Journal of Machine Learning Research 9 (2008) 1647-1678 Submitted 4/07; Revised 2/08; Published 8/08

A New Algorithm for Estimating the Effective Dimension-Reduction Subspace

Arnak S. Dalalyan ARNAK.DALALYAN@UPMC.FR
Université Paris 6 - Pierre et Marie Curie
Laboratoire de Probabilités, B. C. 188
75252 Paris Cedex 05, France

Anatoly Juditsky ANATOLI.IOUDITSKI@IMAG.FR
University Joseph Fourier of Grenoble LMC-IMAG
51 rue des Mathematiques, B. P. 53
38041 Grenoble Cedex 9, France

Vladimir Spokoiny SPOKOINY@WIAS-BERLIN.DE
Weierstrass Institute for Applied Analysis and Stochastics
Mohrenstrasse 39, 10117 Berlin Germany

Editor: Aapo Hyvarinen

Abstract

The statistical problem of estimating the effective dimension-reduction (EDR) subspace in the multi-index regression model with deterministic design and additive noise is considered. A new procedure for recovering the directions of the EDR subspace is proposed. Many methods for estimating the EDR subspace perform principal component analysis on a family of vectors, say $\hat{\beta}_1,\dots,\hat{\beta}_L$, nearly lying in the EDR subspace. This is in particular the case for the structure-adaptive approach proposed by Hristache et al. (2001a). In the present work, we propose to estimate the projector onto the EDR subspace by the solution to the optimization problem

$$\text{minimize} \quad \max_{\ell=1,\dots,L} \hat{\beta}_\ell^\top (I-A)\hat{\beta}_\ell \quad \text{subject to} \ \ A \in \mathcal{A}_{m^*},$$

where $\mathcal{A}_{m^*}$ is the set of all symmetric matrices with eigenvalues in $[0,1]$ and trace less than or equal to m^*, with m^* being the true structural dimension. Under mild assumptions, $\sqrt{n}$-consistency of the proposed procedure is proved (up to a logarithmic factor) in the case when the structural dimension is not larger than 4. Moreover, the stochastic error of the estimator of the projector onto the EDR subspace is shown to depend on L logarithmically. This enables us to use a large number of vectors $\hat{\beta}_\ell$ for estimating the EDR subspace. The empirical behavior of the algorithm is studied through numerical simulations.

Keywords: dimension-reduction, multi-index regression model, structure-adaptive approach, central subspace

1. Introduction

One of the most challenging problems in modern statistics is to find efficient methods for treating high-dimensional data sets. In various practical situations the problem of predicting or explaining a scalar response variable Y by d scalar predictors $X^{(1)},\dots,X^{(d)}$ arises. For solving this problem one should first specify an appropriate mathematical model and then find an algorithm for estimating that model based on the observed data. In the absence of a priori information on the relationship

between Y and $X = (X^{(1)}, \dots, X^{(d)})$, complex models are to be preferred. Unfortunately, the accuracy of estimation is in general a decreasing function of the model complexity. For example, in the regression model with additive noise and two-times continuously differentiable regression function $f : \mathbb{R}^d \to \mathbb{R}$, the most accurate estimators of f based on a sample of size n have a quadratic risk decreasing as $n^{-4/(4+d)}$ when n becomes large. This rate deteriorates very rapidly with increasing d leading to unsatisfactory accuracy of estimation for moderate sample sizes. This phenomenon is called "curse of dimensionality", the latter term being coined by Bellman (1961).

To overcome the "curse of dimensionality", additional restrictions on the candidates f for describing the relationship between Y and X are necessary. One popular approach is to consider the multi-index model with m^* indices: for some linearly independent vectors ϑ_1, ..., ϑ_{m^*} and for some function $g : \mathbb{R}^{m^*} \to \mathbb{R}$, the relation $f(x) = g(\vartheta_1^\top x, \dots, \vartheta_{m^*}^\top x)$ holds for every $x \in \mathbb{R}^d$. Here and in the sequel the vectors are understood as one column matrices and $M^\top$ denotes the transpose of the matrix M. Of course, such a restriction is useful only if $m^* < d$ and the main argument in favor of using the multi-index model is that for most data sets the underlying structural dimension m^* is substantially smaller than d. Therefore, if the vectors ϑ_1, ..., ϑ_{m^*} are known, the estimation of f reduces to the estimation of g, which can be performed much better because of lower dimensionality of the function g compared to that of f.

Another advantage of the multi-index model is that it postulates that only few linear combinations of the predictors may suffice for "explaining" the response Y. Considering these combinations as new predictors leads to a much simpler model (due to its low dimensionality), which can be successfully analyzed by graphical methods, see Cook and Weisberg (1999) and Cook (1998) for more details.

Throughout this work we assume that we are given n observations $(Y_1, X_1), \dots, (Y_n, X_n)$ from the model

$$Y_i = f(X_i) + \varepsilon_i = g(\vartheta_1^\top X_i, \dots, \vartheta_{m^*}^\top X_i) + \varepsilon_i, \tag{1}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are unobserved errors assumed to be mutually independent zero mean random variables, independent of the design $\{X_i, i \le n\}$.

Since it is unrealistic to assume that $\vartheta_1, \dots, \vartheta_{m^*}$ are known, estimation of these vectors from the data is of high practical interest. When the function g is unspecified, only the linear subspace $\mathcal{S}_\vartheta$ spanned by these vectors may be identified from the sample. This subspace is usually called *index space* or *dimension-reduction (DR) subspace*. Clearly, there are many DR subspaces for a fixed model f. Even if f is observed without error, only the smallest DR subspace, henceforth denoted by $\mathcal{S}$, can be consistently identified. This smallest DR subspace, which is the intersection of all DR subspaces, is called *effective dimension-reduction* (EDR) subspace (Li, 1991) or *central mean subspace* (Cook and Li, 2002). We adopt in this paper the former term, in order to be consistent with Hristache et al. (2001a) and Xia et al. (2002), which are the closest references to our work.

The present work is devoted to studying a new algorithm for estimating the EDR subspace. We call it structural adaptation via maximum minimization (SAMM). It can be regarded as a branch of the structure-adaptive (SA) approach introduced in Hristache et al. (2001b) and Hristache et al. (2001a).

Note that a closely related problem is the estimation of the central subspace (CS), see Cook and Weisberg (1999) for its definition. For model (1) with i.i.d. predictors, the CS coincides with the EDR subspace. Hence, all the methods developed for estimating the CS can potentially be applied in our set-up. We refer to Cook and Li (2002) for background on the difference between the CS and

the central mean subspace and to Cook and Ni (2005) for a discussion of the relationship between different algorithms estimating these subspaces.

There are a number of methods providing an estimator of the EDR subspace in our set-up. These include ordinary least square (Li and Duan, 1989), sliced inverse regression (Li, 1991), sliced inverse variance estimation (Cook and Weisberg, 1991), principal Hessian directions (Li, 1992), graphical regression (Cook, 1998), parametric inverse regression (Bura and Cook, 2001a), SA approach (Hristache et al., 2001a), iterative Hessian transformation (Cook and Li, 2002), minimum average variance estimation (MAVE) (Xia et al., 2002), nonparametric linear smoothing for inverse regression (Bura, 2003), minimum discrepancy approach (Cook and Ni, 2005) marginal high moment regression (Yin and Cook, 2006) density based MAVE and outer product of gradient (Xia, 2007), as well as the refinements using contour-projection (Wang et al., 2008), intraslice covariance estimation (Cook and Ni, 2006) and Lasso shrinkage (Ni et al., 2005; Li, 2007).

All these methods, except SA approach and MAVE, rely on the principle of inverse regression (IR). Therefore they inherit its well known limitations. First, they require a hypothesis on the probabilistic structure of the predictors usually called linearity condition. Second, there is no theoretical justification guaranteeing that these methods estimate the whole EDR subspace and not just a part thereof, see Cook and Li (2004, Section 3.1) and the comments on the third example in Hristache et al. (2001a, Section 4). In the same time, they have the advantage of being simple for implementation and for inference.

The two other methods mentioned above—SA approach and MAVE—have much wider applicability including even time series analysis. The inference for these methods is more involved than that of IR based methods, but SA approach and MAVE need no strong requirements on the design of covariates or on the response variable. Moreover, in many cases they provide more accurate estimates of the EDR subspace (Hristache et al., 2001a; Xia et al., 2002; Xia, 2007).

These arguments, combined with empirical experience, indicate the complementarity of different methods designed to estimate the EDR subspace. It turns out that there is no procedure among those cited above that outperforms all the others in plausible settings. Therefore, a reasonable strategy for estimating the EDR subspace is to execute different procedures and to take a decision after comparing the obtained results. In the case of strong contradictions, collecting additional data is recommended.

The algorithm SAMM we introduce here exploits the fact that the gradient ∇f of the regression function f evaluated at any point $x \in \mathbb{R}^d$ belongs to the EDR subspace. The estimation of the gradient being an ill-posed inverse problem, it is better to estimate some linear combinations of $\nabla f(X_1), \ldots, \nabla f(X_n)$, which still belong to the EDR subspace.

Let L be a positive integer. The main idea behind the algorithm proposed in Hristache et al. (2001a) consists in iteratively estimating L linear combinations $\beta_1, \ldots, \beta_L$ of vectors $\nabla f(X_1), \ldots, \nabla f(X_n)$ and then recovering the EDR subspace from the vectors β_ℓ by running a principal component analysis (PCA). The resulting estimator is proved to be $\sqrt{n}$-consistent provided that L is chosen independently of the sample size n. Unfortunately, if L is small with respect to n, the subspace spanned by the vectors $\beta_1, \ldots, \beta_L$ may cover only a part of the EDR subspace. Therefore, empirical experience advocates for large values of L, even if the desirable feature of $\sqrt{n}$-consistency fails in this case.

