N_{h,\Lambda}(x)$ stand for the cardinality of the set

$$B = \{X_i : \ |\Lambda(X_i - x)| \le h\}.$$

One can choose, for instance, $h_k(X_i)$ such that $\lambda_{min}(\overline{V}_{h_k(X_i),\Lambda_k}(X_i)) \ge \lambda_0 > 0$ or such that $N_{h_k(X_i),\Lambda_k} \ge n_0$ (typically, $n_0 > d+1$ would give a non-degenerate matrix $\overline{V}_{h_k(X_i),\Lambda_k}(X_i)$). We realize this idea in the following

Algorithm 2.

```
1 Initialization:  Define the set $\mathcal{H}$ of admissible bandwidths as follows:  set $h_0 = n^{-1/4\vee 1/d}$,
      $$\mathcal{H} = \{h_i = h_0 2^{i/d},\ i = 0, ..., -[d \log_2 h_0] + 1\}$$
  % (here $[\cdot]$ stands for the integer part).
  Put
      $$k = 0,\ \ \rho_k = 1,\ \ \Lambda_k = I,\ \ N_k = 2d.$$
  % Iteration description:
  While $\rho_k > (\ln n/n)^{1/3}$
      2 For each $X_i$ select $h_k(X_i)$ as follows
            $$h(X_i) = \min\{h \in \mathcal{H} :\ N_{h,\Lambda_k}(X_i) \geq N_k\}.$$
      3 Compute the local to $X_i$ solution $\widehat{\beta}_k(X_i)$ of the least-squares problem  (2.4)
            $$\widehat{\beta}_k(X_i) = V^{-1}_{h(X_i),\Lambda_k}(X_i) \sum_{j=1}^{n} (Y_j - Y_i)(X_j - X_i) K \left( \frac{|\Lambda_k (X_j - X_i)|^2}{h^2(X_i)} \right)$$
        with
            $$V_{h(X_i),\Lambda_k}(x) = \sum_{j=1}^{n} (X_j - x)(X_j - x)^T K \left( \frac{|\Lambda_k (X_j - X_i)|^2}{h^2(X_i)} \right);$$
      4 Compute
            $$\lambda(X_i) = \lambda_{\min}\left( V_{h(X_i),\Lambda_k}(X_i) \right)$$
        and the weighted sum $\widehat{\beta}_k$ of $\widehat{\beta}_k(X_i)$:
            $$\widehat{\beta}_k = \sum_{i=1}^{n} \lambda(X_i) \widehat{\beta}_k(X_i). \qquad (4.1)$$
        Set $\widehat{\theta}_k = \widehat{\beta}_k / |\widehat{\beta}_k|$.
      5 Set $k = k + 1$, $\rho_k = \rho_{k-1}/2$, $\Lambda_k = I + \rho_k^{-1} \widehat{\theta}_k \widehat{\theta}_k^T$ and $N_k = 2^d N_{k-1}$.
        Continue with Step 2;
  End While
6 Set $\widehat{\theta} = \widehat{\theta}_k$.  Terminate;
```

Note that in (4.1), Step 4 we compute a weighted sum of $\widehat{\beta}_k(X_i)$. The reason for this modification of the algorithm is rather transparent: the weight of the i-th term in the sum is large if the correspondent matrix $V_{h(X_i),\Lambda_k}(X_i)$ is well conditioned, and, on the contrary, the terms which correspond to ill-conditioned matrices have relatively small weights.

4.2 Simulation results

We provide here an example of the use of Algorithm 2 in the following simulation example. We consider a heteroscedastic single-index regression model

$$Y_i = g\left(X_i^T\theta^*\right) + \varepsilon_i,$$

where

$$\begin{aligned} g(u) &= u^2, \\ \varepsilon_i | X_i &\sim \mathcal{N}\left(0, g(X_i^T\theta^*)\right), \end{aligned}$$

and $X_i = (X_i^{(1)}, \dots, X_i^{(d)})^T \in I\!R^d$ for $d = 4$ and $d = 8$. In the case $d = 4$ we take $\theta^* = (-1,1,1,1)^T/2$. When $d = 8$ $\theta^* = (-1,1,1,1,1,1,1,1)^T/\sqrt{8}$. In both situations $X_i^{(1)} \sim \mathcal{N}(0.5,1)$, $X_i^{(k)} \sim \mathcal{N}(0,1)$ for $k = 2,\dots,4$ and $k = 2,\dots,8$ respectively. The components of X_i are independent.

On Figure 1 we present the dependence of the mean-square error $|\widehat{\theta}_n - \theta^*|$ of the estimate on the sample size n for the 4-dimensional case. The curve is plotted in the log / log axes. The result is averaged over 40 replicates of the observation sequence. The ticks of the Y-axis correspond to $\frac{2}{10+5k}$, the X-axis ticks correspond to $(10+5k)^2$, $k = 1,\dots,5$. Note that the "diagonal" points of the grid lie on the line $Y = \frac{2}{\sqrt{X}}$.

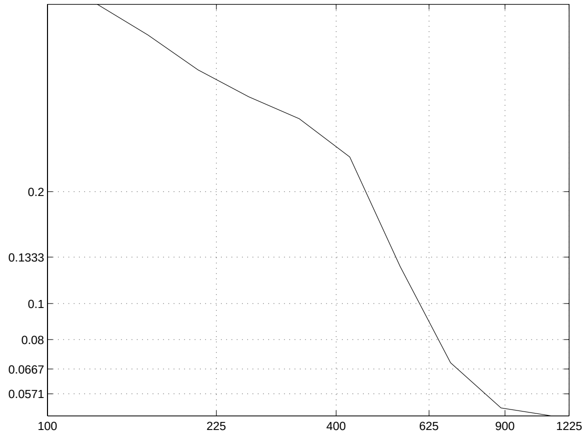

Figure 1. Mean-square error as a function of the sample size n

The result of the analogous experiment for $d = 8$ is presented on Figure 2. The ticks of the Y-axis correspond to $\frac{5}{10+5i}$, the X-axis ticks correspond to $(10+5i)^2$, $i = 1,\dots,7$. Now the "diagonal" corresponds to $Y = \frac{5}{\sqrt{X}}$. The results are clearly in accordance with the root-n consistency of the estimate claimed in Theorem 1.

The simulations were performed in MATLAB on a P2-266 PC. It takes around 20 sec to compute a four-step estimate for the sample size $n = 1000$.

