The Annals of Statistics
2001, Vol. 29, No. 6, 1537–1566

STRUCTURE ADAPTIVE APPROACH FOR DIMENSION REDUCTION[1]

BY MARIAN HRISTACHE, ANATOLI JUDITSKY, JÖRG POLZEHL AND VLADIMIR SPOKOINY

ENSAI, LMC Domaine Universitaire B.P. 53, Weierstrass Institute for Applied Analysis and Stochastics, Weierstrass Institute for Applied Analysis and Stochastics

We propose a new method of effective dimension reduction for a multi-index model which is based on iterative improvement of the family of average derivative estimates. The procedure is computationally straightforward and does not require any prior information about the structure of the underlying model. We show that in the case when the effective dimension m of the index space does not exceed 3, this space can be estimated with the rate $n^{-1/2}$ under rather mild assumptions on the model.

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 the Y_i are 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 $f\colon \mathbb{R}^d \to \mathbb{R}$. We assume that $f(x)$ has the specific structure

$$f(x) = g_0(Tx). \tag{1.2}$$

Here $g_0(\cdot)$ is an unknown m-dimensional *link* function and T is a linear orthonormal mapping from the high-dimensional space $\mathbb{R}^d$ onto the space $\mathbb{R}^m$ with an essentially smaller dimension m, satisfying the condition $TT^\top = I_m$, where $T^\top$ stands for the transpose of T. In the statistical literature, relations as in (1.1) and (1.2) are referred to as *multi-index regression* models. Model (1.2) is a rather general expression of the hypothesis that all the information about $f(x)$ is "concentrated" in a low-dimensional projection Tx. If we adopt such a model, our intention can be both to find the *effective dimension* m and to describe the *index space* $\mathscr{I} = \operatorname{Im} T^\top$ which is also referred to as the *effective dimension space* or the space of *effective dimension reduction* in Li (1991, 1992, 2000) and Cook (1998). In the present paper we propose an algorithm to estimate the index space when the effective dimension m is known a priori. Some extensions are discussed in Section 6.

Received March 2000; revised May 2001.

[1] Support in part by the Deutsche Forschungsgemeinschaft, SFB 373 "Simulation and Quantification of Economic Processes" at Humboldt University.

AMS 2000 *subject classifications*. Primary 62G05; secondary 62H40, 62G20.

Key words and phrases. Dimension-reduction, multi-index model, index space, average derivative estimation, structural adaptation.

Note first that the representation (1.2) is not unique. For instance, if U_m is an orthogonal transform in $\mathbb{R}^m$, then the function f can be rewritten in the form $f(x) = g_1(T_1 x)$ with $g_1(z) = g_0(U_m z)$ and $T_1 = U_m^{\top} T$. Nevertheless, the index space $\mathscr{I}$ is defined uniquely by (1.2) and it contains very important information about the model. As soon as the operator T which maps $\mathbb{R}^d$ onto $\mathbb{R}^m$ is fixed, the link function g_0 can be estimated in a nonparametric way.

Various methods for dimension reduction have been proposed in the literature. Classical theory of *principal component analysis* considers mostly the case of multiple *linear* regression. Brillinger (1983) extended the method to the so-called "generalized linear model" with normally distributed regressors. The underlying idea is to make some data transformation and then to proceed as if the model were linear. Under a similar assumption on the distribution of regressors, Li (1991) offered the so-called "sliced inverse regression" approach. A modification of this method (principal Hessian directions) is explored in Li (1992) and Cook (1998). Samarov (1993) discussed an approach relying on average derivative estimation of some linear functionals of the gradient of the regression function f. However, the conditions for this method to work appear to be quite restrictive in application to real data. The main problem here is that, for large d, the data in the high-dimensional space $\mathbb{R}^d$ is very sparse (the so-called "curse of dimensionality" problem).

Our approach can be seen as an iterative improvement of the average derivative estimator and can be used under weak assumptions on the model. The proposed procedure can be regarded as an extension of the method developed in Hristache, Juditsky and Spokoiny [(2001); henceforth HJS01] for the single-index model to the multi-index situation. In the sequel the latter paper is referred to as HJS01.

The paper is organized as follows: in the next section we discuss the heuristics behind the proposed approach. Then in Section 3 the estimation procedure is presented. The performance of the method is tested for some simulated datasets in Section 4. The theoretic properties of procedure are discussed in Section 5. In particular, it is shown that the procedure leads to root-n consistent estimation of the index space if $m \le 3$. Section 6 briefly summarizes main results and discusses possible extensions and open problems. Finally, the proofs are collected in the Appendix.

2. Basic ideas. Since the gradient $F(X_i) = \nabla f(X_i)$ of the regression function f at every point X_i belongs to the index space $\mathscr{I}$, it seems quite natural to apply the principal component analysis for estimating this space: one can compute the matrix $\mathscr{M}^* = \frac{1}{n}\sum_{i=1}^{n} F(X_i) F^{\top}(X_i)$ and then use the eigenvalue decomposition of $\mathscr{M}^*$, $\mathscr{M}^* = O_d^{\top} \Lambda O_d$. Here O_d is an orthonormal matrix and Λ is a diagonal matrix with decreasing eigenvalues. These matrices deliver important information about model (1.2): the first m columns of O_d (i.e., the first m eigenvectors of $\mathscr{M}^*$) provide an orthonormal basis of the index space $\mathscr{I}$; the corresponding eigenvalues show how much the function f varies in each direction. In particular, the first eigenvector of $\mathscr{M}^*$ is the direction in which f varies most [cf. Samarov (1993)]. This leads to the natural idea to first estimate $\mathscr{M}^*$ from the data $Y_1, \dots, Y_n$ and then to recover the index space $\mathscr{I}$ using this estimate [see, e.g., Fan and Gijbels (1996), page 295 and references therein, or King (1997)]. Note that the matrix $\mathscr{M}^*$ is a quadratic functional of the gradient of the regression function f. There exists some literature on estimation of such functionals in the framework of nonparametric regression. Various estimation algorithms and results on their optimality can be found in Ibragimov, Nemirovskii and Khasmiskii (1986), Donoho and Nussbaum (1990), Fan (1991). The estimators in Samarov (1993) and Doksum and Samarov (1995) are based on kernel estimators of the regression function f. Huang and Fan (1998) applied the local polynomial fit. The procedure from Ibragimov, Nemirovskii and Khasmiskii (1986) is based on the Fourier expansion of the gradient F of the function f. Let us see how this latter idea applies to our problem.

Suppose that we are given a collection $\{\psi_\ell, \ell = 1, \dots, L\}$ of functions $\psi_\ell \colon \mathbb{R}^d \to \mathbb{R}$ which satisfy

$$\sum_{i=1}^n \psi_\ell(X_i)\psi_{\ell'}(X_i) = \delta_{\ell\ell'},$$

where $\delta_{\ell\ell} = 1$ and $\delta_{\ell\ell'} = 0$ for $\ell \neq \ell'$. Now, let β_ℓ^*,

$$\beta_\ell^* = \frac{1}{n}\sum_{i=1}^n F(X_i)\psi_\ell(X_i), \tag{2.1}$$

be the ℓth Fourier coefficient of F with respect to the basis system $\{\psi_\ell\}$. Note that each d-vector β_ℓ^* is a linear functional of the gradient and hence belongs to $\mathscr{I}$. Thus if the dimension of the space spanned by $\beta_1^*, \dots, \beta_L^*$ equals m, this set of vectors completely characterizes the index space $\mathscr{I}$, and one can identify the space $\mathscr{I}$ by looking for the first m principal components of the set $\beta_1^*, \dots, \beta_L^*$.

In order to estimate $\mathscr{M}^*$, one can first construct an estimate $\hat\beta_\ell$ of each Fourier coefficient β_ℓ^*, for example,

$$\hat\beta_\ell = \frac{1}{n}\sum_{i=1}^n \widehat{F}(X_i)\psi_\ell(X_i) \tag{2.2}$$

on the basis of a pilot estimate $\widehat{F}$ of the gradient, and then compose the estimate

$$\widehat{\mathscr{M}_L} = \sum_{\ell=1}^L \hat\beta_\ell \hat\beta_\ell^{\mathsf{T}}$$

of $\mathscr{M}^*$. Note that in order to ensure $\widehat{\mathscr{M}_L}$ to be a consistent estimate of the matrix $\mathscr{M}^*$ the number L of basis functions ψ_ℓ should be taken growing with n. Otherwise $\widehat{\mathscr{M}_L}$ estimates the matrix $\mathscr{M}_L^*$ with

$$\mathscr{M}_L^* = \sum_{\ell=1}^L \beta_\ell^* \beta_\ell^{*\mathsf{T}}.$$

On the other hand, recall that it is the index space $\mathscr{I}$ we are interested in and not the estimation of $\mathscr{M}^*$. It would be sufficient for our purposes to point out a fixed (possibly small) number of "test functions" ψ_ℓ such that $\operatorname{rank}(\mathscr{M}_L^*) = m$ and the value $\|M^* - M_L^*\|$ (that is, the maximal eigenvalue of $M^* - M_L^*$) is not too large. The choice of a proper set of test functions $\psi_\ell, \ell = 1, \dots, L$ is discussed in more details in Section 3.4.

2.1. *Equivalent representation.* As we have already noticed, the model representation (1.2) is not unique. It is more convenient for our purposes to work with another one, which is distinctly defined by the set of test functions $\psi_\ell, \ell = 1, \dots, L$ and the regression function f.

Let us denote by $\mathscr{B}^*$ the $d \times L$ matrix with the columns $\beta_\ell^*, \ell = 1, \dots, L$, where the vectors β_ℓ^* are as in (2.1). Obviously, each vector β_ℓ^* belongs to $\mathscr{I}$ and hence $\operatorname{rank}(\mathscr{B}^*) \le m$. We additionally suppose that $\operatorname{rank}(\mathscr{B}^*) = m$ which means that this matrix completely describes the index space $\mathscr{I}$.

Let $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d$ be the ordered set of eigenvalues of the symmetric $d \times d$-matrix $\mathscr{M}_L^* = \mathscr{B}^*(\mathscr{B}^*)^\top$. Since $\operatorname{rank}(\mathscr{M}_L^*) = m$, only the first m of them are positive and the remaining ones are equal to zero. Without loss of generality we assume that all eigenvalues are different; that is, $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ which ensures that the corresponding eigenvectors of unit length $e_1, \dots, e_m$ are uniquely defined (up to a sign). These vectors belong to the index space $\mathscr{I}$ and can be used as a natural basis in it. We also denote $\theta_k = \sqrt{\lambda_k} e_k$, $k = 1, \dots, m$. Since $\lambda_k = 0$ for $k > m$, it also holds that $\theta_k = 0$ for those k.

We now represent the model (1.1), (1.2) in the form

$$f(x) = g\bigl(\theta_1^\top x, \dots, \theta_m^\top x\bigr), \tag{2.3}$$

where the new link function g is uniquely defined as soon as the vectors $\theta_1, \dots, \theta_m$ are fixed. Usually a similar representation with vectors $e_k = \theta_k / |\theta_k|$ in place of θ_k is used:

$$f(x) = g_1\bigl(e_1^\top x, \dots, e_m^\top x\bigr). \tag{2.4}$$

However, the value λ_k characterizes the variability of the function f in the direction e_k. Thus the function g_1 in (2.4) inherits the inhomogeneity of f in different directions. The benefit of using (2.3) is that the corresponding link function g is homogeneous w.r.t. its variables.

Let $\mathscr{R}^*$ be a $m \times d$-matrix such that its transpose $(\mathscr{R}^*)^\top = (\theta_1, \dots, \theta_m)$ has vectors $\theta_1, \dots, \theta_m$ as columns. Then (2.3) can be rewritten as $f(x) = g(\mathscr{R}^* x)$. The matrix $\mathscr{R}^*$ maps $\mathbb{R}^d$ onto $\mathbb{R}^m$ and determines the required effective dimension space. In what follows we refer to $\mathscr{R}^*$ as the *effective dimension reduction matrix*, or simply the *e.d.r.*

The following well-known matrix result offers an explicit representation of the matrix $\mathscr{R}^*$ via the eigenvalue decomposition of the symmetric $L \times L$-matrix $(\mathscr{B}^*)^\top \mathscr{B}^*$.