The estimator proposed in the present work is designed to provide a remedy for this dissension between the theory and empirical experience. This goal is achieved by introducing a new method of extracting the EDR subspace from the estimators of the vectors $\beta_1, \ldots, \beta_L$. If we think of PCA

as the solution to a minimization problem involving a sum over L terms, see (5) in the next section, then, to some extent, our proposal consists in replacing the sum by the maximum. This motivates the term structural adaptation via maximum minimization. The main advantage of SAMM is that it allows us to deal with the case when L increases polynomially in n and yields an estimator of the EDR subspace which is consistent under a very weak identifiability assumption. In addition, SAMM provides a $\sqrt{n}$-consistent estimator (up to a logarithmic factor) of the EDR subspace when $m^* \leq 4$.

If $m^* = 1$, the corresponding model is referred to as *single-index* regression. There are many methods for estimating the EDR subspace in this case, see Yin and Cook (2005); Delecroix et al. (2006) and the references therein. Note also that the methods for estimating the EDR subspace have often their counterparts in the partially linear regression analysis, see for example Samarov et al. (2005) and Chan et al. (2004).

An interesting problem in the context of dimensionality reduction is the estimation of the true structural dimension m^*. Many approaches exist for constructing estimators of m^*, see (Li, 1991, Section 5), (Xia et al., 2002, Section 2.2), and Bura and Cook (2001b), Bura and Cook (2001a), Bura (2003) and Cook and Li (2004) and the references therein. Here we assume that the structural dimension is known, leaving the development of an extension to the case of unknown m^* for future investigation.

The rest of the paper is organized as follows. We review the structure-adaptive approach and introduce the SAMM procedure in Section 2. Theoretical features including $\sqrt{n}$-consistency of the procedure are stated in Section 3. Section 4 contains an empirical study of the proposed procedure through Monte Carlo simulations. The technical proofs are deferred to Section 5.

2. Structural Adaptation and SAMM

Introduced in Hristache et al. (2001b), the structure-adaptive approach is based on two observations. First, knowing the structural information helps better estimate the model function. Second, improved model estimation contributes to recovering more accurate structural information about the model. These advocate for the following iterative procedure. Start with the null structural information, then iterate the above-mentioned two steps (estimation of the model and extraction of the structure) several times improving the quality of model estimation and increasing the accuracy of structural information during the iteration.

2.1 Purely Nonparametric Local Linear Estimation

When no structural information is available, one can only proceed in a fully nonparametric way. A proper estimation method is based on local linear smoothing (cf. Fan and Gijbels, 1996, for more details): estimators of the function f and its gradient ∇f at a point X_i are given by

$$\begin{aligned}\begin{pmatrix} \hat{f}(X_i) \\ \widehat{\nabla f}(X_i) \end{pmatrix} &= \operatorname*{arg\,min}_{(a,c)^\top} \sum_{j=1}^{n} \big(Y_j - a - c^\top X_{ij}\big)^2 K\big(|X_{ij}|^2/b^2\big) \\ &= \left\{ \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^\top K\left(\frac{|X_{ij}|^2}{b^2}\right) \right\}^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\left(\frac{|X_{ij}|^2}{b^2}\right),\end{aligned}$$

where $X_{ij} = X_j - X_i$, b is a *bandwidth* and $K(\cdot)$ is a univariate kernel supported on $[0,1]$. (For a vector v, $|v|$ stands for its Euclidean norm.) The bandwidth b should be selected so that the ball

with radius b centered at the point of estimation X_i contains at least $d+1$ design points. For large value of d this leads to a large bandwidth and to a strong estimation bias. The goal of the structural adaptation is to diminish this bias using an iterative procedure exploiting the available estimated structural information.

In order to transform these general observations into a concrete procedure, let us describe in the rest of this section how the knowledge of the structure can help to improve the quality of the estimation and how the structural information can be obtained when the function or its estimator is given.

2.2 Model Estimation When an Estimator of $\mathcal{S}$ is Available

Let us start with the case of known $\mathcal{S}$. The function f has the same smoothness as g in the directions of the EDR subspace $\mathcal{S}$ spanned by the vectors $\vartheta_1,\dots,\vartheta_{m^*}$, whereas it is constant (and therefore, infinitely smooth) in all the orthogonal directions. This suggests to apply an anisotropic bandwidth for estimating the model function and its gradient. The corresponding local-linear estimator can be defined by

$$\begin{pmatrix} \hat{f}(X_i) \\ \widehat{\nabla f}(X_i) \end{pmatrix} = \left\{ \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\top} w_{ij}^* \right\}^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} w_{ij}^* \,, \tag{2}$$

with the weights $w_{ij}^* = K(|\Pi^* X_{ij}|^2/h^2)$, where h is some positive real number and Π^* is the orthogonal projector onto the EDR subspace $\mathcal{S}$. This choice of weights amounts to using infinite bandwidth in the directions lying in the orthogonal complement of the EDR subspace.

If only an estimator $\hat{A}$ of the orthogonal projector Π^* is available, a possible strategy is to replace Π^* by $\hat{A}$ in the definitions of the weights w_{ij}^*. This strategy is however too stringent, since it definitely discards the directions belonging to $\hat{\mathcal{S}}^{\perp}$. Being not sure that our information about the structure is exact, it is preferable to define the neighborhoods in a softer way. This is done by setting $w_{ij} = K(X_{ij}^{\top}(I+\rho^{-2}\hat{A})X_{ij}/h^2)$ and by redefining

$$\begin{pmatrix} \hat{f}(X_i) \\ \widehat{\nabla f}(X_i) \end{pmatrix} = \left\{ \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\top} w_{ij} \right\}^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} w_{ij} \,. \tag{3}$$

Here, ρ is a real number from the interval $[0,1]$ measuring the importance attributed to the estimator $\hat{A}$. If we are very confident in our estimator $\hat{A}$, we should choose ρ close to zero.

2.3 Recovering the EDR Subspace from an Estimator of ∇f

Suppose first that the values of the function ∇f at the points X_i are known. Then $\mathcal{S}$ is the linear subspace of $\mathbb{R}^d$ spanned by the vectors $\nabla f(X_i)$, $i=1,\dots,n$. For classifying the directions of $\mathbb{R}^d$ according to the variability of f in each direction and, as a by-product, identifying $\mathcal{S}$, the principal component analysis (PCA) can be used.

Recall that the PCA method is based on the orthogonal decomposition of the matrix $\mathscr{M} = n^{-1}\sum_{i=1}^{n} \nabla f(X_i)\nabla f(X_i)^{\top}$: $\mathscr{M} = O\Lambda O^T$ with an orthogonal matrix O and a diagonal matrix Λ with diagonal entries $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$. Clearly, for the multi-index model with m^*-indices, only the first m^* eigenvalues of $\mathscr{M}$ are positive. The first m^* eigenvectors of $\mathscr{M}$ (or, equivalently, the first m^* columns of the matrix O) define an orthonormal basis in the EDR subspace.

Let L be a positive integer. In Hristache et al. (2001a), a "truncated" matrix $\mathscr{M}_L$ is considered, which coincides with $\mathscr{M}$ if L equals n. Let $\{\psi_\ell, \ell = 1,\dots,L\}$ be a set of vectors of $\mathbb{R}^n$ satisfying the conditions $n^{-1}\sum_{i=1}^n \psi_{\ell,i}\psi_{\ell',i} = \delta_{\ell,\ell'}$ for every $\ell,\ell' \in \{1,\dots,L\}$, with $\delta_{\ell,\ell'}$ being the Kronecker symbol. Define

$$\beta_\ell = n^{-1}\sum_{i=1}^n \nabla f(X_i)\psi_{\ell,i} \tag{4}$$

and $\mathscr{M}_L = \sum_{\ell=1}^L \beta_\ell\beta_\ell^\top$. By the Bessel inequality, it holds $\mathscr{M}_L \preceq \mathscr{M}$. Here and in the sequel, for two symmetric matrices A and B, $A \preceq B$ means that $B-A$ is positive-semidefinite. Moreover, since $\mathscr{M}\mathscr{M}_L = \mathscr{M}_L\mathscr{M}$, any eigenvector of $\mathscr{M}$ is an eigenvector of $\mathscr{M}_L$. Finally, by the Parseval equality, $\mathscr{M}_L = \mathscr{M}$ if $L = n$.

The reason of considering the matrix $\mathscr{M}_L$ instead of $\mathscr{M}$ is that $\mathscr{M}_L$ can be estimated much better than $\mathscr{M}$. In fact, estimators of $\mathscr{M}$ have poor performance for samples of moderate size because of the sparsity of high dimensional data, ill-posedness of the gradient estimation and the non-linear dependence of $\mathscr{M}$ on ∇f. On the other hand, estimation of $\mathscr{M}_L$ reduces to the estimation of L linear functionals of ∇f and may be done with a better accuracy. The obvious limitation of this approach is that it recovers the EDR subspace entirely only if the rank of $\mathscr{M}_L$ coincides with the rank of $\mathscr{M}$, which is equal to m^*. To enhance our chances of seeing the condition $\operatorname{rank}(\mathscr{M}_L) = m^*$ fulfilled, we have to choose L sufficiently large. In practice, L is chosen of the same order as n.

In the case when only an estimator of ∇f is available, the above described method of recovering the EDR directions from an estimator of $\mathscr{M}_L$ has a risk of order $\sqrt{L/n}$ (Hristache et al., 2001a, Theorem 5.1). This fact advocates against using very large values of L. We desire nevertheless to use many linear combinations in order to increase our chances of capturing the whole EDR subspace. To this end, we modify the method of extracting the structural information from the estimators $\hat\beta_\ell$ of vectors β_ℓ.