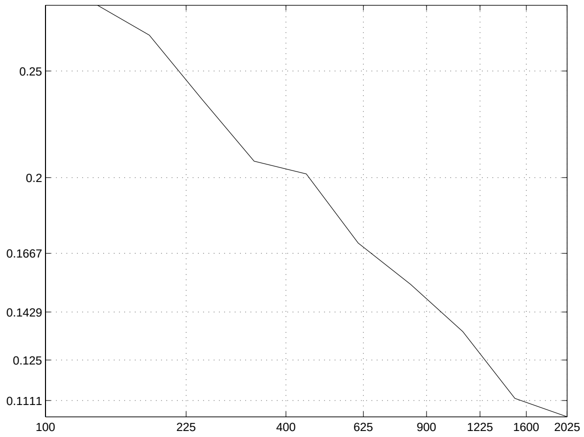

Figure 2. Mean-square error as a function of the sample size n

5 Proofs

Lemma 1 *Let $K(x) \geq \alpha > 0$ for $|x| \leq 1/2$. Then $C_N \leq 8^d/\alpha$.*

Proof: Let Σ be a covering of the ball $B = \{x : |\Lambda_k(x - X_i)| \leq h\}$ with the balls B_k, $k = 1, ..., M$ such that the diameter of each B_k is less than $h/2$; i.e. for any two points x, $y \in B_k$, $|\Lambda_k(x - y)| \leq h/2$. One can easily show that there is such a covering Σ with the cardinality $M = (2 \times 4)^d$. Then if X_j belongs to B_k and

$$B(X_j) = \{x : |\Lambda_k(x - X_j)| \leq h/2\},$$

then $B_k \subseteq B(X_j)$, and due to Assumption 1

$$\sum_{X_j \in B_k} \frac{K\left(\frac{|\Lambda_k(X_j - X_i)|^2}{h_k^2}\right)}{\sum_{k=1}^{n} K\left(\frac{|\Lambda_k(X_k - X_j)|^2}{h_k^2}\right)} \leq \sum_{X_j \in B_k} \frac{K\left(\frac{|\Lambda_k(X_j - X_i)|^2}{h_k^2}\right)}{\sum_{X_j \in B_k} \alpha} \leq \alpha^{-1}.$$

Then

$$\begin{aligned} \sum_{j=1}^{n} \frac{K\left(\frac{|\Lambda_k(X_j - X_i)|^2}{h_k^2}\right)}{\sum_{l=1}^{n} K\left(\frac{|\Lambda_k(X_l - X_j)|^2}{h_k^2}\right)} &\leq \sum_{k=1}^{M} \sum_{X_j \in B_k} \frac{K\left(\frac{|\Lambda_k(X_j - X_i)|^2}{h_k^2}\right)}{\sum_{l=1}^{n} K\left(\frac{|\Lambda_k(X_l - X_j)|^2}{h_k^2}\right)} \\ &\leq \frac{M}{\alpha} \leq \frac{8^d}{\alpha}. \end{aligned}$$

∎

Note that by Assumption 1, due to the continuity of $K(x)$, $K(x) > 0$ for $|x| < 1$ implies that $K(x) \geq \alpha > 0$ for $|x| \leq 1/2$.

5.1 Proof of Theorem 1

In what follows C_i stands for a generic positive constant which value depends only on C_K, C_V, C_N and d. The result of Theorem 1 is heavily based on properties of a single iteration of Algorithm 1. We now turn to the study of the estimate $\widehat{\theta}_k$ obtained after the k-th iteration.

Let a vector θ on the unit sphere be the initial estimate of θ^*. We introduce the following notations:

$$\Lambda_{\rho,\theta} = I + \rho^{-1}\theta\theta^T, \quad \Pi_{\rho,\theta} = \Lambda_{\rho,\theta}^{-1} = I - \frac{\theta\theta^T}{\rho+1}$$

and define

$$\begin{aligned}
V_{h,\Lambda_{\rho,\theta}}(X_i) &= \sum_{j=1}^{n}(X_j - X_i)(X_j - X_i)^T K\left(h^{-2}|\Lambda_{\rho,\theta}(X_j - X_i)|^2\right), \\
\widehat{\beta}(X_i) &= V_{h,\Lambda_{\rho,\theta}}^{-1}(X_i)\sum_{j=1}^{n}(X_j - X_i)(Y_j - Y_i)\, K\left(h^{-2}|\Lambda_{\rho,\theta}(X_j - X_i)|^2\right), \\
\widehat{\beta} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{\beta}(X_i), \\
\widehat{\theta} &= \frac{\widehat{\beta}}{|\widehat{\beta}|}. \qquad (5.1)
\end{aligned}$$

The crux of the proof of Theorem 1 is the following proposition, which is of interest by its own:

Proposition 1 *Suppose that Assumptions 1–5 of Theorem 1 hold true.*

1. *Let $h = h_0$ in (2.5), $\rho = 1$ and $\Lambda_{\rho,\theta} = I$. Then the estimate $\widehat{\theta}$, defined in (5.1), satisfies*

$$\boldsymbol{P}\left(|\widehat{\theta} - \theta^*| \geq \frac{2C_1\sqrt{2\log n}}{h_0\sqrt{n}} + C_2 h_0\right) \leq 1/n \qquad (5.2)$$

where

$$\begin{aligned}
C_1 &= \frac{C_V(1 + C_N)\sigma}{|\beta^*|}, \\
C_2 &= \frac{C_V C_g}{|\beta^*|},
\end{aligned}$$

2. *Let $\rho \le 1/2$ and $\alpha = \gamma/\rho$ satisfy $\alpha(1+\rho+2\alpha) \le 1/6$ and $\alpha < \beta^*/4$. Then it holds for the estimate $\widehat{\theta}$*

$$\boldsymbol{P}\left(\sup_{|\theta-\theta^*|\le\gamma} |\widehat{\theta}-\theta^*| \ge q\left(1+5\frac{q}{\rho}\right)\right) \le \exp(-z^2/2) + \frac{2}{n} \qquad (5.3)$$

where

$$q = \frac{C_1}{h\sqrt{n}}z + 4C_2 h\rho^2 + \frac{C_3\alpha\sqrt{2\log n}}{h\sqrt{n}}$$

with the same C_2 and C_1 and

$$C_3 = \frac{8\, C_w\, C_V\, C_N(2C_V+1)\left(1+\sqrt{2+d}\right)\sigma}{|\beta^*|}$$

The proof of this proposition is placed in the next section. We return now to the proof of Theorem 1.