LEMMA 2.1. *Let* $(\mathscr{B}^*)^\top \mathscr{B}^* = O \Lambda_L O^\top$ *be the eigenvalue decomposition of* $(\mathscr{B}^*)^\top \mathscr{B}^*$ *where* O *is an orthogonal* $L \times L$-*matrix and* Λ_L *is a diagonal matrix* *with nonincreasing eigenvalues* $\lambda'_1 \ge \lambda'_2 \ge \cdots \ge \lambda'_L$. *Let also* O_m *be the block of the first* m *columns of* O. *Then* $\lambda'_k = \lambda_k$ *for* $k \le d$ *and*

$$\mathscr{R}^* = (\mathscr{B}^* O_m)^{\mathsf{T}}. \tag{2.5}$$

Due to this lemma, the model (2.3) can be now rewritten in the form

$$f(x) = g(\mathscr{R}^* x) = g\big((\mathscr{B}^* O_m)^{\mathsf{T}} x\big) \tag{2.6}$$

which is used in the sequel.

2.2. *Gradient estimation.* Next we discuss the problem of estimating each linear functional β^*_ℓ using a nonparametric estimate $\widehat{F}$ of the gradient F; see (2.2). A standard way to estimate both $f(X_i)$ and $F(X_i)$ is to apply the local linear least squares approach,

$$\begin{pmatrix} \hat{f}(X_i) \\ \widehat{F}(X_i) \end{pmatrix} = \operatorname*{arg\,inf}_{c\in\mathbb{R},\, b\in\mathbb{R}^d} \sum_{j=1}^{n} \big[Y_j - c - b^{\mathsf{T}}(X_j - X_i)\big]^2 \, K\bigg(\frac{|X_j - X_i|^2}{h^2}\bigg), \tag{2.7}$$

where $|\cdot|$ means Euclidean norm in $\mathbb{R}^d$ and *a kernel* $K(\cdot)$ is positive and supported on $[0, 1]$, so that the weights of all points X_j outside a spherical neighborhood $U_h(X_i)$ of diameter h around X_i vanish. The solution to this quadratic optimization problem can be represented as

$$\begin{pmatrix} \hat{f}(X_i) \\ \widehat{F}(X_i) \end{pmatrix} = \Bigg\{ \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\mathsf{T}} K\bigg(\frac{|X_{ij}|^2}{h^2}\bigg) \Bigg\}^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\bigg(\frac{|X_{ij}|^2}{h^2}\bigg),$$

where $X_{ij} = X_j - X_i$. As many other nonparametric estimates, the estimate (2.7) suffers from the data sparseness for large d. This phenomenon is often referred to as *curse of dimensionality*. Indeed, one has to select the bandwidth h in a way to provide at least $d + 1$ design points in every (or almost every) spherical neighborhood $U_h(X_i)$. For the case of a random design with a positive density, this implies that a bandwidth h of order $n^{-1/d}$ or even larger should be taken. For large d this leads to a very poor rate $n^{-1/d}$ in estimation of F, and the same applies to the estimation of the vectors β^*_ℓ (see Proposition 5.1 below).

At the same time, suppose for a moment that we know the mapping $T\colon \mathbb{R}^d \to \mathbb{R}^m$. Then we could use this information for estimating the m-dimensional link function g_0 and its gradient ∇g_0. This also provides an estimate of the gradient $F(x) = T^{\mathsf{T}} \nabla g_0(Tx)$ of much better accuracy, which corresponds to an m-dimensional nonparametric problem on the "true" index space, instead of the original d-dimensional nonparametric estimate $\widehat{F}(x)$. More specifically, a function $f(x)$ of the form (2.6) remains constant when x varies in any direction orthogonal to the m-dimensional subspace $\mathscr{I}$. The above considerations leads to another estimate,

$$\begin{pmatrix} \hat{f}(X_i) \\ \widehat{F}(X_i) \end{pmatrix} = \arg\inf_{c\in\mathbb{R},\, b\in\mathbb{R}^d} \sum_{j=1}^{n} \bigl[Y_j - c - b^{\mathsf{T}}(X_j - X_i)\bigr]^2 K\biggl(\frac{|T(X_j - X_i)|^2}{h^2}\biggr)$$

$$= \Biggl\{\sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\mathsf{T}} K\biggl(\frac{|TX_{ij}|^2}{h^2}\biggr)\Biggr\}^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\biggl(\frac{|TX_{ij}|^2}{h^2}\biggr).$$

The latter estimate of $F(X_i)$ is based on averaging over a narrow cylinder $\{x\colon |T(x - X_i)| \le h\}$, centered at X_i, which spans $\mathscr{I}^{\perp}$. For this estimate one can apply an essentially smaller bandwidth h and still have enough design points in every such neighborhood. On the other hand, the smaller bandwidth would decrease drastically the bias of estimation. Unfortunately this "ideal" estimate cannot be implemented in practice since it requires explicit knowledge of the target index space $\mathscr{I}$. A natural idea is to substitute the mapping T by its pilot estimate. This leads to the following *structural adaptation* approach. We proceed iteratively starting with the estimates $\hat{\beta}_\ell = \frac{1}{n}\sum_{i=1}^{n} \widehat{F}(X_i)\psi_\ell(X_i)$, $\ell = 1, \dots, L$ based on the fully nonparametric gradient estimate $\widehat{F}$ with some $h = h_1$; see (2.7). Although this estimate is very rough, it contains some information about the structure of the model function f and, in particular, about the mapping T: all vectors $\hat{\beta}_\ell$ up to the estimation error, belong to the index space $\mathscr{I}$. This information can be used for producing another, more careful estimate of the gradient function and hence, of the vectors β^*_ℓ. More precisely, let $\widehat{\mathscr{B}}_1$ be the matrix composed from the vectors $\hat{\beta}_\ell$, $\ell = 1, \dots, L$. We define the gradient estimate $\widehat{F}_2(X_i)$ at X_i by a local linear fit using the elliptic neighborhood $\{x\colon |S_2(x - X_i)| \le h_2\}$, with $S_2 = (I + \rho_2^{-2}\widehat{\mathscr{B}}_1\widehat{\mathscr{B}}_1^{\mathsf{T}})^{1/2}$ for some $\rho_2 < 1$ 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 all the directions $\hat{\beta}_\ell$ (since $\rho_2 < 1$) and stretch them in all the orthogonal directions (since $h_2 > h_1$),

$$\begin{pmatrix} \hat{f}_2(X_i) \\ \widehat{F}_2(X_i) \end{pmatrix} = \arg\inf_{c\in\mathbb{R},\, b\in\mathbb{R}^d} \sum_{j=1}^{n} \bigl[Y_j - c - b^{\mathsf{T}}(X_j - X_i)\bigr]^2 K\biggl(\frac{|S_2(X_j - X_i)|^2}{h_2^2}\biggr)$$

$$= \Biggl\{\sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\mathsf{T}} K\biggl(\frac{|S_2X_{ij}|^2}{h_2^2}\biggr)\Biggr\}^{-1}$$

$$\times \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\biggl(\frac{|S_2X_{ij}|^2}{h_2^2}\biggr).$$

This leads to the estimates $\hat{\beta}_{2,\ell} = \frac{1}{n}\sum_{i=1}^{n} \widehat{F}_2(X_i)\psi_\ell(X_i)$ of β^*_ℓ producing the matrix $\widehat{\mathscr{B}}_2$. We continue this way each time compressing the averaging windows in the direction of the current estimate $\widehat{\mathscr{B}}_k$ and expanding them in orthogonal directions.

The results presented below show that this procedure allows estimating the index space $\mathscr{I}$ at the rate $n^{-1/2}$ provided that $m < 4$.

3. Estimation procedure. We now present the description of the method. The whole estimation procedure is carried out in two basic steps: estimation of the vectors β_ℓ^* and estimation of the e.d.r. matrix $\mathscr{R}^*$. Below we discuss each step separately.

3.1. *Estimation of* β_ℓ^**'s.* The procedure involves input parameters $h_1 < h_{\max}$ and $\rho_{\min} < \rho_1$, so that ρ decreases geometrically from ρ_1 to $\rho_{\min}$ by the factor $a_\rho < 1$ and h increases geometrically from h_1 to $h_{\max}$ by the factor $a_h > 1$ during iterations. The choice of these parameters as well as the set of basis functions $\{\psi_\ell\}$ will be discussed in the next section. The algorithm reads as follows:

1. Initialization: specify parameters $\rho_1, \rho_{\min}, a_\rho, h_1, h_{\max}, a_h$ and the set of functions $\{\psi_\ell\}$; set $k = 1, \widehat{\mathscr{B}}_0 = 0$;
2. Compute $S_k = (I + \rho_k^{-2}\widehat{\mathscr{B}}_{k-1}\widehat{\mathscr{B}}_{k-1}^\top)^{1/2}$;
3. For every $i = 1, \dots, n$, compute $\widehat{F}_k(X_i)$ from the expression:
$$\begin{pmatrix} \hat{f}_k(X_i) \\ \widehat{F}_k(X_i) \end{pmatrix} = V_k^{-1}(X_i) \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\left(\frac{|S_k X_{ij}|^2}{h_k^2}\right),$$
where $X_{ij} = X_j - X_i$ and $V_k(X_i) = \sum_{j=1}^n \binom{1}{X_{ij}}\binom{1}{X_{ij}}^\top K\Big(\frac{|S_k X_{ij}|^2}{h_k^2}\Big)$;
4. Compute the vectors $\hat{\beta}_{k,\ell} = \frac{1}{n}\sum_{i=1}^n \widehat{F}_k(X_i)\psi_\ell(X_i)$, $\ell = 1, \dots, L$ and compose the matrix $\widehat{\mathscr{B}}_k$ with columns $\hat{\beta}_{k,1}, \dots, \hat{\beta}_{k,L}$;
5. Set $h_{k+1} = a_h h_k, \rho_{k+1} = a_\rho \rho_k$. If $\rho_{k+1} \geq \rho_{\min}$, then set $k = k + 1$ and continue with Step 2; otherwise terminate.

By $k(n)$ we denote the total number of iterations. The estimates $\hat{\beta}_{k(n),\ell}$ from the last iteration are used as the final estimates of β_ℓ^*.

3.2. *Modified estimator.* In the above algorithm, at each step, we use a linear combination of the estimated gradient vectors $\widehat{F}(X_i)$ as the estimate of the vector β_ℓ^*. To guarantee some useful properties of this procedure, the estimates $\widehat{F}(X_i)$ should be well defined, which in turn requires some local regularity of the design in the corresponding neighborhood of the point X_i; see Assumption 5 in Section 5. If such a condition is not satisfied even at a few points, then the corresponding gradient estimates would have a very large standard deviation which may deteriorate the quality of the index estimates $\hat{\beta}_\ell$. We can avoid this problem by weighting each summand in the expression for $\hat{\beta}_{k,\ell}$ with some coefficients which express the degree of local regularity of the design.