Let $m \ge m^*$ be an integer. Observe that the estimator $\widetilde\Pi_m$ of the projector Π^* based on the PCA solves the following quadratic optimization problem:

$$\text{minimize } \sum_\ell \hat\beta_\ell^\top(I-\Pi)\hat\beta_\ell \qquad \text{subject to} \qquad \Pi^2 = \Pi, \quad \operatorname{tr}\Pi \le m, \tag{5}$$

where the minimization is carried over the set of all symmetric $(d\times d)$-matrices. The value m^* can be estimated by looking how many eigenvalues of $\widetilde\Pi_m$ are significant. Let $\mathcal{A}_m$ be the set of $(d\times d)$-matrices defined as follows:

$$\mathcal{A}_m = \{A : A = A^\top,\, 0 \preceq A \preceq I,\, \operatorname{tr}A \le m\}.$$

Define $\hat A_m$ as a minimizer of the maximum of the $\hat\beta_\ell^\top(I-A)\hat\beta_\ell$'s instead of their sum:

$$\hat A_m \in \operatorname*{arg\,min}_{A\in\mathcal{A}_m} \max_\ell \hat\beta_\ell^\top(I-A)\hat\beta_\ell. \tag{6}$$

This is a convex optimization problem that can be effectively solved even for a large d although a closed form solution is not known. It is noteworthy that a solution to (6) is not necessarily a projection matrix. In fact, the matrices from $\mathcal{A}_m$ are symmetric positive-semidefinite with eigenvalues between 0 and 1 and not just 0 or 1. This enlargement of the search space guarantees its convexity, which is needed for the algorithm to be tractable. Moreover, as we will show below, the incorporation of (6) in the structural adaptation yields an algorithm having good theoretical and empirical performance.

3. Theoretical Features of SAMM

Throughout this section the true dimension m^* of the EDR subspace is assumed to be known. Thus, we are given n observations $(Y_1,X_1),\dots,(Y_n,X_n)$ from the model

$$Y_i = f(X_i)+\varepsilon_i = g(\vartheta_1^\top X_i,\dots,\vartheta_{m^*}^\top X_i)+\varepsilon_i,$$

where $\varepsilon_1,\dots,\varepsilon_n$ are independent centered random variables. The vectors ϑ_j are assumed to form an orthonormal basis of the EDR subspace entailing thus the representation $\Pi^* = \sum_{j=1}^{m^*}\vartheta_j\vartheta_j^\top$. In what follows, we mainly consider deterministic design. Nevertheless, the results hold in the case of random design as well, provided that the errors are independent of $X_1,\dots,X_n$. Henceforth, without loss of generality we assume that $|X_i|\le 1$ for any $i=1,\dots,n$.

3.1 Description of the Algorithm

The structure-adaptive algorithm with maximum minimization consists of following steps.

a) Specify positive real numbers a_ρ, a_h, ρ_1 and h_1. Choose an integer L and select a set $\{\psi_\ell,\ \ell\le L\}$ of vectors from $\mathbb{R}^n$ verifying $|\psi_\ell|^2 = n$.

b) Set $k=1$ and $\hat A_0 = 0$.

c) Define the estimators $\widehat{\nabla f}_k(X_i)$ for $i=1,\dots,n$ by formula (3) with $w_{ij} = K\big(X_{ij}^\top(I+\rho_k^{-2}\hat A_{k-1})X_{ij}/h_k^2\big)$. Set

$$\hat\beta_{\ell,k} = \frac{1}{n}\sum_{i=1}^n \widehat{\nabla f}_k(X_i)\psi_{\ell,i},\qquad \ell=1,\dots,L,$$

where $\psi_{\ell,i}$ is the ith coordinate of ψ_ℓ.

d) Define the new value $\hat A_k$ by $\hat A_k \in \arg\min_{A\in\mathcal{A}_{m^*}}\max_\ell \hat\beta_{\ell,k}^\top(I-A)\hat\beta_{\ell,k}$.

e) Set $\rho_{k+1} = a_\rho\cdot\rho_k$, $h_{k+1} = a_h\cdot h_k$ and increase k by one.

f) Stop if $\rho_k<\rho_{\min}$ or $h_k>h_{\max}$, otherwise continue with the step c).

Let $k(n)$ be the total number of iterations. We denote by $\widehat\Pi_n$ the orthogonal projection onto the space spanned by the eigenvectors of $\hat A_{k(n)}$ corresponding to the m^* largest eigenvalues. The estimator of the EDR subspace is then the image of $\widehat\Pi_n$.

Both $\hat A_{k(n)}$ and $\widehat\Pi_n$ are estimators of the projector onto $\mathcal{S}$. Our theoretical results are stated for the estimator $\widehat\Pi_n$, but similar results are valid for $\hat A_{k(n)}$, too. The numerical simulations we made showed that these two estimators have comparable performances.

The described algorithm requires the specification of the parameters ρ_1, h_1, a_ρ and a_h, as well as the choice of the set of vectors $\{\psi_\ell\}$. In what follows we use the values

$$\begin{array}{ccc}\rho_1 = 1, & \rho_{\min} = n^{-1/(3\vee m^*)}, & a_\rho = e^{-1/2(3\vee m^*)},\\ h_1 = C_0 n^{-1/(4\vee d)}, & h_{\max} = 2\sqrt{d}, & a_h = e^{1/2(4\vee d)}.\end{array}$$

This choice of input parameters is up to some minor modifications the same as in Hristache et al. (2001b), Hristache et al. (2001a) and Samarov et al. (2005), and is based on the trade-off between

the bias and the variance of estimation. The constant C_0 will be chosen in a design-dependent manner taking into account the fact that the local neighborhoods used in (2) should contain enough design points to entail the consistency of the estimator. The choice of L and that of vectors ψ_ℓ will be discussed in Section 4.

3.2 Assumptions

Prior to stating rigorous theoretical results we need to introduce a set of assumptions. From now on, we use the notation I for the identity matrix of dimension d, $\|A\|^2$ for the largest eigenvalue of $A^\top A$ and $\|A\|_2$ for the Frobenius norm of A (square root of the sum of squares of elements of A).

We start with the smoothness assumption ensuring the adequacy of the local linear approximation of the regression function.

(A1) There exists a positive real C_g such that $|\nabla g(x)| \le C_g$ and $|g(x) - g(x') - (x-x')^\top \nabla g(x)| \le C_g |x-x'|^2$ for every $x, x' \in \mathbb{R}^{m^*}$.

Unlike the smoothness assumption, the assumptions on the identifiability of the model and the regularity of design are more involved and specific to our algorithm. The formal statements read as follows.

(A2) Let the vectors $\beta_\ell \in \mathbb{R}^d$ be defined by (4) and let $\mathcal{B}^* = \big\{\bar\beta = \sum_{\ell=1}^L c_\ell \beta_\ell : \sum_{\ell=1}^L |c_\ell| \le 1\big\}$. There exist vectors $\bar\beta_1, \dots, \bar\beta_{m^*} \in \mathcal{B}^*$ and constants $\mu_1, \dots, \mu_{m^*}$ such that

$$\Pi^* \preceq \sum_{k=1}^{m^*} \mu_k \bar\beta_k \bar\beta_k^\top. \tag{7}$$

We denote $\mu^* = \mu_1 + \ldots + \mu_{m^*}$.

Remark 1 *Assumption (A2) implies that the subspace $\mathcal{S} = \mathrm{Im}(\Pi^*)$ is the smallest DR subspace, therefore it is the EDR subspace. Indeed, for any DR subspace $\mathcal{S}'$, the gradient $\nabla f(X_i)$ belongs to $\mathcal{S}'$ for every i. Therefore $\beta_\ell \in \mathcal{S}'$ for every $\ell \le L$ and $\mathcal{B}^* \subset \mathcal{S}'$. Thus, for every β° from the orthogonal complement $\mathcal{S}'^\perp$, it holds $|\Pi^* \beta^\circ|^2 \le \sum_k \mu_k |\bar\beta_k^\top \beta^\circ|^2 = 0$. Therefore $\mathcal{S}'^\perp \subset \mathcal{S}^\perp$ implying thus the inclusion $\mathcal{S} \subset \mathcal{S}'$.*

Lemma 2 *If the family $\{\psi_\ell\}$ spans $\mathbb{R}^n$, then assumption (A2) is always satisfied with some μ_k (that may depend on n).*

Proof Set $\Psi = (\psi_1, \dots, \psi_L) \in \mathbb{R}^{n \times L}$, $\nabla f = (\nabla f(X_1), \dots, \nabla f(X_n)) \in \mathbb{R}^{d \times n}$ and write the $d \times L$ matrix $B = (\beta_1, \dots, \beta_L)$ in the form $\nabla f \cdot \Psi$. Recall that if M_1, M_2 are two matrices such that $M_1 \cdot M_2$ is well defined and the rank of M_2 coincides with the number of lines in M_2, then $\mathrm{rank}(M_1 \cdot M_2) = \mathrm{rank}(M_1)$. This implies that $\mathrm{rank}(B) = m^*$ provided that $\mathrm{rank}(\Psi) = n$, which amounts to $\mathrm{span}(\{\psi_\ell\}) = \mathbb{R}^n$.