Let k^* denote the total number of iterations. Recall that this number satisfies $\rho_{k^*} = 2^{-k^*} \ge \left(\frac{\log n}{n}\right)^{1/3}$ and hence $k^* \le \log n$. We set

$$\begin{aligned}
\gamma_1 &= C_2 h_0 + 2\frac{C_1\sqrt{2\log n}}{h_0\sqrt{n}},\\
\alpha_k &= \gamma_k/\rho_k,\\
\gamma_{k+1} &= \delta_{k+1}(1+5\frac{\delta_{k+1}}{\rho_{k+1}}).
\end{aligned}$$

where

$$\delta_{k+1} = 2C_2 h_k \rho_k^2 + \frac{(C_1+C_3\alpha_k)\sqrt{2\log n}}{h_k\sqrt{n}}$$

for $1 \le k < k^*$. We put for some $C_4 > 0$ large enough

$$\alpha_n^* = C_4\left(\frac{\log n}{n}\right)^{1/6}.$$

Evidently, for n large enough α_n^* satisfies $\alpha_n^*(1+\rho+2\alpha^*) \le 1/6$ and $\alpha_n^* \le |\beta^*|/4$.

We first show that $\alpha_k \le \alpha_n^*$ for all $k < k^*$. We proceed by induction. Obviously,

$$\alpha_1 = 2\gamma_1 = 2C_2 h_0 + \frac{4C_1\sqrt{2\log n}}{h_0\sqrt{n}} \le C_5\frac{\log^{1/2} n}{n^{1/4}} = o(\alpha_n^*).$$

Let now $\alpha_k \le \alpha_n^*$. Since $\rho_{k+1} = \rho_k/2$, $h_{k+1} = \min\{1, 2h_k\}$, $h_k \rho_k \le h_0$ and $\rho_{k+1} \ge \rho_{k^*}$, we conclude that

$$\begin{aligned} \frac{\delta_{k+1}}{\rho_{k+1}} &\le \frac{2C_2 h_k \rho_k^2}{\rho_{k+1}} + \frac{(C_1 + C_3\alpha_k)\sqrt{2\log n}}{\rho_{k+1} h_k \sqrt{n}} \\ &\le 4C_2 h_k \rho_k + \frac{(C_1 + C_3\alpha^*)\sqrt{2\log n}}{\rho_{k^*}\sqrt{n}} \\ &\le C_6 \left(\frac{\log n}{n}\right)^{1/6}. \end{aligned}$$

Thus

$$\alpha_{k+1} = \frac{\gamma_{k+1}}{\rho_{k+1}} \le 2\frac{\delta_{k+1}}{\rho_{k+1}} \le C_4 \left(\frac{\log n}{n}\right)^{1/6}$$

as required.

Next we show that the k th estimate $\widehat{\theta}_k$ satisfies

$$\boldsymbol{P}(|\widehat{\theta}_k - \theta^*| \ge \gamma_k) \le \frac{3k}{n}, \tag{5.4}$$

We again proceed by induction. Due to the bound (5.2) in Proposition 1, the initial estimate $\widehat{\theta}$ satisfies

$$\boldsymbol{P}(|\widehat{\theta}_0 - \theta^*| \ge \gamma_0) \le \frac{1}{n}. \tag{5.5}$$

Let now (5.4) be satisfied for some $k < k^* - 1$. Since $\alpha_{k+1} \le \alpha_n^*$, we may apply the bound (5.3) with $z = \sqrt{2\log n}$ to obtain

$$\boldsymbol{P}\left(|\widehat{\theta}_{k+1} - \theta^*| \ge \gamma_{k+1}\right) \le \frac{3k}{n} + \frac{3}{n} = \frac{3(k+1)}{n},$$

as required.

Now we use the bound (5.3) one more time with $\gamma = \gamma_{k^*}$, $h = h_{k^*} = 1$, $\alpha = \alpha_{k^*}$ and $(\frac{\log n}{n})^{1/3} < \rho = \rho_{k^*} \le 2(\frac{\log n}{n})^{1/3}$:

$$\boldsymbol{P}\left(|\widehat{\theta}_{k^*} - \theta^*| \ge q_{k^*}\left(1 + 5\frac{q_{k^*}}{\rho_{k^*}}\right)\right) \le \exp\left(-z^2/2\right) + \frac{3k^*}{n},$$

where

$$q_{k^*} = \frac{C_1}{\sqrt{n}} z + 2C_2 \rho_{k^*}^2 + C_3 \alpha_{k^*} \sqrt{\frac{2\log n}{n}}.$$

Since $h_{k^*-1} = 1$, we get for sufficiently large n

$$\alpha_{k^*} = \gamma_{k^*}/\rho_{k^*} \le 8C_2\, \rho_{k^*-1} + \frac{2(C_1 + C_3\alpha^*)\sqrt{2\log n}}{\rho_{k^*-1}\sqrt{n}} \le 4C_1 \left(\frac{2\log n}{n}\right)^{\frac{1}{6}}.$$

Hence,

$$\begin{aligned} q_{k^*} &= \frac{C_1}{\sqrt{n}} z + 4C_2 \left(\frac{\log n}{n}\right)^{2/3} + C_3 \alpha_{k^*} \sqrt{\frac{2\log n}{n}} \\ &\le \frac{C_1}{\sqrt{n}} z + 4(C_2 + C_3 C_1) \left(\frac{\log n}{n}\right)^{2/3}, \end{aligned}$$

and $\frac{q_{k^*}}{\rho_{k^*}} \le 2C_1 \left(\frac{\log n}{n}\right)^{1/6}$. When summing up we obtain from (5.3)

$$\boldsymbol{P}\left(|\widehat{\theta}_{k^*} - \theta^*| \ge \frac{C_1}{\sqrt{n}} z + 2(4C_2 + 4C_3\, C_1 + 10C_1^2) \left(\frac{2\log n}{n}\right)^{2/3}\right) \\ \le \exp(-z^2/2) + \frac{3\log n}{n}.$$

∎

5.2 Proof of Proposition 1

Given a vector θ, we denote by u the vector $\rho^{-1}\Pi_{\rho,\theta^*}\theta$, and by A_u the matrix $\Pi_{\rho,\theta^*}\Lambda^2_{\rho,\theta}\Pi_{\rho,\theta^*}$,

$$\begin{aligned} u &= \rho^{-1}\Pi_{\rho,\theta^*}\theta, \\ A_u &= \Pi_{\rho,\theta^*}\Lambda^2_{\rho,\theta}\Pi_{\rho,\theta^*}. \end{aligned}$$

If u^* corresponds to θ^*, that is, $u^* = \rho^{-1}\Pi_{\rho,\theta^*}\theta^*$, then obviously $A_{u^*} = I$. We also set $Z_{ij} = h^{-1}\Lambda_{\rho,\theta^*}(X_j - X_i)$, so that

$$K\left(\frac{(X_j - X_i)\Lambda^2_{\rho,\theta}(X_j - X_i)}{h^2}\right) = K\left(Z_{ij}^T A_u Z_{ij}\right).$$