Define $\bar{w}$ as the square root of the minimal eigenvalue of the matrix $\overline{\mathscr{V}}$ with

$$\overline{\mathscr{V}} = \frac{1}{\boldsymbol{E}K(\zeta^\top\zeta)}\boldsymbol{E}\begin{pmatrix} 1 \\ \zeta \end{pmatrix}\begin{pmatrix} 1 \\ \zeta \end{pmatrix}^\top K(\zeta^T\zeta),$$

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

Let also C_w be a positive number. Steps 2–4 of the above algorithm are modified as follows:

2′. Compute $\widehat{\mathscr{M}}_k = \widehat{\mathscr{B}}_{k-1}\widehat{\mathscr{B}}_{k-1}^{\mathsf{T}}$. If $\|\widehat{\mathscr{M}}_k\| > 1$, then normalize it by its maximal eigenvalue: $\widehat{\mathscr{M}}_k := \widehat{\mathscr{M}}_k / \|\widehat{\mathscr{M}}_k\|$; Set $S_k = (I + \rho_k^{-2}\widehat{\mathscr{M}}_k)^{1/2}$;

3′. For every $i = 1, \ldots, n$, compute the matrix $\widehat{\mathscr{V}}_k(X_i)$ with

$$\widehat{\mathscr{V}}_k(X_i) = \frac{1}{\sum_{j=1}^n K(W_{ij,k}^{\mathsf{T}} W_{ij,k})} \sum_{j=1}^n \begin{pmatrix} 1 \\ W_{ij,k} \end{pmatrix} \begin{pmatrix} 1 \\ W_{ij,k} \end{pmatrix}^{\mathsf{T}} K(W_{ij,k}^{\mathsf{T}} W_{ij,k}),$$

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 $\widehat{\mathscr{V}}_k(X_i)$: $w_i^2 = \lambda_{\min}(\widehat{\mathscr{V}}_k(X_i))$; If the condition

$$n^{-1}(w_1 + \cdots + w_n) \geq C_w \bar{w}$$

is fulfilled then compute

$$\begin{pmatrix} \hat{f}_k(X_i) \\ \widehat{F}_k(X_i) \end{pmatrix} = V_k^{-1}(X_i) \sum_{j=1}^n Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\left(\frac{|S_k X_{ij}|^2}{h_k^2}\right),$$

otherwise increase h_k by the factor a_h, that is, $h_k := a_h h_k$. If $h_k > h_{\max}$, then terminate, otherwise repeat this step;

4′. For every $\ell = 1, \ldots, L$, compute the vector $\hat{\beta}_{k,\ell}$

$$\hat{\beta}_{k,\ell} = \left(\sum_{i=1}^n w_i\right)^{-1} \sum_{i=1}^n \widehat{F}_k(X_i)\psi_\ell(X_i) w_i$$

with the previously obtained w_i's. Compose the matrix $\widehat{\mathscr{B}}_k$ with columns $\hat{\beta}_{k,\ell}$, $\ell = 1, \ldots, L$.

3.3. *Computing the effective dimension reduction matrix.* Let $\widehat{\mathscr{B}}$ be an estimate of the matrix $\mathscr{B}^*$ obtained by the previously described iterative procedure. We will see (Theorem 5.3) that this matrix estimates the target matrix $\mathscr{B}^*$ with a reasonable accuracy but it is typically of the rank d and hence, it does not provide any dimension reduction. We estimate the effective dimension reduction matrix $\mathscr{R}^*$ using the singular value decomposition of $\widehat{\mathscr{B}}$ in place of $\mathscr{B}^*$; compare (2.5). Namely, the product $\widehat{\mathscr{B}}^{\mathsf{T}}\widehat{\mathscr{B}}$, being symmetric and nonnegative, can be represented in the form $\widehat{\mathscr{B}}^{\mathsf{T}}\widehat{\mathscr{B}} = \widehat{O}\widehat{\Lambda}\widehat{O}^{\mathsf{T}}$ with the orthogonal $L \times L$-matrix $\widehat{O}$ and the diagonal matrix $\widehat{\Lambda}$: $\widehat{\Lambda} = \operatorname{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_L\}$ with nonincreasing eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_L \geq 0$ (the squared singular values of $\widehat{\mathscr{B}}$). The estimate $\widehat{\mathscr{R}}_m$ of the true e.d.r. matrix $\mathscr{R}^*$ from (2.5) is defined by

$$\widehat{\mathscr{R}}_m = (\widehat{\mathscr{B}}\widehat{O}_m)^{\mathsf{T}} \tag{3.1}$$

where $\widehat{O}_m$ is the submatrix of $\widehat{O}$ composed of its first m columns.

3.4. *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. Some related discussion about this choice can be found in HJS01. The general approach is to provide that at every iteration k there exist enough design points in every or almost every local ellipsoidal neighborhoods $E_k(X_i) = \{x\colon |S_k(x - X_i)| \le h_k\}$.

Note also that assuming the structure of the matrix $\widehat{\mathscr{B}}_{k-1}\widehat{\mathscr{B}}_{k-1}^{\top}$ to follow the structure of the target matrix $\mathscr{M}^*$, 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 directions of the m-dimensional index space $\mathscr{I}$. Therefore, the Lebesgue measure of every such neighborhood is changed each time by the factor $a_h^d a_\rho^m$. This leads to the constraint $a_h^d a_\rho^m > 1$; compare Assumption 4 in Section 5 below. Under the assumption of a random design with a positive density, this would result in an increase of the mean number of design points inside each $E_k(X_i)$.

The main constraint on the set $\{\psi_\ell\}$ is that the matrix $\mathscr{B}^*$ is of the same rank as T and that the function g from the equivalent representation (2.6) is sufficiently smooth; see Assumption 3 below. It can be easily shown that the "ideal" choice of the set $\{\psi_\ell\}$ can be obtained by orthogonalization of the components $F_j = \partial f/\partial x_j$, $j = 1, \dots, d$ of the gradient F. This "ideal" collection of functions ψ_ℓ would contain only m elements. Of course, this choice cannot be realized since it involves the unknown regression function f.

Note next that the functions (vectors) $\psi_1, \dots, \psi_L$ form an orthonormal system in $\mathbb{R}^n$ and β_ℓ^* is the scalar product of the gradient F and the basis function ψ_ℓ. The sum

$$F_L = \sum_{\ell=1}^{L} \beta_\ell^* \psi_\ell$$

is the projection of the gradient F on the linear subspace in $\mathbb{R}^n$ spanned by $\{\psi_\ell\}$. One can easily check that $\mathscr{M}_L^* = \sum_{i=1}^n F_L(X_i)F_L(X_i)^{\top}$. Thus to prevent the loss of information due to the substitution of $\mathscr{M}^*$ for $\mathscr{M}_L^*$, the set $\{\psi_\ell\}$ should be selected rich enough. Our proposals is to define $\{\psi_\ell\}$ by orthogonalizing the set of all polynomials $x_{\ell_1} \cdots x_{\ell_q}$ of the coordinate functions for some $q \ge 1$ and all $1 \le \ell_1 \le \cdots \le \ell_q$. The procedure from HJS01 corresponds to the family of all linear coordinate functions (i.e., $q = 1$). The simulation results are overall in favor of a larger q, for example, $q = 2$.

A suitable alternative, especially for large d, is a basis system constructed by orthogonalizing a fully nonparametric estimate of the gradient.

4. Implementation and simulated results.

In this section we illustrate the performance of the proposed algorithm on some simulated examples. In our simulation study we apply the modified procedure with the following parameter setting:

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

Since $e^{d/2(4\vee d)-(m/6)} > 1$ for all $m \le 3$ and $d > m$, the condition $a_h^d a_\rho^m > 1$ is fulfilled; see Section 3.4 or Assumption 4 in Section 5 below.

We also set $C_w = 4^{-1}$. In case of high dimensionality, that is, $d > 20$, a smaller value of C_w was necessary to guarantee the existence of valid bandwidths h_k. The basis system $\{\psi_\ell\}$ is obtained by orthogonalization of the set of functions $\{1, x_j, x_j x_k, j, k = 1, \dots, d\}$. This setting leads to the number of iterations $k(n) \approx \frac{\log(\rho_1/\rho_{\min})}{\log a_\rho} = 2\log n$.

The performance of the method is illustrated by means of the following examples. We consider the model $Y_i = g(X_i^\mathsf{T}\theta_1, \dots, X_i^\mathsf{T}\theta_m)$ for m between 1 and 3. The design $X_1, \dots, X_n$ is modeled randomly with independent components so that every component of $(X_i + 1)/2$ follows $B(1, \tau)$-distribution. The parameter τ controls the skewness of the beta-distribution with $\tau = 1$ corresponding to the uniform design. We also set:

$m = 1$: $g(u) = u\sin(\sqrt{5}u)$ and $\theta = (1, 2, 0, \dots, 0)^\mathsf{T}/\sqrt{5}$.
$m = 2$: $g(u_1, u_2) = (u_1^3 + u_2)(u_1 - u_2^3)$ and $\theta_1 = (1, 1, 0, \dots, 0)^\mathsf{T}/\sqrt{2}$, $\theta_2 = (1, -1, 0, \dots, 0)^\mathsf{T}/\sqrt{2}$.
$m = 3$: $g(u_1, u_2, u_3) = (u_1^3 + u_2)(u_1 - u_2^3) + u_3$ and $\theta_1 = (1, 1, 1, 0, \dots, 0)^\mathsf{T}/\sqrt{3}$, $\theta_2 = (1, -1, 0, \dots, 0)^\mathsf{T}/\sqrt{2}$, $\theta_3 = (1, 1, -2, \dots, 0)^\mathsf{T}/\sqrt{6}$.

The first situation corresponds essentially to Example 8.2 from Li (1992). The procedure utilizes the biweight kernel $K(x) = (1 - |x|^2)_+^2$. The quality of estimation is measured using the criterion $\|\mathscr{R}^*(I - \widehat{\mathscr{P}}_m)\|_2$ with $\|A\|_2^2 = \operatorname{tr} AA^\mathsf{T}$, where $\widehat{\mathscr{P}}_m$ is the projector on the estimated index space $\widehat{\mathscr{I}}$; see Section 5.2 for more details.

Our objective is to illustrate the following features of the procedure:

How the quality of estimation improves during iteration.
Dependence on the sample size n and the dimensionality d.
How the results depend on skewness of the design and the error variance σ^2.
Relative performance of the method.

For the latter, we compare the performance of our iterative procedure with sliced inverse regression II (SIR II), principal Hessian directions (PHD) [see, e.g., Li (1991, 1992, 2000)], and the estimate coming from the first step of our algorithm, which is actually a version of the usual average derivative estimator (ADE); compare King (1997). The parameters of all the competitors were selected to optimize the criterion at hand while our procedure was implemented with the default parameter choice. Note that for our examples SIR I, which is based on means over different slices, fails to recover the dimension reduction space. We do not report results for SIR and PHD for the third case ($m = 3$) since both methods only recover a two-dimensional subspace.

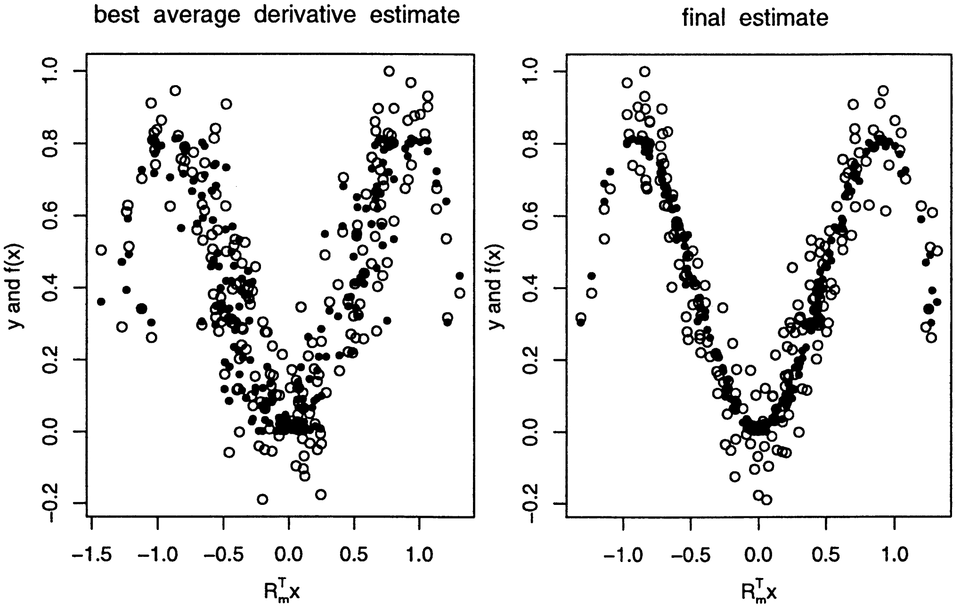

FIG. 1. *Best view for a one-step estimate* (*left*) *and view from the last iteration* (*right*) *for* $g(u) = u\sin(\sqrt{5}u)$; $m = 1, d = 10, n = 200$ *and* $\sigma = 0.1$. *Values of* y *and* $f(x)$ *are indicated by* ○ *and* •, *respectively.*

Figure 1 illustrates the quality of estimation of the index space for $m = 1$, $d = 10$, $n = 200$ and $\sigma = 0.1$, providing the view obtained by a one-step estimate with optimized bandwidth (left) and the view gained from our procedure (right). Simulation results for different dimensionality d and sample size n are given in Tables 1, 2 and 3.