Let now $\tilde\beta_1, \dots, \tilde\beta_{m^*}$ be a linearly independent subfamily of $\{\beta_\ell, \ell \le L\}$. Then the m^*th largest eigenvalue $\lambda_{m^*}(\tilde{\mathscr{M}})$ of the matrix $\tilde{\mathscr{M}} = \sum_{k=1}^{m^*} \tilde\beta_k \tilde\beta_k^\top$ is strictly positive. Moreover, if $v_1, \dots, v_{m^*}$ are the eigenvectors of $\tilde{\mathscr{M}}$ corresponding to the eigenvalues $\lambda_1(\tilde{\mathscr{M}}) \ge \ldots \ge \lambda_{m^*}(\tilde{\mathscr{M}}) > 0$, then

$$\Pi^* = \sum_{k=1}^{m^*} v_k v_k^\top \preceq \frac{1}{\lambda_{m^*}(\tilde{\mathscr{M}})} \sum_{k=1}^{m^*} \lambda_k(\tilde{\mathscr{M}}) v_k v_k^\top = \lambda_{m^*}(\tilde{\mathscr{M}})^{-1} \tilde{\mathscr{M}} = \lambda_{m^*}(\tilde{\mathscr{M}})^{-1} \sum_{k=1}^{m^*} \tilde\beta_k \tilde\beta_k^\top.$$

Hence, inequality (7) is fulfilled with $\mu_k = \lambda_{m^*}(\tilde{\mathscr{M}})^{-1}$ for every $k = 1, \dots, m^*$. ■

These arguments show that the identifiability assumption (A2) is fairly weak. In fact, since we always choose $\{\psi_\ell\}$ so that $\mathrm{span}(\{\psi_\ell\}) = \mathbb{R}^n$, (A2) amounts to requiring that the value μ^* remains bounded when n increases. This assumption is much weaker than the coverage assumption under which the consistency of the inverse regression based methods is proved.

Let us proceed with the assumption on the design regularity. Define $P_1^* = I$ and $P_k^* = (I + \rho_k^{-2}\Pi^*)^{-1/2}$ for every $k \geq 2$. Next, set $Z_{ij}^{(k)} = (h_k P_k^*)^{-1} X_{ij}$ and for any $d \times d$ matrix U put $w_{ij}^{(k)}(U) = K\big((Z_{ij}^{(k)})^\top U Z_{ij}^{(k)}\big)$, $\overline{w}_{ij}^{(k)}(U) = K'\big((Z_{ij}^{(k)})^\top U Z_{ij}^{(k)}\big)$, $N_i^{(k)}(U) = \sum_j w_{ij}^{(k)}(U)$ and

$$V_i^{(k)}(U) = \sum_{j=1}^n \begin{pmatrix} 1 \\ Z_{ij}^{(k)} \end{pmatrix} \begin{pmatrix} 1 \\ Z_{ij}^{(k)} \end{pmatrix}^\top w_{ij}^{(k)}(U).$$

(A3) For some positive constants $C_V, C_K, C_{K'}, C_w$ and for some $\alpha \in]0, 1/2]$, the inequalities

$$\begin{aligned}
&\|V_i^{(k)}(U)^{-1}\| N_i^{(k)}(U) \leq C_V, \qquad i = 1, \dots, n, \\
&\sum_{i=1}^n w_{ij}^{(k)}(U)/N_i^{(k)}(U) \leq C_K, \qquad j = 1, \dots, n, \\
&\sum_{i=1}^n |\overline{w}_{ij}^{(k)}(U)|/N_i^{(k)}(U) \leq C_{K'}, \qquad j = 1, \dots, n, \\
&\sum_{j=1}^n |\overline{w}_{ij}^{(k)}(U)|/N_i^{(k)}(U) \leq C_w \qquad i = 1, \dots, n,
\end{aligned}$$

hold for every $k \leq k(n)$ and for every $d \times d$ matrix U verifying $\|U - I\|_2 \leq \alpha$.

Remark 3 *Note that in (A3) we implicitly assumed that the matrices $V_i^{(k)}$ are invertible, which may be true only if any neighborhood $E^{(k)}(X_i) = \{x : |(I + \rho_k^{-2}\Pi^*)^{-1/2}(X_i - x)| \leq h_k\}$ contains at least d design points different from X_i. The parameters h_1, ρ_1, a_ρ and a_h are chosen so that the volume of ellipsoids $E^{(k)}(X_i)$ is a non-decreasing function of k and $Vol(E^{(1)}(X_i)) = C_0/n$. Therefore, from theoretical point of view, if the design is random with positive density on $[0,1]^d$, it is easy to check that for a properly chosen constant C_0, assumption (A3) is satisfied with a probability close to one. In applications, we define h_1 as the smallest real such that $\min_{i=1,\dots,n} \# E^{(1)}(X_i) = d + 1$ and add to the matrix*

$$\sum_{j=1}^n \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^\top w_{ij},$$

involved in the definition (3), a small full-rank matrix to be sure that the resulting matrix is invertible, see Section 4. This observation also implies that the SAMM procedure can not be applied in the case where the sample size n is smaller than or equal to the dimension d of predictors.

(A4) The errors $\{\varepsilon_i, i \leq n\}$ are centered Gaussian with variance σ^2.

3.3 Risk Bounds for the Projection Matrix Estimation

In this section, we present the main result of the paper assessing the quality of the estimator $\widehat{\Pi}_n$ of the projection matrix Π^* in the asymptotics of large samples. To this end, we assume that the kernel K used in (3) is chosen to be continuous, positive and vanishing outside the interval $[0,1]$. The vectors ψ_ℓ are assumed to verify

$$\max_{\ell=1,\dots,L}\max_{i=1,\dots,n}|\psi_{\ell,i}|<\bar{\psi}, \tag{8}$$

for some constant $\bar{\psi}$ independent of n. In the sequel, we denote by $C,C_1,\dots$ some constants depending only on $m^*,\mu^*,C_g,C_V,C_K,C_{K'},C_w$ and $\bar{\psi}$.

Theorem 4 *Let assumptions **(A1)-(A4)** be fulfilled. There exists $C>0$ such that for any $z\in]0,2\sqrt{\log(nL)}]$ and for sufficiently large values of n, it holds*

$$\mathbf{P}\left(\sqrt{\mathrm{tr}(I-\widehat{\Pi}_n)\Pi^*}>Cn^{-\frac{2}{3\vee m^*}}t_n^2+\frac{2\sqrt{\mu^*}zc_0\sigma}{\sqrt{n(1-\zeta_n)}}\right)\le Lze^{-\frac{z^2-1}{2}}+\frac{3k(n)-5}{n},$$

where $c_0=\bar{\psi}\sqrt{dC_KC_V}$, $t_n=O(\sqrt{\log(Ln)})$ and $\zeta_n=O(t_n n^{-\frac{1}{6\vee m^*}})$.

Corollary 5 *Under the assumptions of Theorem 4, for sufficiently large n, it holds*

$$\mathbf{P}\left(\|\widehat{\Pi}_n-\Pi^*\|_2>Cn^{-\frac{2}{3\vee m^*}}t_n^2+\frac{2\sqrt{2\mu^*}zc_0\sigma}{\sqrt{n(1-\zeta_n)}}\right)\le Lze^{-\frac{z^2-1}{2}}+\frac{3k(n)-5}{n}$$

$$\mathbf{E}(\|\widehat{\Pi}_n-\Pi^*\|_2)\le C\left(n^{-\frac{2}{3\vee m^*}}t_n^2+\frac{\sqrt{\log nL}}{\sqrt{n}}\right)+\frac{\sqrt{2m^*}(3k(n)-5)}{n}.$$

Proof Easy algebra yields

$$\begin{aligned}\|\widehat{\Pi}_n-\Pi^*\|_2^2=\mathrm{tr}(\widehat{\Pi}_n-\Pi^*)^2&=\mathrm{tr}\,\widehat{\Pi}_n^2-2\,\mathrm{tr}\,\widehat{\Pi}_n\Pi^*+\mathrm{tr}\,\Pi^*\\&\le\mathrm{tr}\,\widehat{\Pi}_n+m^*-2\,\mathrm{tr}\,\widehat{\Pi}_n\Pi^*\le 2m^*-2\,\mathrm{tr}\,\widehat{\Pi}_n\Pi^*.\end{aligned}$$

The equality $\mathrm{tr}\,\Pi^*=m^*$ and the linearity of the trace operator complete the proof of the first inequality. The second inequality can be derived from the first one by standard arguments in view of the inequality $\|\widehat{\Pi}_n-\Pi^*\|_2^2\le 2m^*$. ■

According to these results, for $m^*\le 4$, the estimator of the orthogonal projector onto $\mathcal{S}$ provided by the SAMM procedure is $\sqrt{n}$-consistent up to a logarithmic factor. This rate of convergence is known to be optimal for a broad class of semiparametric problems, see Bickel et al. (1998) for a detailed account on the subject.

Remark 6 *The inspection of the proof of Theorem 4 shows that the factor t_n^2 multiplying the "bias" term $n^{-2/(3\vee m^*)}$ disappears when $m^*>3$.*

Remark 7 *The same rate of convergence remains valid in the case when the errors are not necessarily identically distributed Gaussian random variables, but have a bounded exponential moment (uniformly in n). This can be proved along the lines of Proposition 14, see Section 5.*

3.4 Risk Bound for the Estimator of a Basis of the EDR Subspace

The main result of this paper stated in the preceding subsection provides a risk bound for the estimator $\widehat{\Pi}_n$ of Π^*, the orthogonal projector onto $\mathcal{S}$. As a by-product of this result, we show in this section that a similar risk bound holds also for the estimator of an orthonormal basis of $\mathcal{S}$. This means that for an arbitrarily chosen orthonormal basis of the estimated EDR subspace $\hat{\mathcal{S}} = \mathrm{Im}(\widehat{\Pi}_n)$, there is an orthonormal basis of the true EDR subspace $\mathcal{S}$ such that the matrices built from these bases are close in Frobenius norm with a probability tending to one.