It is convenient to denote

$$V_u(X_i) = \sum_{z=1}^{n} Z_{ij} Z_{ij}^T K\left(Z_{ij}^T A_u Z_{ij}\right)$$

and

$$\widehat{b}_u(X_i) = h^{-1} V_u^{-1}(X_i) \sum_{j=1}^{n} Z_{ij}(Y_j - Y_i) K\left(Z_{ij}^T A_u Z_{ij}\right).$$

With this notation, it holds $\Pi_{\rho,\theta^*}\widehat{\beta}(X_i) = \widehat{b}_u(X_i)$. Indeed, for $\Lambda = \Lambda_{\rho,\theta}$,

$$\begin{aligned}
V_{h,\Lambda}(X_i) &= \sum_{j=1}^{n}(X_j - X_i)(X_j - X_i)^T K\left(h^{-2}\left|\Lambda(X_j - X_i)\right|^2\right) \\
&= \sum_{j=1}^{n}(X_j - X_i)(X_j - X_i)^T K\left(Z_{ij}^T A_u Z_{ij}\right) \\
&= h^2 \Pi_{\rho,\theta^*}\sum_{j=1}^{n} Z_{ij} Z_{ij}^T K\left(Z_{ij}^T A_u Z_{ij}\right)\Pi_{\rho,\theta^*} \\
&= h^2 \Pi_{\rho,\theta^*} V_u(X_i)\Pi_{\rho,\theta^*}
\end{aligned}$$

and hence

$$\begin{aligned}
\Pi_{\rho,\theta^*}\widehat{\beta}(X_i) &= \Pi_{\rho,\theta^*}\, V_{h,\Lambda}^{-1}(X_i)\sum_{j=1}^{n}(X_j - X_i)(Y_j - Y_i)K\left(Z_{ij}^T A_u Z_{ij}\right) \\
&= h^{-1} V_u^{-1}(X_i)\sum_{j=1}^{n} Z_{ij}(Y_j - Y_i)K\left(Z_{ij}^T A_u Z_{ij}\right) \\
&= \widehat{b}_u(X_i). \qquad (5.6)
\end{aligned}$$

In the sequel we need the following simple lemma:

Lemma 2

(i) *u^* satisfies: $u^* = (1+\rho)^{-1}\theta^*$ and $|u^*| = 1/(1+\rho)$;*

(ii) *If $|u - u^*| \le \alpha$ then $|u| \le (1+\rho)^{-1} + \alpha$;*

Let $\rho \le 1/2$ and $\alpha = \gamma/\rho$ fulfills $2\alpha(1+\rho+2\alpha) \le 1/3$. Then it holds for every u with $|u - u^*| \le \alpha$

(iii) *for every unit vector v in $I\!\!R^d$*

$$\begin{aligned}
\left|\frac{\partial}{\partial u} v^T A_u v\right| &\le 2(1+\rho+2\alpha), \\
\left|v^T A_u v\right| &\le 1 + 2\alpha(1+\rho+2\alpha) \le 4/3, \\
\left|v^T A_u^{-1} v\right| &\le \frac{1}{1 - 2\alpha(1+\rho+2\alpha)} \le 3/2;
\end{aligned}$$

(iv) *if $z^T A_u z \le 1$ for some vector z in $I\!\!R^d$, then*

$$\begin{aligned}
|z|^2 &\le 3/2; \\
\left|\frac{\partial}{\partial u} z^T A_u z\right| &\le \sqrt{12},
\end{aligned}$$

Proof: (i). By definition

$$\Pi_{\rho,\theta^*}\theta^* = \theta^* - (1+\rho)^{-1}\theta^*\theta^{*T}\theta^* = \rho(1+\rho)^{-1}\theta^*$$

so that $u^* = (1+\rho)^{-1}\theta^*$ and (i) follows. (ii) is the straightforward consequence of (i) and the inequality $|u-u^*| \le \alpha$.

(iii). One has

$$A_u = \Pi_{\rho,\theta^*}\left(\rho^{-1}\theta\theta^T + I\right)^2 \Pi_{\rho,\theta^*} = (1+2\rho)uu^T + \Pi^2_{\rho,\theta^*}$$

and, since $|v|=1$, using also (ii) we obtain

$$\left|\frac{\partial}{\partial u}v^T A_u v\right| = 2(1+2\rho)\left|v^T uv\right| \le 2(1+2\rho)|u| \le 2(1+2\rho)\left((1+\rho)^{-1}+\alpha\right).$$

Now the fist statement in (iii) follows from the trivial inequality

$$(1+2\rho)\left((1+\rho)^{-1}+\alpha\right) \le 1+\rho+2\alpha$$

and the lemma conditions. The other two inequalities follow from the first one in view of $v^T A_{u^*} v = 1$.

(iv). Note first that by (iii) the inequality $z^T A_u z \le 1$ implies $|z|^2 \le |A_u^{-1}| \le 3/2$. Now, let $v = z/|z|$. Since $1 \ge z^T A_u z \ge (1+2\rho)|z^T u|^2$, we get

$$\begin{aligned}\left|\frac{\partial}{\partial u}z^T A_u z\right|^2 &= \left|2(1+2\rho)z^T uz\right|^2 = 4(1+2\rho)^2|z^T u|^2|z|^2 \\ &\le 4(1+2\rho)|z|^2 \le 12\end{aligned}$$

and (iv) follows in view of $\rho \le 1/2$ and $|z|^2 \le 3/2$. ∎

In the next technical lemma we collect some useful properties of the matrices $V_u(X_i)$. We use the notation

$$N_u(X_i) = \sum_{j=1}^n K\left(Z_{ij}^T A_u Z_{ij}\right).$$

Lemma 3 *Let* $\rho \le 1/2$, $|u-u^*| \le \alpha$ *and* $\alpha(1+\rho+\alpha) \le 1/6$. *Then for all* i

(i) $\left|V_{u^*}^{-1}(X_i)\right| \le C_V N_{u^*}(X_i)$;

(ii) $|V_u^{-1}(X_i)| \le \frac{4}{3}C_V\, N_u^{-1}(X_i)$;

(iii) $\left|\frac{\partial}{\partial u}V_u(X_i)\right| \le 3\sqrt{3}\,C_w\, N_u(X_i)$;

(iv) $\left|\frac{\partial}{\partial u}V_u^{-1}(X_i)\right| \le \frac{16}{\sqrt{3}}\,C_w\, C_V^2 N_u^{-1}(X_i)$;

Proof: (i) and (ii). For any unit vector $v \in I\!\!R^d$, we get from the definitions of $V_u(X_i)$ and Assumption 5 that

$$N_u(X_i)\, v^T V_u^{-1}(X_i) v = v^T \Pi_{\rho,\theta^*} \Lambda_{\rho,\theta} \overline{V}_{h,\Lambda_{\rho,\theta}}^{-1} \Lambda_{\rho,\theta} \Pi_{\rho,\theta^*} v \le C_V \left| \Lambda_{\rho,\theta} \Pi_{\rho,\theta^*} v \right|^2,$$

and, in particular, $v^T V_{u^*}^{-1}(X_i) v \le C_V N_{u^*}(X_i)$. Next, by Lemma 2, (iii)

$$\left| \Lambda_{\rho,\theta} \Pi_{\rho,\theta^*} v \right|^2 = v^T \Pi_{\rho,\theta^*} \Lambda_{\rho,\theta} \Lambda_{\rho,\theta} \Pi_{\rho,\theta^*} v = v^T A_u v \le 4/3$$

as required.