All results show a considerable gain using the proposed iterative method. This gain increases drastically as the dimensionality d grows. The results from Table 2 for $d = 10$ and different σ-values clearly illustrate the bias-variance trade-off. For the first step estimate as well as for the "best" such estimate with the optimal bandwidth, the bias dominates and the quality of estimation only weakly depends on the noise variance while for our procedure the bias is essentially reduced during iteration and the final quality of estimation is proportional to the standard deviation σ. We also observe a very stable performance of the procedure in case of moderate error variance and design asymmetry. The results are also uniformly (and essentially) better than for the other considered methods like SIR II or PHD. One reason could be that the assumption on the design required for the SIR or PHD to work is not fulfilled in our example.

The box plots in Figure 2 provide some information about the distribution of the criterion $\sqrt{n}\|\mathscr{R}^*(I-\widehat{\mathscr{P}_m})\|_2$ for the "best" one-step estimate and after the first, second, fourth, eighth and final iteration for $d = 10$, $m = 2$ and different

TABLE 1

Case $m = 1$: *mean loss* $\|\mathscr{R}^*(I - \widehat{\mathscr{P}}_m)\|_2/\|\mathscr{R}^*\|_2$ *for the first, second, fourth, eighth and final iteration, the "best" one step estimate (ADE), SIR II and PHD. Results are obtained from* $N = 250$ *simulations. The interquartile range of the losses is given in parentheses*

d	n	σ	τ	**1st**	**2nd**	**4th**	**8th**	**Final**	**ADE**	**SIR II**	**PHD**
3	200	0.1	1	0.0508	0.0419	0.0359	0.0271	0.0236	0.0442	0.106	0.113
				(0.038)	(0.031)	(0.026)	(0.019)	(0.014)	(0.032)	(0.050)	(0.072)
4	200	0.1	1	0.0606	0.0484	0.0417	0.0339	0.0309	0.0558	0.121	0.122
				(0.033)	(0.024)	(0.025)	(0.02)	(0.018)	(0.034)	(0.061)	(0.066)
6	200	0.1	1	0.0829	0.0631	0.0536	0.0437	0.0389	0.0807	0.150	0.159
				(0.034)	(0.024)	(0.024)	(0.02)	(0.018)	(0.036)	(0.059)	(0.066)
10	100	0.1	1	0.341	0.208	0.146	0.105	0.0903	0.341	0.283	0.323
				(0.14)	(0.083)	(0.067)	(0.047)	(0.04)	(0.14)	(0.107)	(0.121)
10	200	0.1	1	0.173	0.109	0.0854	0.0646	0.0537	0.172	0.205	0.220
				(0.065)	(0.036)	(0.026)	(0.02)	(0.017)	(0.066)	(0.058)	(0.067)
10	400	0.1	1	0.103	0.0698	0.0573	0.0438	0.0369	0.101	0.150	0.158
				(0.031)	(0.024)	(0.019)	(0.015)	(0.012)	(0.029)	(0.045)	(0.046)
10	800	0.1	1	0.0642	0.0479	0.0409	0.032	0.0271	0.0619	0.122	0.122
				(0.019)	(0.015)	(0.013)	(0.011)	(0.0084)	(0.019)	(0.031)	(0.033)

sample sizes n. Results displayed are obtained from $N = 250$ simulations. The results confirm the root-n consistence of the final estimate as claimed by Theorem 5.1 from Section 5. Note that the losses even being multiplied by $\sqrt{n}$ are still slightly improved with growing n.

5. Main results.

In this section we present some results describing the properties of the previously introduced basic procedure. The modified procedure can be considered similarly.

5.1. *Assumptions.*

We consider the following assumptions.

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

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 (2.6) is two times differentiable with a bounded second derivative, so that, for some constants C_g and for all $u, v \in \mathbb{R}^m$,

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

ASSUMPTION 4 (Range of parameters h_k, ρ_k). The parameters of the procedure satisfy $\rho_1 = 1, \rho_{\min} = n^{-1/3}, h_1 = C_0 n^{-1/(4\vee d)}$ with a constant $C_0 \geq 1$, $h_{\max} \geq 1$ and $a_h^d a_\rho^m \geq 1$.

TABLE 2

Case $m = 2$: mean loss $\|\mathscr{R}^*(I-\widehat{\mathscr{P}}_m)\|_2/\|\mathscr{R}^*\|_2$ for the first, second, fourth, eighth and final iteration, the "best" one step estimate (ADE), SIR II and PHD. Results are obtained from 250 *simulations* (100 *for* $d > 10$). *The interquartile range of the losses is given in parentheses*

d	n	σ	τ	**1st**	**2nd**	**4th**	**8th**	**Final**	**ADE**	**SIR II**	**PHD**
3	200	0.1	1	0.0207 (0.016)	0.0142 (0.01)	0.0124 (0.0085)	0.0116 (0.0074)	0.0114 (0.0076)	0.0203 (0.015)	0.0647 (0.042)	0.0728 (0.056)
4	200	0.1	1	0.0398 (0.019)	0.0273 (0.013)	0.0224 (0.011)	0.0203 (0.01)	0.0208 (0.0099)	0.0398 (0.019)	0.102 (0.045)	0.11 (0.055)
6	200	0.1	1	0.0837 (0.034)	0.058 (0.021)	0.048 (0.019)	0.037 (0.016)	0.0313 (0.014)	0.0832 (0.033)	0.14 (0.049)	0.162 (0.052)
10	100	0.1	1	0.33 (0.095)	0.223 (0.072)	0.189 (0.062)	0.181 (0.08)	0.182 (0.087)	0.327 (0.092)	0.315 (0.083)	0.37 (0.093)
10	200	0.1	1	0.18 (0.046)	0.11 (0.033)	0.0897 (0.027)	0.0616 (0.019)	0.0472 (0.016)	0.18 (0.046)	0.209 (0.051)	0.246 (0.06)
10	400	0.1	1	0.109 (0.025)	0.0617 (0.016)	0.0484 (0.014)	0.0289 (0.009)	0.0216 (0.0062)	0.109 (0.025)	0.146 (0.038)	0.169 (0.039)
10	800	0.1	1	0.0636 (0.014)	0.0404 (0.0092)	0.0325 (0.0083)	0.0192 (0.0056)	0.012 (0.0033)	0.0636 (0.014)	0.105 (0.023)	0.114 (0.026)
20	800	0.1	1	0.166 (0.021)	0.107 (0.013)	0.0821 (0.014)	0.0462 (0.0088)	0.0227 (0.0047)	0.162 (0.022)	0.157 (0.027)	0.18 (0.03)
50	800	0.1	1	0.617 (0.15)	0.349 (0.056)	0.252 (0.03)	0.146 (0.033)	0.0623 (0.011)	0.617 (0.15)	0.265 (0.031)	0.324 (0.033)
10	400	0.05	1	0.107 (0.024)	0.0577 (0.015)	0.0444 (0.014)	0.0237 (0.0075)	0.0141 (0.004)	0.107 (0.024)	0.141 (0.036)	0.168 (0.037)
10	400	0.2	1	0.117 (0.028)	0.0766 (0.02)	0.0622 (0.016)	0.0444 (0.012)	0.0397 (0.01)	0.117 (0.028)	0.161 (0.04)	0.172 (0.041)
10	400	0.1	0.75	0.102 (0.025)	0.0628 (0.019)	0.0531 (0.018)	0.0306 (0.012)	0.0191 (0.0054)	0.102 (0.023)	0.153 (0.037)	0.165 (0.039)
10	400	0.1	1.5	0.115 (0.027)	0.0784 (0.024)	0.0789 (0.031)	0.0662 (0.039)	0.0424 (0.018)	0.11 (0.029)	0.19 (0.048)	0.197 (0.055)

Our last assumption concerns the design properties. In what follows we assume a deterministic design, 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 that $X_1, \dots, X_n$ are i.i.d. random points in $\mathbb{R}^d$ with a design density $p(x)$. Then all the results should be understood to hold conditionally on the design.

In order for the algorithm to work, we have to suppose that the design points (X_i) are "well diffused" and, as a consequence, all the matrices $V_k(X_i)$ are well defined.

The estimation procedure utilizes the matrices S_k with

$$S_k^2 = I + \rho_k^{-2}\widehat{\mathscr{B}}_{k-1}\widehat{\mathscr{B}}_{k-1}^{\mathsf{T}}$$

where $\widehat{\mathscr{B}}_{k-1}$ is the estimate of the matrix $\mathscr{B}^*$ constructed at the preceding iteration step. We also introduce an "ideal" matrix $S_k^* = (I + \rho_k^{-2}\mathscr{B}^*(\mathscr{B}^*)^{\mathsf{T}})^{1/2}$

TABLE 3

Case $m = 3$: mean loss $\|\mathscr{R}^*(I - \widehat{\mathscr{P}}_m)\|_2/\|\mathscr{R}^*\|_2$ for the first, second, fourth, eighth and final iteration. Results are obtained from $N = 250$ simulations. The interquartile range of the losses is given in parentheses

d	n	σ	τ	1st	2nd	4th	8th	Final	ADE
10	800	0.1	1	0.0614 (0.013)	0.0454 (0.01)	0.036 (0.0084)	0.0229 (0.0061)	0.017 (0.0036)	0.0614 (0.012)
10	800	0.1	0.75	0.0677 (0.015)	0.054 (0.013)	0.0476 (0.012)	0.0345 (0.011)	0.018 (0.0054)	0.0660 (0.016)
10	800	0.1	1.5	0.0701 (0.016)	0.0571 (0.013)	0.0532 (0.011)	0.0472 (0.013)	0.0293 (0.0093)	0.0697 (0.015)

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 matrix $\mathscr{B}^*$ by $\widehat{\mathscr{B}}_{k-1}$. If $\widehat{\mathscr{B}}_{k-1} = \mathscr{B}^*$, then $U_k = I$. We shall see that these matrices U_k are typically close to I. Define now, given a matrix U and $k \le k(n)$,

$$Z_{ij,k} = h_k^{-1} S_k^*(X_j - X_i), \qquad i, j = 1, \dots, n,$$

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

$$\mathscr{V}_{i,k}(U) = \sum_{j=1}^{n} \begin{pmatrix} 1 \\ Z_{ij,k} \end{pmatrix} \begin{pmatrix} 1 \\ Z_{ij,k} \end{pmatrix}^{\mathsf{T}} K(Z_{ij,k}^{\mathsf{T}} U Z_{ij,k}), \qquad i = 1, \dots, n.$$

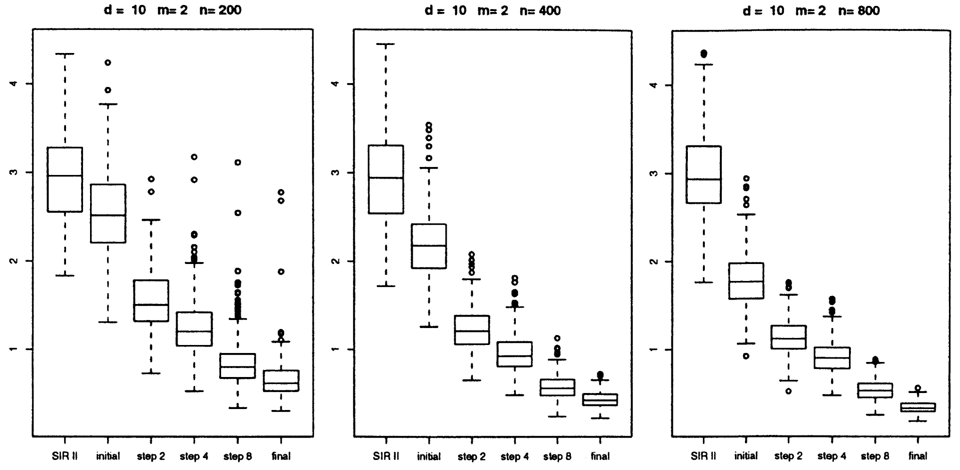

FIG. 2. *Simulation results in terms of $\sqrt{n}\|\mathscr{R}^*(I - \widehat{\mathscr{P}}_m)\|_2/\|\mathscr{R}^*\|_2$ for $m = 2$, $d = 10$ and $n = 200, 400, 800$ for the estimates obtained by SIR II, the initial estimate, second, fourth, eighth and final iteration.*

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 $i \le n$.

We use below the notation $\|A\|$ for the sup-norm of A: $\|A\| = \sup_\lambda |A\lambda|/|\lambda|$.