Proposition 8 *Let the assumptions of Theorem 4 be fulfilled. For any orthonormal basis $\hat{\vartheta}_1, \ldots, \hat{\vartheta}_{m^*}$ of the estimated EDR subspace $\hat{\mathcal{S}} = \mathrm{Im}(\widehat{\Pi}_n)$ there exists an orthonormal basis $\vartheta_1, \ldots, \vartheta_{m^*}$ of the true EDR subspace $\mathcal{S} = \mathrm{Im}(\Pi^*)$ such that, for sufficiently large n, it holds*

$$\mathbf{P}\left(\|\widehat{\Theta}_n - \Theta\|_2 > C n^{-\frac{2}{3\vee m^*}} t_n^2 + \frac{2(\sqrt{m^*}+1)\sqrt{2\mu^*} z c_0 \sigma}{\sqrt{n(1-\zeta_n)}} \right) \leq L z e^{-\frac{z^2-1}{2}} + \frac{3k(n)-5}{n} ,$$

where $\widehat{\Theta}_n$ (resp. Θ) is the $d \times m^*$ matrix whose jth column is $\hat{\vartheta}_j$ (resp. ϑ_j).

Proof Using the singular value decomposition, we write $\Pi^* \widehat{\Theta}_n = U \Lambda V^\top$, where U and V are orthogonal matrices and Λ is a diagonal matrix. Let us denote by λ_j, u_j, v_j respectively the jth diagonal entry of Λ, the jth column of U and the jth column of V. Since $\Pi^* \widehat{\Theta}_n v_j = \lambda_j u_j$, we have $\lambda_j = |\lambda_j u_j| = |\Pi^* \widehat{\Theta}_n v_j| \leq 1$. On the other hand,

$$\lambda_j = |\Pi^* \widehat{\Theta}_n v_j| \geq |\widehat{\Theta}_n v_j| - |(\widehat{\Pi}_n - \Pi^*)\widehat{\Theta}_n v_j| \geq 1 - \|\widehat{\Pi}_n - \Pi^*\|,$$

where we used the fact that $|\widehat{\Theta}_n v_j|^2 = v_j^\top \widehat{\Theta}_n^\top \widehat{\Theta}_n v_j = v_j^\top v_j = 1$. Let us define the matrix Θ as follows: $\Theta = U I_{d \times m^*} V^\top$, where $I_{d \times m^*}$ is the $d \times m^*$ diagonal matrix with all diagonal entries equal to one. One easily checks that Θ is orthogonal, that is $\Theta^\top \Theta = I_{m^*}$. Moreover, we have $\Theta = \Pi^* \hat{\Theta}_n V \Lambda^- V^\top$, where Λ^- is the $m^* \times m^*$ diagonal matrix having λ_j^{-1} as jth diagonal entry. Note that if the norm of $\hat{\Pi}_n - \Pi^*$ is less than 1, the eigenvalues λ_j are strictly positive. In this case, Λ^- is well defined and we obviously have $\Pi^* \Theta = \Theta$. Thus the columns of Θ form an orthonormal basis of $\mathrm{Im}(\Pi^*)$. Furthermore, we have

$$\begin{aligned}
\|\widehat{\Theta}_n - \Theta\|_2 &\leq \|\Theta - \Pi^* \widehat{\Theta}_n\|_2 + \|(\Pi^* - \widehat{\Pi}_n)\Theta\|_2 \\
&\leq \|U(I_{d \times m^*} - \Lambda) V^\top\|_2 + \|\Pi^* - \widehat{\Pi}_n\|_2 \\
&\leq (\sum_{j=1}^{m^*} (\lambda_j - 1)^2)^{1/2} + \|\Pi^* - \widehat{\Pi}_n\|_2 \\
&\leq (\sqrt{m^*} + 1)\|\Pi^* - \widehat{\Pi}_n\|_2,
\end{aligned}$$

provided that $\|\Pi^* - \widehat{\Pi}_n\| < 1$. This implies that for every $\mathsf{d} \in (0,1)$ the event $\{\|\Pi^* - \widehat{\Pi}_n\|_2 \leq \mathsf{d}\}$ is included in $\{\|\widehat{\Theta}_n - \Theta\|_2 \leq (\sqrt{m^*}+1)\mathsf{d}\}$. By virtue of this inclusion, the assertion of the proposition follows from Corollary 5. ■

4. Simulation Results

The aim of this section is to demonstrate on several examples how the performance of the algorithm SAMM depends on the sample size n, the dimension d and the noise level σ. We also show that our procedure can be successfully applied in autoregressive models. Many unreported results show that in most situations the performance of SAMM is comparable to the performance of SA approach based on PCA and to that of MAVE. A thorough comparison of the numerical virtues of these methods being out of scope of this paper, we simply show on some examples that SAMM may substantially outperform MAVE in the case of large "bias". Our results also show that SAMM and MAVE provide more accurate estimates of the EDR subspace than inverse regression based methods : inverse regression based on Minimum Discrepancy Approach (MDA) introduced in Cook and Ni (2005) and Sliced Average Variance Estimation (SAVE) of Cook and Weisberg (1991). In all simulations for inverse regression based methods, the number of slices is chosen to minimise the risk.

The computer code of the procedure SAMM is distributed freely, it can be downloaded from `http://code.google.com/p/samm07/`. It requires the MATLAB packages SDPT3 and YALMIP. We are grateful to Professor Yingcun Xia for making the computer code of MAVE available to us.

To obtain higher stability of the algorithm, we preliminarily standardize the response Y and the predictors $X^{(j)}$. More precisely, we deal with $\tilde{Y}_i = Y_i/\sigma_Y$ and $\tilde{X} = \mathrm{diag}(\Sigma_X)^{-1/2}X$, where σ_Y^2 is the empirical variance of Y, Σ_X is the empirical covariance matrix of X and $\mathrm{diag}(\Sigma_X)$ is the $d \times d$ matrix obtained from Σ_X by replacing the off-diagonal elements by zero. To preserve consistency, we set $\tilde{\beta}_{\ell,k(n)} = \mathrm{diag}(\Sigma_X)^{-1/2}\hat{\beta}_{\ell,k(n)}$, where $\hat{\beta}_{\ell,k(n)}$ is the last-step estimate of β_ℓ, and define $\widehat{\Pi}_{k(n)}$ as the solution to (6) with $\hat{\beta}_\ell$ replaced by $\tilde{\beta}_{\ell,k(n)}$. Furthermore, we add the small full-rank matrix I_{d+1}/n to $\sum_{j=1}^n \binom{1}{X_{ij}}\binom{1}{X_{ij}}^\top w_{ij}$ in (3).

In all examples presented below the number of replications is $N = 250$; for each replication, a new sample of the design and the error vector $(\varepsilon_1, \ldots, \varepsilon_n)$ has been generated at random. The mean loss $\overline{\mathrm{er}_N} = \frac{1}{N}\sum_j \mathrm{er}_j$ and the standard deviation $\sqrt{\frac{1}{N}\sum_j (\mathrm{er}_j - \overline{\mathrm{er}_N})^2}$ are reported, where $\mathrm{er}_j = \|\widehat{\Pi}^{(j)} - \Pi^*\|$ with $\widehat{\Pi}^{(j)}$ being the estimator of Π^* for jth replication.

4.1 Choice of $\{\psi_\ell, \ell \le L\}$

The set $\{\psi_\ell\}$ plays an essential role in the algorithm. The optimal choice of this set is an important issue that needs further investigation. We content ourselves with giving one particular choice which agrees with theory and leads to nice empirical results.

Let $\mathfrak{S}_j$, $j \le d$, be the permutation of the set $\{1, \ldots, n\}$ satisfying $X^{(j)}_{\mathfrak{S}_j(1)} \le \ldots \le X^{(j)}_{\mathfrak{S}_j(n)}$. Let $\mathfrak{S}_j^{-1}$ be the inverse of $\mathfrak{S}_j$, that is, $\mathfrak{S}_j(\mathfrak{S}_j^{-1}(k)) = k$ for every $k = 1, \ldots, n$. Define $\{\psi_\ell\}$ as the set of vectors

$$\left\{ \begin{pmatrix} \cos\big(\frac{2\pi(k-1)\mathfrak{S}_j^{-1}(1)}{n}\big), \ldots, \cos\big(\frac{2\pi(k-1)\mathfrak{S}_j^{-1}(n)}{n}\big) \end{pmatrix}^\top \\ \begin{pmatrix} \sin\big(\frac{2\pi k\mathfrak{S}_j^{-1}(1)}{n}\big), \ldots, \sin\big(\frac{2\pi k\mathfrak{S}_j^{-1}(n)}{n}\big) \end{pmatrix}^\top \quad , k \le [n/2],\ j \le d \right\}$$

normalized to satisfy $\sum_{i=1}^n \psi_{\ell,i}^2 = n$ for every ℓ. It is easily seen that these vectors satisfy conditions (8) and $\mathrm{span}(\{\psi_\ell\}) = \mathbb{R}^n$, so the conclusion of Lemma 2 holds. Above, $[n/2]$ is the integer part of $n/2$ and k and j are positive integers.