(iii). Clearly,

$$\frac{\partial}{\partial u} V_u(X_i) v = \sum_{j=1}^n Z_{ij} Z_{ij}^T K' \left(Z_{ij}^T A_u Z_{ij} \right) \frac{\partial}{\partial u} (Z_{ij}^T A_u Z_{ij}) v.$$

Since the kernel K vanishes outside $[0,1]$, we may consider only those j that $Z_{ij}^T A_u Z_{ij} \le 1$ which by Lemma 2, (iv) implies $\left| Z_{ij} Z_{ij}^T \right| \le 3/2$. Now in view of Assumption 5

$$\sum_{j=1}^n \left| K' \left(Z_{ij}^T A_u Z_{ij} \right) \right| \le C_w N_u(X_i).$$

Hence, using Lemma 2, (iv), we derive

$$\left| \frac{\partial}{\partial u} V_u(X_i) v \right| \le 3\sqrt{3} \sum_{j=1}^n \left| K' \left(Z_{ij}^T A_u Z_{ij} \right) \right| \le 3\sqrt{3}\, C_w\, N_u(X_i).$$

(iv). The second and third statements yield

$$\begin{aligned} \left| \frac{\partial}{\partial u} V_u^{-1}(X_i) v \right| &= \left| V_u^{-1}(X_i) \frac{\partial}{\partial u} V_u(X_i) v\, V_u^{-1}(X_i) \right| \\ &\le \left| V_u^{-1}(X_i) \right|^2 \left| \frac{\partial}{\partial u} V_u(X_i) v \right| \\ &\le \frac{1}{N_u(X_i)} (4/3)^2 C_V^2\, 3\sqrt{3} C_w \\ &= \frac{16}{\sqrt{3}} C_V^2 C_w. \end{aligned}$$

■

Now we turn directly to the proof of Proposition 1. By (5.6) we have the following decomposition for $\widehat{b}_u(X_i) = \Pi_{\rho,\theta^*} \widehat{\beta}(X_i)$:

$$\begin{aligned} \widehat{b}_u(X_i) &= h^{-1} V_u^{-1}(X_i) \sum_{j=1}^n [f(X_j) - f(X_i)]\, Z_{ij} K \left(Z_{ij}^T A_u Z_{ij} \right) \\ &\quad + h^{-1} V_u^{-1}(X_i) \sum_{|j=1}^n (\varepsilon_j - \varepsilon_i) Z_{ij} K \left(Z_{ij}^T A_u Z_{ij} \right) \\ &= b_u(X_i) + \zeta_u(X_i). \end{aligned}$$

Let β^* stand for the averaged derivative of f, i.e.

$$\beta^* = \frac{1}{n}\sum_{i=1}^n \nabla f(X_i).$$

Then

$$\widehat{b}_u(X_i) - \Pi_{\rho,\theta^*}\beta^* = \frac{1}{n}\sum_{i=1}^n \bigl(b_u(X_i) - \Pi_{\rho,\theta^*}\nabla f(X_i)\bigr) + \frac{1}{n}\sum_{i=1}^n \zeta_u(X_i). \tag{5.7}$$

We denote

$$\begin{aligned} \Delta_u &= \frac{1}{n}\sum_{i=1}^n \bigl(b_u(X_i) - \Pi_{\rho,\theta^*}\nabla f(X_i)\bigr), \\ \zeta_u &= \frac{1}{n}\sum_{i=1}^n \zeta_u(X_i). \end{aligned}$$

Lemma 4 *Suppose that* $\rho \le 1/2$ *and* $\alpha = \gamma/\rho$ *fulfills* $\alpha(1+\rho+\alpha) \le 1/6$ *, then*

$$|\Delta_{u^*}| \le 0.5 h\rho^2 C_V C_g,$$
$$\sup_{u:|u-u^*|\le\alpha} |\Delta_u| \le 2h\rho^2 C_V C_g.$$

Proof: By definition of $V_u(X_i)$

$$\begin{aligned} I &= \left(V_u(X_i)\Pi_{\rho,\theta^*}\right)^{-1} V_u(X_i)\Pi_{\rho,\theta^*} \\ &= h^{-1}\Lambda_{\rho,\theta^*} V_u^{-1}(X_i) \sum_{j=1} Z_{ij}(X_j - X_i)^T K(Z_{ij}^T A_u Z_{ij}) \end{aligned}$$

and we get

$$\begin{aligned} \Pi_{\rho,\theta^*}\nabla f(X_i) &= \Pi_{\rho,\theta^*} g'(X_i^T\theta^*)\theta^* \\ &= h^{-1} V_u^{-1}(X_i) \sum_{j=1}^n g'(X_i^T\theta^*)\, Z_{ij}(X_j - X_i)^T \theta^* K(Z_{ij}^T A_u Z_{ij}) \end{aligned}$$

and hence

$$\begin{aligned} &b_u(X_i) - \Pi_{\rho,\theta^*}\nabla f(X_i) = h^{-1} V_u^{-1}(X_i) \\ &\qquad \times \sum_{j=1}^n Z_{ij}\left[f(X_j) - f(X_i) - g'(X_i^T\theta^*)\,(X_j - X_i)^T\theta^*\right] K(Z_{ij}^T A_u Z_{ij}). \end{aligned} \tag{5.8}$$

Since the kernel K vanishes outside $[-1,1]$, we may consider in this sum only those X_j 's that $|Z_{ij}^T A_u Z_{ij}| \le 1$. For every such X_j, it follows from Lemma 2, (iv) that

$$\begin{aligned} |(X_i - X_j)^T \theta^*|^2 &= |(\Lambda_{\rho,\theta^*}(X_i - X_j))^T \Pi_{\rho,\theta^*} \theta^*|^2 \\ &= h^2 \rho^2 |Z_{ij}^T u^*|^2 \\ &\le h^2 \rho^2 |Z_{ij}|^2 |u^*|^2 \\ &\le \frac{3}{2} h^2 \rho^2 (1+\rho)^{-2}. \end{aligned}$$