ASSUMPTION 5 (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:

1. The inverse matrices $\mathscr{V}_{i,k}(U)^{-1}$ are well defined and
$$N_{i,k}(U)\|\mathscr{V}_{i,k}(U)^{-1}\| \le C_V, \qquad i = 1,\dots,n;$$
2. For $j = 1,\dots,n$,
$$\sum_{i=1}^{n} \frac{1}{N_{i,k}(U)} K\bigl(Z_{ij,k}^{\mathsf{T}} U Z_{ij,k}\bigr) \le C_K,$$
$$\sum_{i=1}^{n} \frac{1}{N_{i,k}(U)} \bigl|K'\bigl(Z_{ij,k}^{\mathsf{T}} U Z_{ij,k}\bigr)\bigr| \le C_{K'}.$$

Here K' means the derivative of the kernel K.

REMARK 5.1. One can easily check that for the case of a random design with a continuous positive density, one can fix some constant C_V, C_K and $C_{K'}$ depending on the dimension d and design density only and such that the conditions from Assumption 5 are fulfilled with a high probability converging exponentially fast to 1 as n grows. Some results on semiparametric M-estimation in the single-index model only require that the projection of the design on the e.d.r. space has a continuous density; see, for example, Carroll (1997). This condition is not sufficient for us since the procedure estimates the gradient of the regression function which is impossible if, for example, the design is concentrated on a low-dimensional subspace.

In what follows by $C, C_1, C_2, \dots$ we denote generic constants depending on $d, C_g, C_V, C_K, C_{K'}, \psi_\ell, L$ and σ only.

5.2. *Loss of information caused by estimated e.d.r.* An important characteristic of the estimated e.d.r. $\widehat{\mathscr{R}}_m$ is the loss of information caused by this reduction. Due to the representation (2.6), the information contained in a unit vector $v \in \mathbb{R}^d$ can be measured by the value $|\mathscr{R}^* v|$. A loss of information occurs if $|\mathscr{R}^* v| > 0$ but $|\widehat{\mathscr{R}}_m v| = 0$. Let Π^* be the projector in $\mathbb{R}^d$ onto the true index space $\mathscr{I}$ and similarly, $\widehat{\mathscr{P}}_m$ denotes the projector in $\mathbb{R}^d$ onto the estimated index space $\widehat{\mathscr{I}}$ corresponding to the e.d.r. $\widehat{\mathscr{R}}_m$; that is, $\widehat{\mathscr{I}} = \operatorname{Im} \widehat{\mathscr{R}}_m^{\mathsf{T}}$. Then the total loss of information by e.d.r. $\widehat{\mathscr{R}}_m$ can be measured by the value
$$\|\mathscr{R}^*(I - \widehat{\mathscr{P}}_m)\|_2,$$

where $\|A\|_2$ is the Euclidean norm of the matrix A; that is, $\|A\|_2^2 = \operatorname{tr} AA^{\mathsf{T}} = \operatorname{tr} A^{\mathsf{T}}A$. In the sequel we use the following obvious inequalities: $\|A\| \le \|A\|_2 \le \sqrt{m}\|A\|$ where m is the rank of A.

The next result claims that the loss of information caused by the e.d.r. $\widehat{\mathscr{R}}_m$ is of order $n^{-1/2}$.

THEOREM 5.1. *Let $\widehat{\mathscr{R}}_m$ be defined by* (3.1). *For $m \le 3$, there exists a sequence $\varkappa_n \to 0$ as $n \to \infty$ such that under Assumptions* 1 *through* 5, *for sufficiently large n and every $z \ge 1$,*

$$\boldsymbol{P}\left(\|\widehat{\mathscr{R}}_m(I - \Pi^*)\|_2 > \frac{2zH_1}{\sqrt{n}} + Ct_n^2 n^{-2/3}\right) < ze^{-(z^2-1)/2} + \frac{3k(n)}{n},$$

$$\boldsymbol{P}\left(\|\mathscr{R}^*(I - \widehat{\mathscr{P}}_m)\|_2 > \frac{2zH_1}{\sqrt{n(1-\varkappa_n)}} + Ct_n^2 n^{-2/3}\right) < ze^{-(z^2-1)/2} + \frac{3k(n)}{n},$$

with $t_n = (1 + 2\log n + 2\log\log n)^{1/2}$ *and*

$$H_1 = \sqrt{2}\sigma C_V C_K \bar{\psi}\sqrt{L},$$
$$\bar{\psi} = \max_{i=1,\dots,n}\ \max_{\ell=1,\dots,L} |\psi_\ell(X_i)|. \tag{5.1}$$

5.3. *Estimation of the index space.* By construction, $\mathscr{R}^*$ is an orthogonal mapping from $\mathbb{R}^d$ to $\mathbb{R}^m$, that is, $\mathscr{R}^*(\mathscr{R}^*)^{\mathsf{T}}$ is a diagonal $m \times m$-matrix with the diagonal elements $\lambda_1, \dots, \lambda_m$. Moreover, the product $\Pi^* = (\mathscr{R}^*)^{\mathsf{T}}(\mathscr{R}^*(\mathscr{R}^*)^{\mathsf{T}})^{-1} \times \mathscr{R}^*$ is the projector in $\mathbb{R}^d$ onto the corresponding index space $\mathscr{I}$. Similarly, $\widehat{\mathscr{P}}_m = \widehat{\mathscr{R}}_m^{\mathsf{T}}(\widehat{\mathscr{R}}_m\widehat{\mathscr{R}}_m^{\mathsf{T}})^{-1}\widehat{\mathscr{R}}_m$ is the projector onto the estimated e.d.r. space. Thus the quality of the identification of the true index space can be measured by the error of estimating Π^* with $\widehat{\mathscr{P}}_m$. We encounter the following identifiability problem: if, for instance, the last eigenvalue λ_m is (close to) zero, then the corresponding eigenvector e_m is not uniquely defined. The next result states that if the eigenvalue λ_m is separated away from zero, the estimated projector $\widehat{\mathscr{P}}_m$ recovers Π^* at the rate $n^{-1/2}$.

THEOREM 5.2. *Let $m \le 3$ and Assumptions* 1 *through* 5 *hold. For n sufficiently large,*

$$\boldsymbol{P}\left(\left\|\Pi^* - \widehat{\mathscr{P}}_m\right\|_2 > \frac{2\sqrt{2}\lambda_m^{-1/2} zH_1}{\sqrt{n(1-\varkappa_n)}} + Ct_n^2 n^{-2/3}\right) \le ze^{-(z^2-1)/2} + \frac{3k(n)}{n}$$

with $\varkappa_n$ and H_1 as in Theorem 5.1.

REMARK 5.2. Since λ_m is the mth eigenvalue of the matrix $\mathscr{M}_L^*$, the condition $\lambda_m > 0$ relies both on the model function f and on the basis system $\psi_1, \dots, \psi_L$. If λ_m^* is the mth eigenvalue of $\mathscr{M}^*$, then the ratio λ_m/λ_m^* characterizes the quality of the basis $\{\psi_\ell\}$. This value typically approaches one as L grows. Our numerical examples are also in favour of a larger L.

5.4. *Estimation of the matrix $\mathscr{B}^*$.* In this section we present some results describing the quality of estimating the vectors β_ℓ^* by the proposed estimation procedure. The first result describes the accuracy of the first step estimate, and the next result describes the quality of the final estimate.

5.4.1. *The first-step approximation.* Let $\hat{\beta}_{1,\ell}, \ell = 1, \dots, L$ be the family of the estimates obtained at the first step of the iterative procedure with $\rho_1 = 1, S_1 = I$ and some h_1.

PROPOSITION 5.1. *Under Assumptions* 1 *through* 5, *for every* $\ell \le L$,

$$\hat{\beta}_{1,\ell} - \beta_\ell^* = C_{1,\ell} h_1 + \frac{\xi_{1,\ell}}{h_1\sqrt{n}},$$

where $C_{1,\ell}$ *is a constant and* $\xi_{1,\ell}$ *is a zero mean normal random vector in* $\mathbb{R}^d$ *satisfying*

$$C_{1,\ell} \le \sqrt{2} C_g C_V \bar{\psi}_\ell,$$
$$\boldsymbol{E}|\xi_{1,\ell}|^2 \le 2\sigma^2 C_V^2 C_K^2 \bar{\psi}_\ell^2.$$

REMARK 5.3. The bandwidth h_1 should be at least of order $n^{-1/d}$ to provide at least $d+1$ design points in almost every spherical neighborhood of radius h_1. The optimization of the risk of the first step estimate under the constraint $h_1 \ge$ Const. $h^{-1/d}$ leads to the following rule for the choice of h_1: $h_1 =$ Const. $n^{-1/(4\vee d)}$. Hence, we get the accuracy for $\hat{\beta}_{1,\ell}$,

$$|\hat{\beta}_{1,\ell} - \beta_\ell^*| \le \text{Const. } n^{-(1/4\wedge 1/d)}.$$

5.4.2. *Accuracy of the final estimate.* Let $\hat{\beta}_\ell$'s be the estimates of β_ℓ^*'s obtained at the last iteration, $\ell = 1, \dots, L$. As previously, $\widehat{\mathscr{B}}$ denotes the matrix composed by the vectors $\hat{\beta}_\ell$. It turns out that the quality of estimation delivered by $\widehat{\mathscr{B}}$ is not homogeneous w.r.t. to the orientation in the space $\mathbb{R}^d$. This heterogeneity is caused by application of elliptic windows for estimating the gradient vectors $F(X_i)$. To mimic this property, we introduce for every $k \le k(n)$ an operator ($d \times d$-matrix) $P_{\rho_k}^* = (I + \rho_k^{-2}\mathscr{B}^*(\mathscr{B}^*)^\top)^{-1/2} = (S_k^*)^{-1}$ which, roughly speaking, multiplies by the factor ρ_k within the index space $\mathscr{I}$ while, being restricted to the orthogonal subspace $\mathscr{I}^\perp$, it coincides with the identity mapping.

THEOREM 5.3. *Let* $m \le 3$ *and Assumptions* 1 *through* 5 *hold. There exists a Gaussian zero mean random* $d \times L$*-matrix* $\xi^* \in \mathbb{R}^{dL}$ *such that, with* $\rho = \rho_{k(n)}$ *and* n *large enough*,

$$\boldsymbol{P}\left(\left\| P_\rho^*(\widehat{\mathscr{B}} - \mathscr{B}^*) - \frac{\xi^*}{\sqrt{n}} \right\|_2 > C_1 t_n^2 n^{-2/3}\right) \le \frac{3k(n) - 1}{n}$$

and

$$\boldsymbol{E}\|\xi^*\|_2^2 \le 2\sigma^2\bar{\psi}^2 L C_V^2 C_K^2 = H_1^2.$$

COROLLARY 5.1. *Under the conditions of Theorem* 5.3, *for every* $z \ge 1$,

$$\boldsymbol{P}\biggl(\|P_\rho^*(\widehat{\mathscr{B}} - \mathscr{B}^*)\|_2 > \frac{zH_1}{\sqrt{n}} + C_1 t_n^2 n^{-2/3}\biggr) \le ze^{-(z^2-1)/2} + \frac{3k(n)-1}{n}.$$

6. Conclusions and outlook. We introduce a new method of dimension reduction based on the idea of structural adaptation. The method applies for a very broad class of regression models under mild assumptions on the underlying regression function and the regression design. The procedure is fully adaptive and does not require any prior information. The results claim that the proposed procedure delivers the optimal rate $n^{-1/2}$ of estimating the index space provided that the effective dimensionality of the model is not larger than 3. The simulation results demonstrate an excellent performance of the procedure for all situations considered. An important feature of the method is that it is very stable with respect to high dimensionality and for a nonregular design.

It is worth noting that the basic iterative procedure does not rely on m. This value is used only for the last step of describing the m-dimensional e.d.r. If the effective dimension m exceeds 4, then the procedure continues to apply and it allows estimating the index space, but the corresponding accuracy would be worse than $n^{-1/2}$. One more open question concerns the case of an unknown effective dimension. Note first that if we apply some m which is smaller than the real effective dimension m^* when describing the e.d.r. space, then the best m-index approximation of the original model is expected to be obtained. In practical applications, the following two problems arise: estimation of m and testing a m-index hypothesis. The matrix $\widehat{\mathscr{B}}$ from the last step of the algorithm can be used for answering the above mentioned problems. A detailed study of this situation is an important topic for further research.