The idea behind the above described choice of vectors ψ_ℓ is the following: if the design is uniformly distributed in $[0,1]^d$ and $H(x)$ is a function $\mathbb{R}^d \to \mathbb{R}$ depending only on one coordinate of x, the projections of the vector $g = (H(X_1),\dots,H(X_n))^\top$ on some of directions ψ_ℓ are nearly equal to the Fourier coefficients of H. Indeed, for n odd and for every fixed j, the vectors $\{e_{k,j} = \big(\phi_k(\mathfrak{S}_j^{-1}(1)/n), \phi_k(\mathfrak{S}_j^{-1}(2)/n),\dots,\phi_k(\mathfrak{S}_j^{-1}(n)/n)\big)^\top; 1 \le k \le n\}$ with $\{\phi_k\}$ being the trigonometric basis (that is $\phi_{2p}(x) = \sqrt{2}\sin(2\pi px)$ and $\phi_{2p+1}(x) = \sqrt{2}\cos(2\pi px)$ for every $p \in \mathbb{N}$) form an orthonormal basis of $\mathbb{R}^n$. Therefore, for any function H from $\mathbb{R}^d$ to $\mathbb{R}$, which depends exclusively on the j^{th} coordinate $x^{(j)}$ of x, one has $g^\top e_{kj} = \sum_{i=1}^n H_0(X_i^{(j)})\phi_k(\mathfrak{S}_j^{-1}(i)/n) = \sum_{i=1}^n H_0(X^{(j)}_{\mathfrak{S}_j(i)})\phi_k(i/n)$ for some function $H_0 : \mathbb{R} \to \mathbb{R}$. Since for a sample $X_1^{(j)},\dots,X_n^{(j)}$ drawn from uniform distribution in $[0,1]$ the order statistics are nearly equal to i/n, we get $\frac{1}{n} g^\top e_{kj} \approx \frac{1}{n}\sum_{i=1}^n H_0(i/n)\phi_k(i/n) \approx \langle H_0, \phi_k\rangle_{L^2[0,1]}$. Note that although this explanation is valid only for uniform design and a function H depending only on one coordinate, empirical results show that this choice leads to satisfactory results in more general situations.

4.2 Example 1 (Single-index)

We set $d = 5$ and $f(x) = g(\vartheta^\top x)$ with

$$g(t) = 4|t|^{1/2}\sin^2(\pi t), \quad \text{and} \quad \vartheta = (1/\sqrt{5}, 2/\sqrt{5}, 0, 0, 0)^\top \in \mathbb{R}^5.$$

We ran SAMM, MAVE, MDA and SAVE procedures on the data generated by the model

$$Y_i = f(X_i) + 0.5\cdot\varepsilon_i,$$

where the design X is such that the coordinates $(X_i^{(j)}, j \le 5, i \le n)$ are i.i.d. uniform in $[-1,1]$, and the errors ε_i are i.i.d. standard Gaussian independent of the design.

Table 1 contains the average loss for different values of the sample size n for the first step estimator by SAMM, the final estimator provided by SAMM and the estimators based on MAVE, MDA and SAVE. The first observation is that inverse regression based methods are not consistent in this case. We plot in Figure 1 the average loss normalized by the square root of the sample size n versus n. It is clearly seen that the iterative procedure improves considerably the quality of estimation and that the final estimator provided by SAMM is $\sqrt{n}$-consistent. In this example, MAVE method often fails to recover the EDR subspace. However, the number of failures decreases very rapidly with increasing n. This is the reason why the curve corresponding to MAVE in Figure 1 decreases with a strong slope.

4.3 Example 2 (Double-index)

For $d \ge 2$ we set $f(x) = g(\vartheta^\top x)$ with

$$g(x) = (x_1 - x_2^3)(x_1^3 + x_2);$$

and $\vartheta_1 = (1,0,\dots,0) \in \mathbb{R}^d$, $\vartheta_2 = (0,1,\dots,0) \in \mathbb{R}^d$. We ran SAMM, MAVE, MDA and SAVE procedures on the data generated by the model

$$Y_i = f(X_i) + 0.1\cdot\varepsilon_i, \qquad i = 1,\dots,300,$$

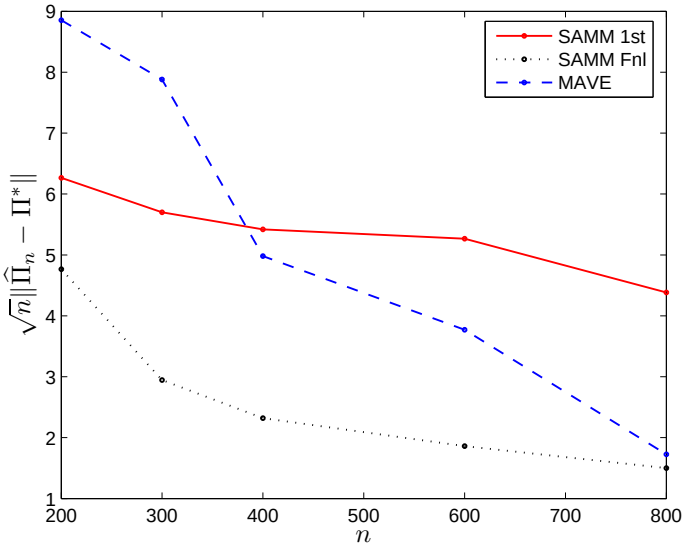

Figure 1: Average loss multiplied by $\sqrt{n}$ versus n for the first step (solid line) and the final (dotted line) estimators provided by SAMM and for the estimator by MAVE (dashed line) in Example 1.

n	200	300	400	600	800
SAMM, 1st	0.443	0.329	0.271	0.215	0.155
	(.211)	(.120)	(.115)	(.095)	(.079)
SAMM, Fnl	0.337	0.170	0.116	0.076	0.053
	(.273)	(.147)	(.104)	(.054)	(.031)
MAVE	0.626	0.455	0.249	0.154	0.061
	(.363)	(.408)	(.342)	(.290)	(.161)
MDA	0.882	0.885	0.890	0.885	0.882
	(.144)	(.141)	(.130)	(.142)	(.148)
SAVE	0.857	0.847	0.832	0.818	0.782
	(.145)	(.144)	(.154)	(.168)	(.169)

Table 1: Average loss $\|\widehat{\Pi} - \Pi^*\|$ of the estimators obtained by SAMM, MAVE, MDA and SAVE procedures in Example 1. The standard deviation is given in parentheses.

where the design X is such that the coordinates $(X_i^{(j)}, j \le d, i \le n)$ are i.i.d. uniform in $[-40,40]$, and the errors ε_i are i.i.d. standard Gaussian independent of the design. The results of simulations for different values of d are reported in Table 2.

As expected, we found that (cf. Figure 2) the quality of SAMM, as well as the quality of SAVE, deteriorated linearly in d as d increased. This agrees with our theoretical results. It should be noted that in this case MAVE and MDA fail to find the EDR subspace.

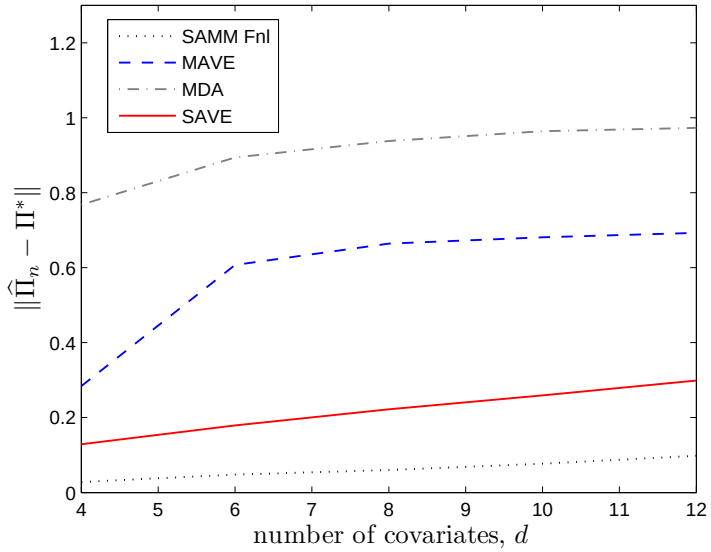

Figure 2: Average loss versus d for the estimators provided by SAMM (dotted line), by MAVE (dashed line), by MDA (dash-dot line) and by SAVE (solid line) in Example 2.

d	4	6	8	10	12
SAMM 1st	0.154	0.242	0.296	0.365	0.421
	(.063)	(.081)	(.071)	(.087)	(.095)
SAMM, Fnl	0.028	0.048	0.060	0.077	0.098
	(.011)	(.020)	(.021)	(.026)	(.037)
MAVE	0.284	0.607	0.664	0.681	0.693
	(.147)	(.073)	(.052)	(.054)	(.044)
MDA	0.768	0.894	0.938	0.964	0.973
	(.232)	(.142)	(.095)	(.062)	(.049)
SAVE	0.129	0.179	0.222	0.259	0.299
	(.048)	(.047)	(.050)	(.058)	(.071)

Table 2: Average loss $\|\widehat{\Pi} - \Pi^*\|$ of the estimators obtained by SAMM, MAVE and MDA procedures in Example 2. The standard deviation is given in parentheses.