Then, due to Assumption 3,

$$\begin{aligned} r_{i,j} &\stackrel{\Delta}{=} |g(X_j^T \theta^*) - g(X_i^T \theta^*) - g'(X_i^T \theta^*)(X_j - X_i)^T \theta^*| \\ &\le \frac{C_g}{2} |(X_j - X_i)^T \theta^*|^2 \\ &\le \frac{3}{4} C_g h^2 \rho^2, \end{aligned}$$

and by Assumption 5 we get from (5.8) and Lemma 3:

$$\begin{aligned} &\left| b_u(X_i) - \Pi_{\rho,\theta^*} \nabla f(X_i) \right| \\ &\quad \le h^{-1} \left| V_u^{-1}(X_i) \sum_{j=1}^n Z_{ij} K(Z_{ij}^T A_u Z_{ij}) r_{i,j} \right| \\ &\quad \le h^{-1} \left| V_u^{-1}(X_i) \right| \left| \sum_{j=1}^n |Z_{ij}| K(Z_{ij}^T A_u Z_{ij}) r_{i,j} \right| \\ &\quad \le h^{-1} \frac{4}{3} C_V \frac{3}{4} C_g h^2 \rho^2 \sqrt{3/2} \\ &\quad \le 2 h \rho^2 C_V C_g \end{aligned}$$

as required.

Similar and even simpler calculations with $u = u^*$ and $A_{u^*} = I$ lead to the bound $|\Delta_{u^*}| \le 0.5 h \rho^2 C_V C_g$. ∎

We now turn to the study of the stochastic part of the error. We have the following decomposition for ζ_u:

$$\begin{aligned} \zeta_u &= \frac{1}{nh} \sum_{i=1}^n \sum_{j=1}^n V_u^{-1}(X_i)(\varepsilon_j - \varepsilon_i)\, Z_{ij} K(Z_{ij}^T A_u Z_{ij}) \\ &= \frac{1}{nh} \sum_{i=1}^n \varepsilon_i \left(\sum_{j=1}^n V_u^{-1}(X_i)\, Z_{ij}\, K(Z_{ij}^T A_u Z_{ij}) + \sum_{j=1}^n V_u^{-1}(X_j)\, Z_{ij}\, K(Z_{ij}^T A_u Z_{ij}) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left(c_u^{(1)}(X_i) + c_u^{(2)}(X_i) \right) \end{aligned}$$

with

$$
\begin{aligned}
c_u^{(1)}(X_i) &= h^{-1} V_u^{-1}(X_i) \sum_{j=1}^n Z_{ij}\, K(Z_{ij}^T A_u Z_{ij}),\\
c_u^{(2)}(X_i) &= h^{-1} \sum_{j=1}^n V_u^{-1}(X_j)\, Z_{ij}\, K(Z_{ij}^T A_u Z_{ij}).
\end{aligned}
$$

Let us show that $c_u^{(\ell)}(X_i)$ are uniformly bounded and Lipschitz-continuous in u, $\ell = 1, 2$.

Lemma 5 *Let $\rho \le 1/2$ and $\alpha = \gamma/\rho$ satisfy $\alpha(1+\rho+\alpha) \le 1/6$. Then for all $i \le n$*

$$
|c_{u^*}^{(1)}(X_i)| \le \frac{C_V}{h}, \qquad |c_{u^*}^{(2)}(X_i)| \le \frac{C_V C_N}{h}
$$

and

$$
\sup_{u:|u-u^*|\le\alpha} \left| \frac{\partial}{\partial u} c_u^{(\ell)}(X_i) \right| \le 4\sqrt{2}\, C_w\, C_V\, C_N (2C_V + 1) h^{-1}, \qquad \ell = 1, 2
$$

Proof: We have by Lemma 3, (i)

$$
\begin{aligned}
h|c_{u^*}^{(1)}(X_i)| &= \left| V_{u^*}^{-1}(X_i) \sum_{j=1}^n Z_{ij}\, K(Z_{ij}^T Z_{ij}) \right| \\
&\le \left| V_{u^*}^{-1}(X_i) \right| \sum_{j=1}^n |Z_{ij}|\, K(Z_{ij}^T Z_{ij}) \\
&\le \frac{C_V}{N_{u^*}(X_i)} \sum_{j=1}^n K(Z_{ij}^T Z_{ij}) \\
&\le C_V.
\end{aligned}
$$

In the same way we have for $c_{u^*}^{(2)}(X_i)$ by Assumption 5

$$
\begin{aligned}
h|c_{u^*}^{(2)}(X_i)| &\le \left| \sum_{j=1}^n V_{u^*}^{-1}(X_j)\, Z_{ij}\, K(Z_{ij}^T Z_{ij}) \right| \\
&\le C_V \sum_{j=1}^n \frac{1}{N_{u^*}(X_j)} |Z_{ij}|\, K(Z_{ij}^T Z_{ij}) \\
&\le C_V C_N.
\end{aligned}
$$

Now we compute the derivative of $c_u^{(2)}(X_i)$. (The proof for $c_u^{(1)}(X_i)$ can be carried out in an analogous way.) First we observe that

$$\begin{aligned}
& h \left| \frac{\partial}{\partial u} c_u^{(2)}(X_i) \right| \\
&= \left| \frac{\partial}{\partial u} \sum_{j=1}^{n} V_u^{-1}(X_j)\, Z_{ij}\, K(Z_{ij}^T A_u Z_{ij}) \right| \\
&\le \sum_{j=1}^{n} \left| \frac{\partial}{\partial u} V_u^{-1}(X_j) Z_{ij} \right| K(Z_{ij}^T A_u Z_{ij}) + \sum_{j=1}^{n} |V_u^{-1}(X_j) Z_{ij}| \left| \frac{\partial}{\partial u} K(Z_{ij}^T A_u Z_{ij}) \right| \\
&= \delta_1 + \delta_2 .
\end{aligned}$$

When using Lemma 2, (iv), Lemma 3, (iv) and Assumption 5 we bound

$$\begin{aligned}
\delta_1 &\le \frac{16}{\sqrt{3}} C_w\, C_V^2 \sqrt{3/2} \sum_{j=1}^{n} N_u^{-1}(X_j) K(Z_{ij}^T A_u Z_{ij}) \\
&\le 8\sqrt{2}\, C_w\, C_V^2\, C_N
\end{aligned}$$

and similarly

$$\begin{aligned}
\delta_2 &\le \frac{4}{3} C_V \sqrt{3/2} \sum_{j=1}^{n} N_u^{-1}(X_j) 4 |K'(Z_{ij}^T A_u Z_{ij})| \\
&\le 4\sqrt{2}\, C_V\, C_N\, N_u^{-1}(X_i) \sum_{j=1}^{n} |K'(Z_{ij}^T A_u Z_{ij})| \\
&\le 4\sqrt{2}\, C_V\, C_N\, C_w
\end{aligned}$$

and the assertion follows. ∎

Let $c_u(X_i) = c_u^{(1)}(X_i) + c_u^{(2)}(X_i)$. Then $\zeta_u = \sum_{i=1}^{n} c_u(X_i)\varepsilon_i$ and it follows from Lemma 5 that

$$|c_{u^*}(X_i)| \le C_V(1 + C_N) h^{-1}, \qquad \left| \frac{\partial}{\partial u} c_u(X_i) \right| \le 8\sqrt{2}\, C_w\, C_V\, C_N (2C_V + 1) h^{-1}.$$

To bound the stochastic term ζ_u we use the following general result.