The procedure can be easily extended to the situation with a multivariate response variable $Y \in \mathbb{R}^p$ with $p > 1$. The underlying multi-index assumption remains of the same functional form: $\boldsymbol{E}(Y \mid X) = f(x) = g(X^\top\theta_1, \dots, X^\top\theta_m)$ where g is a vector function on $\mathbb{R}^m$ with values in $\mathbb{R}^p$. This means that the gradient $F_j = \nabla f_j$ of each component f_j of f belongs to the index space spanned by vectors $\theta_1, \dots, \theta_m$ and one can utilize the same ideas as previously for estimating the index space $\mathscr{I}$. The only difference is that the basis functions $\{\psi_\ell\}$ should also be vectors in $\mathbb{R}^p$. A reasonable example corresponds to the procedure which estimates for every component f_j, $j = 1, \dots, p$, of the regression function $f \in \mathbb{R}^p$ the vectors $\beta^*_{1,j}, \dots, \beta^*_{L,j}$ with

$$\beta^*_{\ell,j} = \frac{1}{n}\sum_{i=1}^n F_j(X_i)\psi_\ell(X_i), \qquad \ell = 1, \dots, L,$$

and the same ψ_ℓ's and then utilizes the total collections of the vectors $\{\hat{\beta}_{\ell,j}\}$ with $\ell = 1, \dots, L$ and $j = 1, \dots, p$ for estimating the index space $\mathscr{I}$.

Another interesting issue arises when considering multiple time series and especially financial data. We regard such extensions as topics for further research.

One more important question is semiparametric efficiency. Our procedure is shown to be rate optimal, at least for $m \le 3$. In the single-index situation there are several methods which are also asymptotically efficient in the semiparametric sense; see, for example, Carroll, Fan, Gijbels and Wand (1997). Our procedure is not expected to achieve the semiparametric efficiency in the single-index model, but it can be used for constructing a semiparametrically efficient estimator by one-step improvement.

APPENDIX A

Proofs. Here we collect the proofs of the assertions formulated previously. All our results are based on the following technical assertion describing an improvement of the estimate $\widehat{\mathscr{B}}$ at each iteration step.

A.1. *One-step improvement.* Suppose that we are given some fixed numbers h and ρ (which mean the current values h_k and ρ_k) and a fixed $d \times L$-matrix B which can be viewed as an approximation $\widehat{\mathscr{B}}_{k-1}$ of $\mathscr{B}^*$ obtained at the previous step. Set also

$$S_B = (I + \rho^{-2} B B^{\mathsf{T}})^{1/2},$$

$$V_B(X_i) = \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\mathsf{T}} K\biggl(\frac{|S_B X_{ij}|^2}{h^2} \biggr),$$

$$\begin{pmatrix} \hat{f}_B(X_i) \\ \widehat{F}_B(X_i) \end{pmatrix} = V_B(X_i)^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\biggl(\frac{|S_B X_{ij}|^2}{h^2} \biggr), \tag{A.1}$$

$$\hat{\beta}_{B,\ell} = \frac{1}{n} \sum_{i=1}^{n} \widehat{F}_B(X_i)\, \psi_\ell(X_i), \tag{A.2}$$

where, recall, $X_{ij} = X_j - X_i$, and define the matrix $\widehat{\mathscr{B}}_B$ with columns $\hat{\beta}_{B,\ell}$, $\ell = 1, \dots, L$. We aim to evaluate the estimation errors $\widehat{\mathscr{B}}_B - \mathscr{B}^*$. To describe the results, we introduce the matrix (linear operator) $P_\rho^* = (I + \rho^{-2} \mathscr{B}^* (\mathscr{B}^*)^{\mathsf{T}})^{-1/2}$. Define also for some positive $\delta < \rho/4$, the set $\mathfrak{B}_{\delta,\rho}$ by

$$\mathfrak{B}_{\delta,\rho} = \bigl\{ B \colon \| P_\rho^* (B - \mathscr{B}^*) \|_2 \le \delta \bigr\}.$$

PROPOSITION A.1. *Let Assumptions* 1 *through* 5 *hold. Then there exists Gaussian random* $d \times L$*-matrix* ξ *such that, with* $\alpha = 2\delta/\rho + \delta^2/\rho^2$,

$$\boldsymbol{P}\biggl(\sup_{B \in \mathfrak{B}_{\delta,\rho}} \biggl\| P_\rho^* (\widehat{\mathscr{B}}_B - \mathscr{B}^*) - \frac{\xi}{h\sqrt{n}} \biggr\|_2 > \frac{\sqrt{2} C_g C_V \bar{\psi} \sqrt{L}}{(1-\alpha)^{3/2}} \rho^2 h + \frac{\sigma \bar{\psi} \sqrt{L} C_{\alpha,n} \alpha}{h\sqrt{n}} \biggr) \le \frac{2}{n},$$

where

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

and

$$\boldsymbol{E}\|\xi\|_2^2 \le 2\sigma^2 C_V^2 C_K^2 \bar{\psi}^2 L. \tag{A.4}$$

Before proving this statement, we present one straightforward corollary.

COROLLARY A.1. *Under Assumptions* 1 *through* 5 *for every* $z \ge 1$,

$$\boldsymbol{P}\left(\sup_{B\in\mathfrak{B}_{\delta,\rho}} \|P_\rho^*(\widehat{\mathscr{B}}_B - \mathscr{B}^*)\|_2 > \bar{\psi}\sqrt{L}\left(\frac{\sqrt{2}C_g C_V \rho^2 h}{(1-\alpha)^{3/2}} + \frac{z\sqrt{2}\sigma C_V C_K}{h\sqrt{n}} + \frac{\sigma C_{\alpha,n}\alpha}{h\sqrt{n}}\right)\right)$$
$$\le z e^{-(z^2-1)/2} + 2/n.$$

Indeed, the Gaussian vector $\xi \in \mathbb{R}^{dL}$ satisfies for every $z \ge 1$,

$$\boldsymbol{P}\left(\|\xi\|_2 \ge z\sqrt{\boldsymbol{E}\|\xi\|_2^2}\right) \le z e^{-(z^2-1)/2}$$

(see Lemma 9 in HJS01), and the assertion follows from Proposition A.1.

PROOF OF PROPOSITION A.1. We follow the line of the proof of Proposition 2 in HJS01 and focus here only on essential points, omitting technical details.

It is useful to define

$$u = \rho^{-1}P_\rho^* B, \qquad U = P_\rho^*(I + \rho^{-2}B B^\top)P_\rho^* = (P_\rho^*)^2 + uu^\top$$

and similarly,

$$u^* = \rho^{-1}P_\rho^*\mathscr{B}^*, \qquad U^* = P_\rho^*\left(I + \rho^{-2}\mathscr{B}^*(\mathscr{B}^*)^\top\right)P_\rho^* = I$$

so that u, u^* are $d \times L$-matrices and U, U^* are $d \times d$ symmetric matrices. Clearly $B = \mathscr{B}^*$ implies $U = I$ and the condition $\|B - \mathscr{B}^*\|_2 \le \delta$ implies $\|u - u^*\|_2 \le \delta/\rho$, that is, the inclusion $B \in \mathfrak{B}_{\delta,\rho}$ is equivalent to $u \in \{u\colon \|u - u^*\|_2 \le \delta/\rho\}$. Due to Lemma B.1, it also follows that $\|U - U^*\| = \|uu^\top - u^*(u^*)^\top\| \le \alpha = 2\delta/\rho + \delta^2/\rho^2$ for all such u.

Next, for every $i, j \le n$, define

$$Z_{ij} = h^{-1}(P_\rho^*)^{-1}(X_j - X_i),$$

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

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

It is easy to check that $\hat{s}_i(U) = \begin{pmatrix} h^{-1}\hat{f}_B(X_i) \\ P_\rho^* \widehat{F}_B(X_i) \end{pmatrix}$ and hence,

$$P_\rho^* \hat{\beta}_{B,\ell} = \mathscr{E}_d n^{-1} \sum_{i=1}^{n} \hat{s}_i(U)\psi_\ell(X_i),$$

where $\mathscr{E}_d$ denotes the projector from $\mathbb{R}^{d+1}$ onto $\mathbb{R}^d$ keeping the last d coordinates.

The model equation (1.2) implies

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

with

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

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

so that

$$P_\rho^*\bigl(\hat{\beta}_{B,\ell} - \beta_\ell^*\bigr) = \frac{1}{n}\sum_{i=1}^{n}\bigl\{\mathscr{E}_d s_i(U) - P_\rho^* F(X_i)\bigr\}\psi_\ell(X_i)$$
$$+ \mathscr{E}_d n^{-1} \sum_{i=1}^{n} \zeta_i(U)\psi_\ell(X_i).$$

Clearly $\xi_\ell(U) = \mathscr{E}_d n^{-1} \sum_{i=1}^{n} \zeta_i(U)\psi_\ell(X_i)$ is for every U a linear combination of the Gaussian errors ε_i and therefore it is also a Gaussian vector in $\mathbb{R}^d$. We define $\xi(U)$ to be the $d \times L$ matrix with columns $\xi_\ell(U)$ and set $\xi = \xi(U^*)$. It is easy to see that the following three statements imply the desired result.

$$\text{(A.5)} \qquad \sup_{u:\ \|u-u^*\|_2 \le \delta/\rho} |\mathscr{E}_d s_i(U) - P_\rho^* F(X_i)| \le \frac{\sqrt{2} C_g C_V}{(1-\alpha)^{3/2}} h\rho^2, \quad i = 1, \dots, n,$$

$$\text{(A.6)} \qquad \boldsymbol{P}\Biggl(\sup_{u:\ \|u-u^*\|_2 \le \delta/\rho} \|\xi(U) - \xi(U^*)\|_2 > \frac{\sigma C_{\alpha,n}\alpha}{h\sqrt{n}}\Biggr) \le 2/n$$

with $U = (P_\rho^*)^2 + uu^\top$ and $U^* = I$, and for all $\ell = 1, \dots, L$,

$$\text{(A.7)} \qquad \boldsymbol{E}|\xi_\ell(U^*)|^2 \le \frac{2\sigma^2 C_V^2 C_K^2 \bar{\psi}_\ell^2}{h^2 n}.$$

To check these statements, the following lemma will be useful.

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

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

$$\left|Z_{ij}^{\top} U Z_{ij} - |Z_{ij}|^2\right| = \left|Z_{ij}^{\top}(U - I) Z_{ij}\right| \le \alpha |Z_{ij}|^2$$

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

First we evaluate the "bias" term $\mathscr{E}_d s_i(U) - P_\rho^* F(X_i)$. Since

$$\begin{aligned}\begin{pmatrix} h^{-1} f(X_i) \\ P_\rho^* F(X_i) \end{pmatrix} &= \mathscr{V}_i(U)^{-1} \sum_{j=1}^{n} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix}^{\top} \begin{pmatrix} h^{-1} f(X_i) \\ P_\rho^* F(X_i) \end{pmatrix} K(Z_{ij}^{\top} U Z_{ij}) \\ &= h^{-1} \mathscr{V}_i(U)^{-1} \sum_{j=1}^{n} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} \left\{ f(X_i) + (X_j - X_i)^{\top} F(X_i) \right\} K\left(Z_{ij}^{\top} U Z_{ij}\right)\end{aligned}$$

it follows that

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

where in view of (2.6),

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

The use of $P_\rho^* \mathscr{B}^* (\mathscr{B}^*)^{\top} P_\rho^* = \rho^2 (I - (P_\rho^*)^2)$ and $\|I - (P_\rho^*)^2\| \le 1$ yields

$$\begin{aligned} |(\mathscr{B}^*)^{\top} X_j - (\mathscr{B}^*)^{\top} X_i|^2 &= (X_j - X_i)^{\top} \mathscr{B}^* (\mathscr{B}^*)^{\top} (X_j - X_i) \\ &= \left((P_\rho^*)^{-1}(X_j - X_i)\right)^{\top} P_\rho^* \mathscr{B}^* (\mathscr{B}^*)^{\top} P_\rho^* (P_\rho^*)^{-1} (X_j - X_i) \\ &= h^2 \rho^2 Z_{ij}^{\top} \left(I - (P_\rho^*)^2\right) Z_{ij} \\ &\le h^2 \rho^2 |Z_{ij}|^2 \end{aligned}$$

which also implies

$$|\mathscr{R}^* X_j - \mathscr{R}^* X_i| = |(\mathscr{B}^* O_m)^{\top} X_j - (\mathscr{B}^* O_m)^{\top} X_i|^2 \le h^2 \rho^2 |Z_{ij}|^2.$$

This yields by Lemma A.1 and Assumption 3 for every pair (i, j) with $Z_{ij}^{\top} U Z_{ij} \le 1$,

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

and using Assumptions 5 we bound

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

and (A.5) follows.