4.4 Example 3

For $d = 5$ we set $f(x) = g(\vartheta^\top x)$ with

$$g(x) = (1+x_1)(1+x_2)(1+x_3)$$

and $\vartheta_1 = (1,0,0,0,0)$, $\vartheta_2 = (0,1,0,0,0)$, $\vartheta_3 = (0,0,1,0,0)$. We ran SAMM, MAVE, MDA and SAVE procedures on the data generated by the model

$$Y_i = f(X_i) + \sigma \cdot \varepsilon_i, \qquad i = 1, \ldots, 250,$$

σ	200	150	100	50	25	10
SAMM 1st	0.227	0.177	0.141	0.119	0.113	0.106
	(.092)	(.075)	(.055)	(.051)	(.048)	(.043)
SAMM, Fnl	0.125	0.084	0.057	0.039	0.034	0.030
	(.076)	(.037)	(.026)	(.019)	(.021)	(.018)
MAVE	0.103	0.087	0.073	0.062	0.063	0.059
	(.041)	(.035)	(.027)	(.023)	(.024)	(.023)
MDA	0.854	0.850	0.867	0.862	0.858	0.873
	(.167)	(.173)	(.157)	(.159)	(.171)	(.159)
SAVE	0.510	0.511	0.496	0.505	0.496	0.490
	(.208)	(.204)	(.207)	(.197)	(.196)	(.199)

Table 3: Average loss $\|\widehat{\Pi} - \Pi^*\|$ of the estimators obtained by SAMM and MAVE procedures in Example 3. The standard deviation is given in parentheses.

where the design X is such that the coordinates $(X_i^{(j)}, j \le d, i \le n)$ are i.i.d. uniform in $[0,20]$, and the errors ε_i are i.i.d. standard Gaussian independent of the design.

Figure 3 shows that the qualities of both SAMM and MAVE deteriorate linearly in σ, when σ increases. These results also demonstrate that, thanks to an efficient bias reduction, the SAMM procedure outperforms MAVE when stochastic error is small, whereas MAVE works better than SAMM in the case of dominating stochastic error (that is when σ is large).

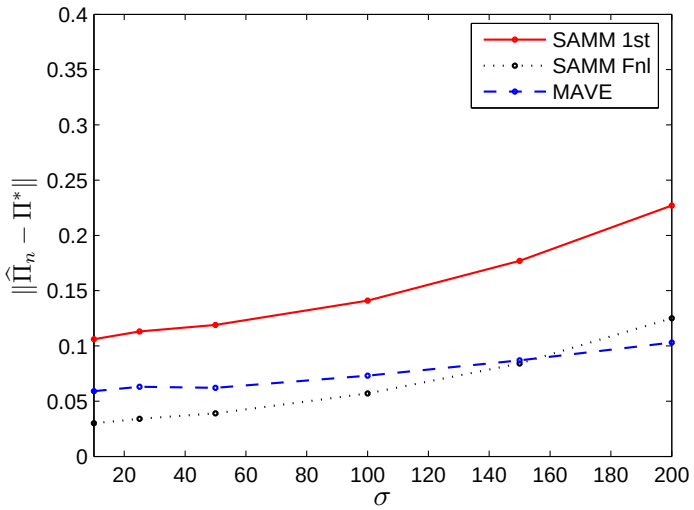

Figure 3: Average loss versus σ for the first step (solid line) and the final (dotted line) estimators provided by SAMM and for the estimator based on MAVE (dashed line) in Example 3.

4.5 Example 4 (Time Series)

Let now $T_1,\dots,T_{n+6}$ be generated by the autoregressive model

$$T_{i+6} = f(T_{i+5},T_{i+4},T_{i+3},T_{i+2},T_{i+1},T_i) + 0.2\cdot\varepsilon_i, \qquad i=1,\dots,n,$$

with initial variables $T_1,\dots,T_6$ being independent standard normal independent of the innovations ε_i, which are i.i.d. standard normal as well. Let now $f(x) = g(\vartheta^\top x)$ with

$$g(x) = -1 + 0.6x_1 - \cos(0.5\pi x_2) + e^{-x_3^2},$$

and

$$\begin{aligned}\vartheta_1 &= (1,0,0,2,0,0)/\sqrt{5},\\ \vartheta_2 &= (0,0,1,0,0,2)/\sqrt{5},\\ \vartheta_3 &= (-2,2,-2,1,-1,1)/\sqrt{15}.\end{aligned}$$

We ran SAMM and MAVE procedures on the data (X_i,Y_i), $i=1,\dots,250$, where $Y_i = T_{i+6}$ and $X_i = (T_i,\dots,T_{i+5})^\top$. The results of simulations reported in Table 4 show that the qualities of SAMM and MAVE are comparable, with SAMM being slightly better. SAVE is better than MDA, but both of them are far less accurate than SAMM and MAVE.

n	300	400	500	600
SAMM, 1st	0.391	0.351	0.334	0.293
	(.172)	(.161)	(.137)	(.132)
SAMM, Fnl	0.220	0.186	0.174	0.146
	(.119)	(.123)	(.102)	(.089)
MAVE	0.268	0.231	0.209	0.182
	(.209)	(.170)	(.159)	(.122)
MDA	0.914	0.915	0.913	0.912
	(.115)	(.107)	(.119)	(.119)
SAVE	0.617	0.515	0.428	0.369
	(.200)	(.184)	(.151)	(.138)

Table 4: Average loss $\|\widehat{\Pi} - \Pi^*\|$ of the estimators obtained by SAMM, MAVE, MDA and SAVE procedures in Example 4. The standard deviation is given in parentheses.

5. Proofs

Since the proof of the main result is carried out in several steps, we give a short road map for guiding the reader throughout the proof. The main idea is to evaluate the accuracy of the first step estimators of β_ℓ and, given the accuracy of the estimator at the step k, evaluate the accuracy of the estimators at the step $k+1$. This is done in Subsections 5.2 and 5.1. These results are based on a maximal inequality proved in Subsection 5.4 and on some properties of the solution to (6) proved in Subsection 5.5. The proof of Theorem 4 is presented in Subsection 5.3, while some technical lemmas are postponed to Subsection 5.6.

5.1 One Step Improvement

Let $\{\delta_k\}$ be a sequence of positive numbers to be chosen later and let $\mathscr{P}_k = \big\{A \in \mathcal{A}_{m^*} : \operatorname{tr}(I-A)\Pi^* \le \delta_k^2\big\}$. Recall that we use the following notation:

$$P_k^* = (I+\rho_k^{-2}\Pi^*)^{-1/2}, \quad Z_{ij}^{(k)} = (h_k P_k^*)^{-1} X_{ij}, \quad w_{ij}^{(k)}(U) = K\big((Z_{ij}^{(k)})^\top U Z_{ij}^{(k)}\big)$$
$$N_i^{(k)}(U) = \sum_j w_{ij}^{(k)}(U), \qquad V_i^{(k)}(U) = \sum_{j=1}^n \begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}\begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}^\top w_{ij}^{(k)}(U),$$

where U is a $d\times d$ symmetric positive-semidefinite matrix. Let us define $S_k = (I+\rho_k^{-2}\hat{A}_{k-1})^{1/2}$ and $U_k = P_k^* S_k^2 P_k^*$.

One easily checks that the estimator $\widehat{\nabla f}_k(X_i)$ is given by

$$\begin{pmatrix}\hat{f}_k(X_i)\\ \widehat{\nabla f}_k(X_i)\end{pmatrix} = \Big\{\sum_{j=1}^n \begin{pmatrix}1\\ X_{ij}\end{pmatrix}\begin{pmatrix}1\\ X_{ij}\end{pmatrix}^\top w_{ij}^{(k)}(U_k)\Big\}^{-1} \sum_{j=1}^n Y_j \begin{pmatrix}1\\ X_{ij}\end{pmatrix} w_{ij}^{(k)}(U_k),$$

Simple algebra yields

$$\begin{pmatrix}h_k^{-1}\hat{f}_k(X_i)\\ P_k^*\widehat{\nabla f}_k(X_i)\end{pmatrix} = h_k^{-1} V_i^{(k)}(U_k)^{-1} \sum_{j=1}^n Y_j \begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix} w_{ij}^{(k)}(U_k).$$

In order to study the behavior of $\widehat{\nabla f}_k$, we will proceed in a first step as if U_k were deterministic. For this reason, the notation

$$\begin{pmatrix}h_k^{-1}\overline{f}_k(X_i)\\ P_k^*\overline{\nabla f}_k(X_i)\end{pmatrix} = h_k^{-1} V_i^{(k)}(U_k)^{-1} \sum_{j=1}^n f(X_j) \begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix} w_{ij}^{(k)}(U_k),$$

will be useful. In fact, $\overline{\nabla f}_k(X_i)$ defined as above would be the expectation of $\widehat{\nabla f}_k(X_i)$ if U_k were deterministic.

Proposition 9 *Let assumptions (A1)-(A4) be fulfilled. If for some integer $k \in [2, k(n)]$ the real number $\alpha_k = 2\delta_{k-1}^2\rho_k^{-2} + 2\delta_{k-1}\rho_k^{-1}$ is less than the constant α appearing in assumption (A3), then there exist Gaussian vectors $\xi_{1,k}^*, \dots, \xi_{L,k}^* \in \mathbb{R}^d$ such that $\max_{1\le\ell\le L}\mathbf{E}[|\xi_{\ell,k}^*|^2] \le c_0^2\sigma^2$ and*

$$\mathbf{P}\Big(\max_{1\le\ell\le L}\Big|P_k^*(\hat{\beta}_{\ell,k}-\beta_\ell) - \frac{\xi_{\ell,k}^*}{\sqrt{n}h_k}\Big| \ge \Upsilon_k,\ \hat{A}_{k-1}\in\mathscr{P}_{k-1}\Big) \le \frac{2}{n}\ ,$$

where we used the notation $\Upsilon_k = \sqrt{C_V}C_g(\rho_k+\delta_{k-1})^2 h_k + c_1\sigma\alpha_k t_n/(\sqrt{n}h_k)$ with $t_n = 4+(3\log(Ln)+\frac{3}{2}d^2\log n)^{1/2}$, $c_0 = \bar{\psi}(dC_KC_V)^{1/2}$ and $c_1 = 15\bar{\psi}(C_w^2C_V^4C_K^2 + C_V^2C_{K'}^2)^{1/2}$.