Lemma 6 *Let $0 < \alpha \le 1/2$ and let functions $a_i(u)$ obey the conditions*

$$|a_i(u^*)| \le \kappa_1 \tag{5.9}$$

$$\sup_{|u-u^*| \le \alpha} \left| \frac{\partial}{\partial u} a_i(u) \right| \le \kappa_2, \qquad i = 1, \ldots, n. \tag{5.10}$$

If ε_i are independent $\mathcal{N}(0,\sigma^2)$ -distributed random variables, then it holds for each $z \ge 1$

$$\boldsymbol{P}\left(\sup_{|u-u^*|\le\alpha} \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n} a_i(u)\varepsilon_i\right| > z\sigma\kappa_1 + \sigma\kappa_2\alpha\left(2+\sqrt{(2+d)\log n}\right)\right) \le \exp\left(-z^2/2\right) + \frac{2}{n}.$$

Proof: Let B_α be the ball $\{u:\ |u-u^*|\le\alpha\}$ and Σ_α be the ϵ-net on B_α such that for any $u\in B_\alpha$ there is an element u_ℓ of Σ_α such that $|u-u_\ell|\le\frac{\alpha}{\sqrt{n}}$. It is easy to see that such a net with cardinality $N(\Sigma_\alpha)\le(4n)^{d/2}$ can be constructed. For a $u_\ell\in\Sigma_\alpha$ we denote

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

Then by (5.10)

$$\boldsymbol{E}|\eta(u_\ell)|^2=\frac{\sigma^2}{n}\sum_{i=1}^{n}|a_j(u_\ell)-a_j(u^*)|^2\le\sigma^2\kappa_2^2\alpha^2,$$

and for any $t\ge1$

$$\boldsymbol{P}(|\eta(u_\ell)|>t)\le\exp\left(-\frac{t^2}{2\boldsymbol{E}|\eta(u_\ell)|^2}\right)\le\exp\left(-\frac{t^2}{2\sigma^2\kappa_2^2\alpha^2}\right).$$

Hence, if $t=\sigma\kappa_2\alpha\sqrt{2\log nN(\Sigma_\alpha)}$,

$$\begin{aligned}\boldsymbol{P}\left(\sup_{u_\ell\in\Sigma_\alpha}|\eta(u_\ell)|>t\right) &\le \sum_{\ell=1}^{N(\Sigma_\alpha)}\boldsymbol{P}(|\eta(u_\ell)|>t)\\ &\le N(\Sigma_\alpha)\exp\left(-\log nN(\Sigma_\alpha)\right)=\frac{1}{n}.\end{aligned} \tag{5.11}$$

Meanwhile, by construction of the net Σ_α, for any $u\in B_\gamma$ there is $u_\ell(u)\in\Sigma_\alpha$ such that $|u-u_\ell(u)|\le\frac{\alpha}{\sqrt{n}}$. Then we have by the Cauchy-Schwarz inequality and (5.10):

$$|\eta(u)-\eta(u_\ell(u))|^2\le\frac{1}{n}\sum_{i=1}^{n}|a_i(u_\ell(u))-a_i(u)|^2\sum_{i=1}^{n}\varepsilon_i^2\le\frac{\kappa_2^2\alpha^2}{n}\sum_{i=1}^{n}\varepsilon_i^2.$$

However, the probability

$$\boldsymbol{P}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2>4\sigma^2\right)$$

is certainly less than n^{-1}. Thus

$$\boldsymbol{P}\left(\sup_{u\in B_\alpha}|\eta(u)-\eta(u_\ell(u))|>2\kappa_2\sigma\alpha\right)\le\frac{1}{n}. \tag{5.12}$$

Then in an obvious way we have from (5.11) and (5.12) using
$nN(\Sigma_\alpha)\le n^{1+d/2}$

$$\begin{aligned}
&\boldsymbol{P}\left(\sup_{u\in B_\alpha}|\eta(u)|>\kappa_2\sigma\alpha(2+\sqrt{(2+d)\log n})\right)\\
&\quad\le\ \boldsymbol{P}\left(\sup_{u_\ell\in\Sigma_\alpha}|\eta(u_\ell)|>\kappa_2\sigma\alpha\sqrt{\log(2+d)\log n}\right)\\
&\qquad\quad+\boldsymbol{P}\left(\sup_{u\in B_\alpha}|\eta(u)-\eta(u_\ell(u))|>2\kappa_2\sigma\alpha\right)\le\frac{2}{n}.
\end{aligned}$$

Since the sum $n^{-1/2}\sum_{i=1}^n a_i(u^*)\varepsilon_i$ is a Gaussian random variable and by (5.9)

$$\boldsymbol{E}\left(n^{-1/2}\sum_{i=1}^n a_i(u^*)\varepsilon_i\right)^2=n^{-1}\sigma^2\sum_{i=1}^n a_i^2(u^*)\le\kappa_1^2\sigma^2$$

we also get

$$\boldsymbol{P}\left(\left|n^{-1/2}\sum_{i=1}^n a_i(u^*)\varepsilon_i\right|>z\,\sigma\,\kappa_1\right)\le\exp\left(-z^2/2\right)$$

and the lemma follows. ■

The results of Lemmas 5 and 6 lead to the following bound for the stochastic term ζ_u:

$$\begin{aligned}
&\boldsymbol{P}\left(|\zeta_{u^*}|>z\frac{\sigma C_V(1+C_N)}{h\sqrt{n}}+\right)\le\exp\left(-z^2/2\right)\\
&\boldsymbol{P}\left(\sup_{|u-u^*|\le\alpha}|\zeta_u|>z\frac{\sigma C_V(1+C_N)}{h\sqrt{n}}+\sigma\kappa_2\alpha\left(2+\sqrt{(2+d)\log n}\right)\right)\\
&\qquad\le\exp\left(-z^2/2\right)+\frac{2}{n}
\end{aligned}$$

with

$$\kappa_2=\frac{8\sqrt{2}\,C_w\,C_V\,C_N(2C_V+1)}{h\sqrt{n}}.$$

When summing up this result and that of Lemma 4, we get from (5.7) the bounds for $\widehat{\beta}$ for two different cases: the first one corresponds to the initial estimate