Further, we study the stochastic components $\xi_\ell(U)$. It follows directly from the definition that there are vector coefficients $c_{i,\ell}(U)$ such that

$$
\xi_\ell(U) = \sum_{i=1}^n c_{i,\ell}(U) \varepsilon_i .
$$

We now apply the following two technical results from HJS01, see Lemmas 3, 10 there for a particular case with $L=1$ and $\psi_\ell \equiv 1$. Extension to general L and ψ_ℓ's is straightforward.

LEMMA A.2.

(i)

$$
\sum_{i=1}^n |c_{i,\ell}(U^*)|^2 \le \frac{2 C_V^2 C_K^2 \bar{\psi}_\ell^2}{h^2 n};
$$

(ii)

$$
\sup_{U:\ \|U-I\|\le\alpha} \sum_{i=1}^n |c_{i,\ell}(U)|^2 \le \frac{2 C_V^2 C_K^2 \bar{\psi}_\ell^2}{(1-\alpha) h^2 n};
$$

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

$$
\sup_{U:\ \|U-I\|\le\alpha} \left\| \frac{d}{dU} e^\top c_{i,\ell}(U) \right\| \le \frac{\kappa_\alpha \bar{\psi}_\ell}{nh}
$$

with

$$
\kappa_\alpha = \sqrt{2}(1-\alpha)^{-3/2} C_V C_{K'} + 2\sqrt{2}(1-\alpha)^{-5/2} C_V^2 C_{K'} C_K;
$$

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

$$
\sup_{u:\ \|u-u^*\|_2 \le \delta/\rho} \left\| \frac{d}{du} e^\top c_{i,\ell}(U) \right\| \le \frac{\kappa'_\alpha \bar{\psi}_\ell}{nh}
$$

with $U = U_u = (P_\rho^*)^2 + u u^\top$ *and*

$$
\kappa'_\alpha = \kappa_\alpha (1-\alpha)^{-1/2} = \sqrt{2}(1-\alpha)^{-2} C_V C_{K'} + 2\sqrt{2}(1-\alpha)^{-3} C_V^2 C_{K'} C_K .
$$

LEMMA A.3. *Let $r \ge 0$ and let vector functions $a_i(u)$ with $u \in \mathbb{R}^p$ obey the conditions*

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

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

$$\boldsymbol{P}\biggl(\sup_{|u-u^*|\le r}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\{a_i(u)-a_i(u^*)\}\varepsilon_i\right| > \sigma\kappa r\Bigl(2+\sqrt{(3+p)\log(4n)}\Bigr)\biggr) \le \frac{2}{n}.$$

Lemma A.2(i) implies (A.7). The statement (A.6) follows from Lemma A.2(iv), and Lemma A.3 applied to the matrix $\xi(U) \in \mathbb{R}^{dL}$ with columns $\xi_\ell(U)$ and with $U = U_u = (P_\rho^*)^2 + uu^\mathsf{T}$; for details see again HJS01. □

A.2. *Proof of Theorem* 5.3. To be able to apply Proposition A.1 to the estimates $\hat\beta_{k,\ell}$ at step k, we need that the matrix $B = \widehat{\mathscr{B}}_{k-1}$ coming as the result of the preceding iteration belongs to the set $\mathfrak{B}_{\rho,\delta}$ with $\rho = \rho_k$ and some $\delta < \rho/4$. Since the matrix $\widehat{\mathscr{B}}_{k-1}$ is random, we have to check that the probability of the event $\{\widehat{\mathscr{B}}_{k-1} \in \mathfrak{B}_{\rho_k,\delta}\} = \{B\colon \|P_\rho^*(B-\mathscr{B}^*)\|_2 \le \rho\}$ is sufficiently large. Further we show that this property is fulfilled if n is large enough.

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

$$\delta_k = \bar\psi\sqrt{L}\biggl(\frac{\sqrt{2}C_g C_V}{(1-\alpha_k)^{3/2}}h_k\rho_k^2 + \frac{\sqrt{2}\sigma C_V C_K t_n}{h_k\sqrt{n}} + \frac{\sigma C_{\alpha_k,n}\alpha_k}{2h_k\sqrt{n}}\biggr),$$

$$\alpha_{k+1} = \rho_{k+1}^{-2}(2\delta_k\rho_k + \delta_k^2),$$

where $t_n = (1+2\log n + 2\log\log n)^{1/2}$.

LEMMA A.4. *For $m \le 3$ and n sufficiently large, the values α_k's satisfy $\max_{k\le k(n)}\alpha_k < 1/4$. In addition, for the last iteration $k(n)$,*

$$\mu_n := \bar\psi\sqrt{L}\biggl(\frac{\sqrt{2}C_g C_V}{(1-\alpha_{k(n)})^{3/2}}h_{k(n)}\rho_{k(n)}^2 + \frac{\sigma C_{\alpha_{k(n)},n}\alpha_{k(n)}}{h_{k(n)}\sqrt{n}}\biggr) \le C_1 t_n^2 n^{-2/3}.$$

For the proof, see Lemma 5 in HJS01.

Next, successive application of the results of Proposition A.1 and Corollary A.1 with $t_n = (1+2\log n+2\log\log n)^{1/2}$ leads to the following.

LEMMA A.5. *Let n be sufficiently large. There exist random sets $\mathscr{A}_1 \supseteq \cdots \supseteq \mathscr{A}_{k(n)}$ such that*

$$\boldsymbol{P}(\mathscr{A}_k) \ge 1 - \frac{3k}{n}$$

and on $\mathscr{A}_k$

$$\bigl\|P_{\rho_{k+1}}^*(\widehat{\mathscr{B}}_k - \mathscr{B}^*)\bigr\|_2 \le \delta_k, \qquad k=1,\dots,k(n)-1.$$

For the proof, see Lemma 6 in HJS01.

Now the result of Theorem 5.3 can be proved by one more application of Proposition A.1 to the last step estimate $\widehat{\mathscr{B}} = \widehat{\mathscr{B}}_{k(n)}$ with $h = h_{k(n)} \geq 1$ and $\rho = \rho_{k(n)} \approx n^{-1/3}$; see again HJS01 for the detailed derivation.

A.3. *Proof of Theorem* 5.1. Let $\widehat{\mathscr{B}}$ be the last step estimate of the matrix $\mathscr{B}^*$. We know from Theorem 5.3 that, with probability close to 1, $\widehat{\mathscr{B}}$ satisfies the conditions

$$\|P_\rho^*(\widehat{\mathscr{B}} - \mathscr{B}^*)\|_2 \leq \tau, \tag{A.8}$$

with $\rho = \rho_{k(n)}$ and some small τ. This implies, by Lemma B.1,

$$\left\|\widehat{\mathscr{B}} - \Pi^* \widehat{\mathscr{B}}\right\|_2 \leq \tau. \tag{A.9}$$

where Π^* denotes the projector on the index space $\mathscr{I}$.

Recall that $\widehat{\mathscr{B}}$ approximates the $d \times L$-matrix $\mathscr{B}^*$ of rank m. However, it is typically of rank d. It is useful to introduce another $d \times L$-matrix $\widehat{\mathscr{B}}_m$ of rank m which minimizes the expression $\left\|\widehat{\mathscr{B}} - \widehat{\mathscr{B}}_m\right\|_2$ over all such matrices. The solution to this optimization problem can be described explicitly via the eigenvalue decomposition of the matrix $\widehat{\mathscr{B}}^\top \widehat{\mathscr{B}} = \widehat{O} \widehat{\Lambda}_L \widehat{O}^\top$ with an orthogonal matrix $\widehat{O}$ and a diagonal matrix $\widehat{\Lambda}_L$ with nonincreasing eigenvalues (cf. Lemma 2.1). We use the notation I_m for the diagonal $L \times L$-matrix with the first m diagonal elements equal to 1 and the remaining ones equal to zero.

LEMMA A.6 [Harville (1997), Theorem 21.12.4]. *The* $d \times L$-*matrix* $\widehat{\mathscr{B}}_m = \widehat{\mathscr{B}} \widehat{O} I_m \widehat{O}^\top$ *minimizes the norm* $\|B - \widehat{\mathscr{B}}\|_2$ *over all* $d \times L$-*matrices* B *of rank* m:

$$\widehat{\mathscr{B}}_m = \widehat{\mathscr{B}} \widehat{O} I_m \widehat{O}^\top = \operatorname*{arg\,inf}_{B \in \mathfrak{B}_m} \|\widehat{\mathscr{B}} - B\|_2, \tag{A.10}$$

where $\mathfrak{B}_m$ *denotes the set of* $d \times L$-*matrices of rank* m.

PROOF. Let $\widehat{\mathscr{B}}^\top \widehat{\mathscr{B}} = \widehat{O} \widehat{\Lambda}_L \widehat{O}^\top$. Then, for the $d \times L$-matrix $\widetilde{\mathscr{B}} = \widehat{\mathscr{B}} \widehat{O}$

$$\widetilde{\mathscr{B}}^\top \widetilde{\mathscr{B}} = O^\top \widehat{\mathscr{B}}^\top \widehat{\mathscr{B}} \widehat{O} = \widehat{O}^\top \widehat{O} \widehat{\Lambda}_L \widehat{O}^\top \widehat{O} = \widehat{\Lambda}_L;$$

that is, the columns of the matrix $\widetilde{\mathscr{B}}$ are orthogonal and they are arranged in a way that their norms decrease. This clearly implies

$$\operatorname*{arg\,inf}_{B \in \mathfrak{B}_m} \|\widetilde{\mathscr{B}} - B\|_2 = \widetilde{\mathscr{B}} I_m$$

and the assertion of the lemma follows by a change-of-basis argument. □

Recall that we define the e.d.r. matrix $\widehat{\mathscr{R}}_m$ by $\widehat{\mathscr{R}}_m=(\widehat{\mathscr{B}}\widehat{O}_m)^{\mathsf{T}}$; see (3.1). It follows from the last lemma that $\widehat{\mathscr{R}}_m=(\widehat{\mathscr{B}}_m\widehat{O}_m)^{\mathsf{T}}$. Also, (A.9) and the definition of $\widehat{\mathscr{B}}_m$ [see (A.10)] imply

$$\|\widehat{\mathscr{B}}-\widehat{\mathscr{B}}_m\|_2\le\|\widehat{\mathscr{B}}-\Pi^*\widehat{\mathscr{B}}\|_2\le\tau,$$

and, since $\|P_\rho^*\|\le 1$,

$$\big\|P_\rho^*(\widehat{\mathscr{B}}_m-\mathscr{B}^*)\big\|_2\le\|\widehat{\mathscr{B}}-\widehat{\mathscr{B}}_m\|_2+\big\|P_\rho^*(\widehat{\mathscr{B}}-\mathscr{B}^*)\big\|_2\le 2\tau. \tag{A.11}$$

This implies by Lemma B.2,

$$\big\|P_{\rho,m}(\widehat{\mathscr{B}}_m-\mathscr{B}^*)\big\|_2\le 2\tau(1-\varkappa)^{-1/2}, \tag{A.12}$$

where $P_{\rho,m}=(I+\rho^{-2}\widehat{\mathscr{B}}_m\widehat{\mathscr{B}}_m^{\mathsf{T}})^{-1/2}$ and $\varkappa=4\tau/\rho+4\tau^2/\rho^2$. Now the result of Theorem 5.1 is a straightforward application of Theorem 5.3 and Lemma B.3.

A.4. *Proof of Theorem* 5.2. Let $\widehat{\mathscr{B}}$ be the last step estimate of the matrix $\mathscr{B}^*$. We know from Theorem 5.3 that, with probability close to one, $\widehat{\mathscr{B}}$ satisfies the condition (A.8) with $\rho=\rho_{k(n)}$ and some small τ. Next, let the matrices $\widehat{\mathscr{B}}_m$, and $\widehat{\mathscr{R}}_m$ of rank m be defined as in the proof of Theorem 5.1 so that the condition (A.11) is satisfied. The projectors Π^* and $\widehat{\mathscr{P}}_m$ are defined as

$$\Pi^*=(\mathscr{R}^*)^{\mathsf{T}}\big(\mathscr{R}^*(\mathscr{R}^*)^{\mathsf{T}}\big)^{-1}\mathscr{R}^*,$$
$$\widehat{\mathscr{P}}_m=\widehat{\mathscr{R}}_m^{\mathsf{T}}\big(\widehat{\mathscr{R}}_m\widehat{\mathscr{R}}_m^{\mathsf{T}}\big)^{-1}\widehat{\mathscr{R}}_m.$$

The use of Lemma B.5 provides

$$\big\|\Pi^*-\widehat{\mathscr{P}}_m\big\|_2\le\sqrt{2}\lambda_m^{-1/2}2\tau(1-4\tau/\rho-4\tau^2/\rho^2)^{-1/2}$$

and we end up as in the proof of Theorem 5.1. □

APPENDIX B

Some matrix inequalities. Let B and B_1 be two $d\times L$-matrices and ρ be some positive number. Define the $d\times d$-matrix P_ρ as

$$P_\rho=(I+\rho^{-2}BB^{\mathsf{T}})^{-1/2}.$$

Here we collect some facts which can be obtained from the inequality

$$\big\|P_\rho(B_1-B)\big\|_2\le\delta \tag{B.1}$$

with some small $\delta\ge 0$. Here and in what follows $\|A\|_2$ denotes the L_2-norm of the matrix A, that is, $\|A\|_2^2=\operatorname{tr}AA^{\mathsf{T}}$, and $\|A\|$ is the sup-norm: $\|A\|=\sup_{v\in\mathbb{R}^d}|Av|/|v|$.

LEMMA B.1. *Condition* (B.1) *implies*

$$\big\|P_\rho\big(BB^{\mathsf{T}}-B_1B_1^{\mathsf{T}}\big)P_\rho\big\|\le 2\rho\delta+\delta^2.$$

PROOF. Since

$$\|P_\rho B\|^2 = \|P_\rho B B^\top P_\rho\| = \big\|(I + \rho^{-2} B B^\top)^{-1} B B^\top\big\| \le \rho^2,$$

(B.1) yields

$$\begin{aligned}\big\|P_\rho(B_1 B_1^\top - B B^\top)P_\rho\big\| &\le 2\big\|P_\rho(B_1 - B)B^\top P_\rho\big\| + \big\|P_\rho(B_1 - B)(B_1 - B)^\top P_\rho\big\| \\ &\le 2\big\|P_\rho(B_1 - B)\big\|_2 \big\|P_\rho B\big\| + \big\|P_\rho(B_1 - B)\big\|_2^2 \\ &\le 2\delta\rho + \delta^2\end{aligned}$$

as required. □

Define also

$$P_{\rho,1} = \big(I + \rho^{-2} B_1 B_1^\top\big)^{-1/2}.$$

LEMMA B.2. *Let B and B_1 satisfies* (B.1) *for some $\delta < \rho/4$. Then*

$$\big\|P_{\rho,1}(B - B_1)\big\|_2 \le \frac{\delta}{\sqrt{1 - 2\delta/\rho - \delta^2/\rho^2}}.$$

PROOF. Let $\alpha = 2\delta/\rho + \delta^2/\rho^2$. By Lemma B.1,

$$\big\|P_\rho P_{\rho,1}^{-2} P_\rho - I\big\| = \rho^{-2}\big\|P_\rho(B B^\top - B_1 B_1^\top)P_\rho\big\| \le \alpha$$

and hence,

$$\begin{aligned}\big\|P_{\rho,1}^{-1} P_\rho\big\|^2 &= \big\|P_\rho P_{\rho,1}^{-2} P_\rho\big\| \le 1 + \alpha, \\ \big\|P_{\rho,1} P_\rho^{-1}\big\|^2 &= \big\|(P_\rho P_{\rho,1}^{-2} P_\rho)^{-1}\big\| \le (1 - \alpha)^{-1}.\end{aligned}$$

Now

$$\begin{aligned}\big\|P_{\rho,1}(B - B_1)\big\|_2 &= \big\|P_{\rho,1} P_\rho^{-1} P_\rho(B - B_1)\big\|_2 \\ &\le \big\|P_{\rho,1} P_\rho^{-1}\big\| \big\|P_\rho(B - B_1)\big\|_2 \le \big\|P_{\rho,1} P_\rho^{-1}\big\| \delta \le \delta(1 - \alpha)^{-1/2}. \quad \square\end{aligned}$$

Next we consider the situation when both matrices B and B_1 are of rank m with some $m < d$. By Π we denote the projector in $\mathbb{R}^d$ onto the subspace $\mathscr{L} = \operatorname{Im} B$. Similarly Π_1 is the projector in $\mathbb{R}^d$ onto the subspace $\mathscr{L}_1 = \operatorname{Im} B_1$.

LEMMA B.3. *Let $d \times L$-matrices B and B_1 of rank m satisfy $\|P_\rho(B - B_1)\|_2 \le \delta$. Then*

$$\|(I - \Pi)B_1\|_2 \le \delta.$$

PROOF. Since P_ρ is the unity operator within the subspace $\mathscr{L}^\perp = \operatorname{Im}(I - \Pi)$, it easily follows that $(I - \Pi)P_\rho = I - \Pi$ (this fact is obvious when $BB^\top$ and hence P_ρ is a diagonal matrix, and the general case can be reduced to that one by an orthogonal transform). Since also $(I-\Pi)B=0$, we derive

$$\begin{aligned} B_1 &= (\Pi + I - \Pi)B_1 \\ &= \Pi B_1 + (I-\Pi)(B_1 - B) \\ &= \Pi B_1 + (I-\Pi)P_\rho(B_1 - B) \end{aligned}$$

so that $\|(I-\Pi)B_1\|_2 \le \|P_\rho(B_1-B)\|_2 \le \delta$. □

LEMMA B.4. *Let Π and Π_1 be two projectors in $\mathbb{R}^d$ of rank $m<d$. Then*

$$\|\Pi_1 - \Pi\|_2 = \sqrt{2}\|\Pi(I-\Pi_1)\|_2.$$

PROOF. Note first that since Π and $I-\Pi$ are orthogonal,

$$\|\Pi_1-\Pi\|_2^2 = \|\Pi_1(I-\Pi)-(I-\Pi_1)\Pi\|_2^2 = \|\Pi_1(I-\Pi)\|_2^2 + \|(I-\Pi_1)\Pi\|_2^2.$$

Now, since $\|\Pi\|_2^2 = \|\Pi_1\|_2^2 = m$, we derive

$$\begin{aligned} \|\Pi_1(I-\Pi)\|_2^2 &= \|\Pi_1\|_2^2 - \|\Pi_1\Pi\|_2^2 = m - \|\Pi_1\Pi\|_2^2, \\ \|(I-\Pi_1)\Pi\|_2^2 &= \|\Pi\|_2^2 - \|\Pi_1\Pi\|_2^2 = m - \|\Pi_1\Pi\|_2^2, \end{aligned}$$

so that $\|\Pi_1(I-\Pi)\|_2 = \|(I-\Pi_1)\Pi\|_2$ and the assertion follows. □

Now let $B^\top B = O\Lambda O^\top$ be the single value decomposition (SVD) of the matrix B where O is the unitary $L\times L$-matrix and Λ is the diagonal matrix with nonincreasing eigenvalues. Then let the $m\times d$ matrix R be constructed using (2.5) with $\mathscr{B}^*$ replaced by B on the base of B; that is, $R=(BO_m)^\top$ where O_m is the block of the first m columns of O. Clearly it holds $|Rv| = |v^\top B|$ for every $v\in\mathbb{R}^d$. Similarly we define R_1 via the SVD of B_1.

The projector Π in $\mathbb{R}^d$ onto the value space of B, can be represented in the form $\Pi = R^\top(RR^\top)^{-1}R$. Similarly $\Pi_1 = R_1^\top(R_1R_1^\top)^{-1}R_1$. Let λ_m denote the smallest eigenvalue of $RR^\top$.

LEMMA B.5. *Let the matrices B, B_1 of rank m satisfy* (B.1) *with some $\delta < \rho/4$. Then the associated projectors Π and Π_1 satisfy*

$$\|\Pi - \Pi_1\|_2 \le \sqrt{2}\lambda_m^{1/2}\delta_1,$$

where $\delta_1 = \delta(1-2\delta/\rho-\delta^2/\rho^2)^{-1/2}$.

PROOF. Condition (B.1) implies by Lemma B.2 $\left\|P_{\rho,1}(B-B_1)\right\|_2 \le \delta_1$ which yields by Lemma B.3,

$$\|R_1(I-\Pi)\|_2 = \|(I-\Pi)B_1\|_2 \le \delta_1.$$

This and Lemma B.4 yields

$$\begin{aligned}\|\Pi - \Pi_1\|_2 &= \sqrt{2}\|\Pi(I - \Pi_1)\|_2 = \sqrt{2}\big\|R^{\mathsf{T}}(RR^{\mathsf{T}})^{-1}R(I - \Pi_1)\big\|_2 \\ &\le \sqrt{2}\big\|R^{\mathsf{T}}(RR^{\mathsf{T}})^{-1}\big\|\,\|R_1(I - \Pi)\|_2 = \delta_1\sqrt{2}\big\|R^{\mathsf{T}}(RR^{\mathsf{T}})^{-1}\big\|.\end{aligned}$$

It remains to note that

$$\big\|R^{\mathsf{T}}(RR^{\mathsf{T}})^{-1}\big\|^2 = \big\|(RR^{\mathsf{T}})^{-1}RR^{\mathsf{T}}(RR^{\mathsf{T}})^{-1}\big\| = \big\|(RR^{\mathsf{T}})^{-1}\big\| = \lambda_m^{-1}$$

and the assertion follows. □

REFERENCES

BRILLINGER, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Eric L. Lehmann* (P. J. Bickel, K. A. Doksum and J. A. Hodges, eds.) 97–114. Wadsworth, Belmont, CA.

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.

COOK, D. (1998). Principal Hessian directions revisited. *J. Amer. Statist. Assoc.* **93** 84–93.

DOKSUM, K. and SAMAROV, A. (1995). Nonparametric estimation of global functionals and a measure of explanatory power of covariates in regression. *Ann. Statist.* **23** 1443–1473.

DONOHO, D. L. and NUSSBAUM, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323.

FAN, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19** 1273–1294.

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

HORN, R. A. and JOHNSON, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press.

HARVILLE, D. A. (1997). *Matrix Analysis from a Statistitian's Perspective*. Springer, New York.

HRISTACHE, M., JUDITSKY, A. and SPOKOINY, V. (2001). Direct estimation of the index coefficients in a single-index model. *Ann. Statist.* **29** 1–32.

HUANG, L.-S. and FAN, J. (1998). Nonparametric estimation of quadratic regression functionals. *Benoulli* **5** 927–949.

IBRAGIMOV, I., NEMIROVSKII, A. and KHASMINSKI, R. (1986). Some problems on noparametric estimation in Gaussian white noise. *Theory Probab. Appl.* **31** 391–406.

KING, M. S. (1997). Local likelihood and local partial likelihood in hazard regression. Ph.D. dissertation, Univ. North Carolina.

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.

LI, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87** 1025–1039.

LI, K.-C. (2000). High dimensional data analysis via the SIR/PHD approach. Available at http://www.stat.ucla.edu/ kcli/sir-PHD.ps.gz.

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

M. HRISTACHE
ENSAI, CAMPUS KER LANN
RUE B. PASCAL
35170 BRUZ
FRANCE
E-MAIL: hristach@ensai.fr

A. JUDITSKY
LMC DOMAINE UNIVERSITAIRE B.P.53
38041 GRENOBLE CEDEX 9
FRANCE
E-MAIL: anatoli.iouditski@inrialpes.fr

J. POLZEHL
V. SPOKOINY
WEIERSTRASS-INSTITUTE
MOHRENSTR. 39
10117 BERLIN
GERMANY
E-MAIL: polzehl@wias-berlin.de
spokoiny@wias-berlin.de