$\widehat{\beta} = \widehat{\beta}_0$ with $u = u^*$, $\rho = 1$ and $A_{u^*} = I$, and the second one is for the general situation,

$$\boldsymbol{P}\left(|\widehat{\beta} - \beta^*)| \geq 0.5h\, C_V\, C_g + \frac{\sigma C_V(C_N+1)}{h\sqrt{n}} z\right) \leq \exp(-z^2/2)$$

$$\boldsymbol{P}\left(\sup_{|\theta-\theta^*|\leq\gamma} |\Pi_{\rho,\theta^*}(\widehat{\beta} - \beta^*)| \geq 2h\rho^2\, C_V\, C_g + \frac{\sigma C_V(C_N+1)}{h\sqrt{n}} z \right.$$
$$\left. + \frac{8\sigma\alpha\, C_w\, C_V\, C_N(2C_V+1)\left(1+\sqrt{(2+d)}\right)\sqrt{2\log n}}{h\sqrt{n}}\right)$$
$$\leq \exp(-z^2/2) + \frac{2}{n}.$$

Now the both statements of the proposition follows from the following simple

Lemma 7 *Let $\theta = \beta/|\beta|$ and $\theta^* = \beta^*/|\beta^*|$, where β and $\beta^* \in I\!\!R^d$. 1) If $|\beta - \beta^*| \leq \delta \leq 1/4$. Then*

$$|\theta - \theta^*| \leq 2\frac{\delta}{|\beta^*|}.$$

2) If $|\Pi_{\rho,\theta^*}(\beta - \beta^*)| \leq \delta$, $\rho \leq 1/2$ and $\delta(1+\rho^{-1}) \leq |\beta^*|/2$, then

$$|\theta - \theta^*| \leq \frac{\delta}{|\beta^*|}\left(1 + \frac{5\delta}{\rho|\beta^*|}\right).$$

Proof: To show 1) we write

$$|\theta - \theta^*| = 2\sin\frac{(\theta,\theta^*)}{2} \leq \sqrt{2}\sin(\beta,\beta^*) \tag{5.13}$$

(we have used here that $\sin\frac{\alpha}{2} \leq \frac{\sin\alpha}{\sqrt{2}}$ for $0 \leq \alpha \leq \pi/2$). Furthermore,

$$\sin(\beta,\beta^*) \leq \frac{|\beta - \beta^*|}{\min(|\beta|, |\beta^*|)} \leq \frac{\delta}{|\beta^* - \delta|} \leq \frac{4\delta}{3|\beta^*|}.$$

Now we conclude 1) from (5.13).

To prove 2) we remark first that $|\Pi_{\rho,\theta^*}(\beta - \beta^*)| \leq \delta$ implies that

$$|\beta^* - \beta| \leq \delta(1+\rho^{-1}) \leq |\beta^*|/2. \tag{5.14}$$

Note now that $\theta^*(\theta^*)^T x$ is the projection of x on θ^*, thus

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

On the other hand, $|(I-\theta^*(\theta^*)^T)(\beta-\beta^*)| = |(I-\theta^*(\theta^*)^T)\beta| = |\beta|\sin(\theta,\theta^*)$. This implies that

$$\sin(\theta,\theta^*) \le \frac{\delta}{|\beta|}.$$

Then the orthogonal decomposition

$$\theta^* - \theta = (I-\theta^*(\theta^*)^T)\theta + \theta^*((\theta^*)^T\theta - 1),$$

along with $|(I-\theta^*(\theta^*)^T)\theta^*|^2 = \sin^2(\theta,\theta^*)$ and $|\theta^T\theta^*| = \cos(\theta,\theta^*)$ yields

$$\begin{aligned} |\theta^* - \theta|^2 &= \sin^2(\theta,\theta^*) + (1-\cos(\theta,\theta^*))^2 \\ &\le \sin^2(\theta,\theta^*) + 4\sin^4\frac{(\theta,\theta^*)}{2} \\ &\le \frac{\delta^2}{|\beta|^2}\left(1+\frac{\delta^2}{|\beta|^2}\right). \end{aligned}$$

We have now

$$|\theta^* - \theta| \le \frac{\delta}{|\beta|}\left(1+\frac{\delta^2}{2|\beta|^2}\right) \le \frac{\delta}{|\beta|}\left(1+2\frac{\delta^2}{|\beta^*|^2}\right). \tag{5.15}$$

However, by (5.14) we obtain

$$\left|\frac{\delta}{|\beta|} - \frac{\delta}{|\beta^*|}\right| \le \frac{\delta|\beta-\beta^*|}{|\beta|\,|\beta^*|} \le \frac{2\delta^2(1+\rho^{-1})}{|\beta^*|^2}.$$

When substituting this into (5.15), we conclude

$$|\theta^* - \theta| \le \frac{\delta}{|\beta^*|}\left(1+2\frac{\delta^2}{|\beta^*|^2} + \frac{2\delta(1+\rho^{-1})}{|\beta^*|}\right) \le \frac{\delta}{|\beta^*|}\left(1+\frac{5\delta}{\rho|\beta^*|}\right),$$

as required. ∎

References

Bonneu, M., Delecroix, M., and Hristache, M. (1997). Semiparametric estimation of generalized linear models. submitted.

Delecroix, M., Härdle, W., and Hristache, M. (1997). Efficient estimation in single-index regression. submitted.

Delecroix, M. and Hristache, M. (1998). M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique. *The Bulletin of the Belgian Mathematical Society.* to appear.

FRIEDMAN, F. H. AND STUETZLE, W. (1981). A projection pursuit regression. *Journal of the American Statistical Association*, 79:599–608.

HALL, P. (1989). On projection pursuit regression. *Annals of Statistics*, 17:573–588.

HÄRDLE, W., HALL, P., AND ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, 21:157–178.

HÄRDLE, W. AND HOROWITZ, J. (1997). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91:1632–1640.

HÄRDLE, W. AND TSYBAKOV, A. (1993). How sensitive are average derivatives. *Journal of Econometrics*, 58:31–48.

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

KLEIN, R. AND SPADY, R. (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. *Annals of Statistics*, 17:1009–1052.

MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

POWELL, J., STOCK, J., AND STOKER, T. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57:1403–1430.

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