Proof Let us start with evaluating the "bias" term $|P_k^*(\bar{\beta}_{\ell,k}-\beta_\ell)|$, where the vectors $\bar{\beta}_{\ell,k}$ are defined as $\frac{1}{n}\sum_{i=1}^n \overline{\nabla f}_k(X_i)\psi_{\ell,i}$. According to the Cauchy-Schwarz inequality, it holds

$$\begin{aligned}\big|P_k^*\big(\bar{\beta}_{\ell,k}-\beta_\ell\big)\big|^2 &= n^{-2}\Big|\sum_{i=1}^n P_k^*\big(\overline{\nabla f}_k(X_i) - \nabla f(X_i)\big)\psi_{\ell,i}\Big|^2\\ &\le n^{-2}\sum_{i=1}^n\big|P_k^*\big(\overline{\nabla f}_k(X_i)-\nabla f(X_i)\big)\big|^2\sum_{i=1}^n\psi_{l,i}^2\\ &\le \max_{i=1,\dots,n}\big|P_k^*\big(\overline{\nabla f}_k(X_i)-\nabla f(X_i)\big)\big|^2.\end{aligned}$$

Simple computations show that

$$\begin{aligned}\left|P_k^*\left(\overline{\nabla f}_k(X_i)-\nabla f(X_i)\right)\right| &\le \left|\begin{pmatrix} h_k^{-1}\overline{f}_k(X_i)\\ P_k^*\overline{\nabla f}_k(X_i)\end{pmatrix}-\begin{pmatrix} h_k^{-1}f(X_i)\\ P_k^*\nabla f(X_i)\end{pmatrix}\right| \\ &= \left|h_k^{-1}V_i^{(k)}(U_k)^{-1}\sum_{j=1}^n f(X_j)\begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}w_{ij}^{(k)}(U_k)-\begin{pmatrix} h_k^{-1}f(X_i)\\ P_k^*\nabla f(X_i)\end{pmatrix}\right| \\ &= h_k^{-1}\left|V_i^{(k)}(U_k)^{-1}\sum_{j=1}^n r_{ij}\begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}w_{ij}^{(k)}(U_k)\right| := b(X_i),\end{aligned}$$

where $r_{ij}=f(X_j)-f(X_i)-X_{ij}^\top\nabla f(X_i)$. Define $v_j=V_i^{(k)}(U_k)^{-1/2}\begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}\sqrt{w_{ij}^{(k)}(U_k)}$, $\lambda_j=h_k^{-1}r_{ij}\sqrt{w_{ij}^{(k)}(U_k)}$ and $\lambda=(\lambda_1,\dots,\lambda_n)^\top$. Then $\sum_j v_j v_j^\top=I_{d+1}$ and

$$b(X_i)=\left|V_i^{(k)}(U_k)^{-1/2}\sum_{j=1}^n\lambda_j v_j\right|\le\left\|V_i^{(k)}(U_k)^{-1/2}\right\|\cdot|\lambda|.$$

Note now that in view of Lemma 21, $\|U_k-I\|_2\le\alpha_k$ on the event $\{\hat{A}_{k-1}\in\mathscr{P}_{k-1}\}$. Therefore,

$$\begin{aligned}b(X_i)^2 &\le h_k^{-2}\left\|V_i^{(k)}(U_k)^{-1/2}\right\|^2\cdot\sum_{j=1}^n r_{ij}^2 w_{ij}^{(k)}(U_k)\\ &\le h_k^{-2}\max_{j:w_{ij}^{(k)}(U_k)\ne0} r_{ij}^2\left\|V_i^{(k)}(U_k)^{-1}\right\|\cdot\sum_{j=1}^n w_{ij}^{(k)}(U_k)\le C_V h_k^{-2}\max_{j:w_{ij}^{(k)}(U_k)\ne0} r_{ij}^2.\end{aligned}$$

Let us denote by Θ the $(d\times m^*)$ matrix having ϑ_l as lth column. Then $\Pi^*=\Theta\Theta^\top$ and therefore, in view of (A1),

$$\begin{aligned}|r_{ij}| &= |f(X_j)-f(X_i)-X_{ij}^\top\nabla f(X_i)|\\ &= |g(\Theta^\top X_j)-g(\Theta^\top X_i)-(\Theta^\top X_{ij})^\top\nabla g(\Theta^\top X_i)|\\ &\le C_g|\Theta^\top X_{ij}|^2=C_g|\Pi^*X_{ij}|^2.\end{aligned}$$

Since the weights $w_{ij}^{(k)}$ are defined via the kernel function K vanishing on the interval $[1,\infty[$, we have $\max_{j:w_{ij}^{(k)}(U_k)\ne0} r_{ij}^2=\max\{r_{ij}^2:|S_kX_{ij}|\le h_k\}$. By Corollary 19, the inequality $|S_kX_{ij}|\le h_k$ implies $|\Pi^*X_{ij}|\le(\rho_k+\delta_{k-1})h_k$. On the other hand, $|\Pi^*X_{ij}|\le|X_{ij}|\le|S_kX_{ij}|\le h_k$. These estimates yield $|b(X_i)|\le\sqrt{C_V}C_g\{(\rho_k+\delta_{k-1})\wedge1\}^2h_k$, and consequently,

$$\max_{\ell=1,\dots,L}\left|P_k^*\left(\overline{\beta}_{\ell,k}-\beta_\ell\right)\right|\le\max_i b(X_i)\le\sqrt{C_V}C_g\{(\rho_k+\delta_{k-1})\wedge1\}^2h_k. \tag{9}$$

Let us evaluate now the "stochastic" error $P_k^*\left(\hat{\beta}_{\ell,k}-\overline{\beta}_{\ell,k}\right)$. Define E_1 as the $d\times(d+1)$ matrix $(0\ I)$, where 0 stands for the vector all coordinates of which are zero and I is the $d\times d$ identity matrix. Using this notation, we have $P_k^*\left(\hat{\beta}_{\ell,k}-\overline{\beta}_{\ell,k}\right)=\sum_{j=1}^n c_{j,\ell}(U_k)\,\varepsilon_j$, where

$$c_{j,\ell}(U_k)=\frac{1}{nh_k}\sum_{i=1}^n E_1V_i^{(k)}(U_k)^{-1}\begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}w_{ij}^{(k)}(U_k)\psi_{\ell,i}.$$

Let us define $\xi^*_{\ell,k} = \sqrt{n}h_k \sum_{j=1}^n c_{j,\ell}(I)\,\varepsilon_j$. Clearly, the vectors $\xi^*_{\ell,k}$ are centered Gaussian and, in view of Lemma 22, they satisfy $\mathbf{E}[|\xi^*_{\ell,k}|^2] \le nh_k^2\sigma^2 \sum_j |c_{j,\ell}(I)|^2 \le c_0^2\sigma^2$.

By virtue of Lemma 21, on the event $\{\hat{A}_{k-1} \in \mathscr{P}_{k-1}\}$, for any $\ell = 1,\dots,L$ we have

$$\left| P_k^*(\hat{\beta}_{\ell,k} - \bar{\beta}_{\ell,k}) - \frac{\xi^*_{\ell,k}}{\sqrt{n}h_k} \right| \le \sup_{\|U-I\|_2 \le \alpha_k} \left| \sum_{j=1}^n \big(c_{j,\ell}(U) - c_{j,\ell}(I)\big)\,\varepsilon_j \right|.$$

Set $a_{j,\ell}(U) = c_{j,\ell}(U) - c_{j,\ell}(I)$. Lemma 23 and inequality (12) imply that Proposition 14 can be applied with $\kappa_0 = \frac{c_1\alpha_k}{\sqrt{n}h_k}$ and $\kappa_1 = \frac{c_1}{\sqrt{n}h_k}$. Setting $\varepsilon = 2\alpha_k/\sqrt{n}$ we get that the probability of the event

$$\left\{ \sup_{U,\ell} \left| \sum_{j=1}^n \big(c_{j,\ell}(U) - c_{j,\ell}(I)\big)\,\varepsilon_j \right| \ge \frac{c_1\sigma\alpha_k(4 + \sqrt{3\log(Ln) + 3d^2\log(\sqrt{n})})}{\sqrt{n}h_k} \right\}$$

is less than $2/n$. This completes the proof of the proposition. ■

Corollary 10 *If $nL \ge 6$ and the assumptions of Proposition 9 are fulfilled, then*

$$\mathbf{P}\left(\max_\ell \left|P_k^*(\hat{\beta}_{\ell,k} - \beta_\ell)\right| \ge \Upsilon_k + \frac{\sigma c_0 z}{\sqrt{n}h_k}, \hat{A}_{k-1} \in \mathscr{P}_{k-1} \right) \le Lze^{-\frac{z^2-1}{2}}.$$

In particular, if $nL \ge 6$, the probability of the event

$$\left\{ \max_\ell \left|P_k^*(\hat{\beta}_{\ell,k} - \beta_\ell)\right| \ge \Upsilon_k + \frac{2\sigma c_0\sqrt{\log(Ln)}}{\sqrt{n}h_k} \right\} \cap \{\hat{A}_{k-1} \in \mathscr{P}_{k-1}\}$$

does not exceed $3/n$, where Υ_k and c_0 are defined in Proposition 9.

Proof In view of Lemma 7 in Hristache et al. (2001b), we have

$$\mathbf{P}\big(\max_{\ell=1,\dots,L} \left|\xi^*_{\ell,k}\right| \ge zc_0\sigma\big) \le \sum_{\ell=1}^L \mathbf{P}\big(\left|\xi^*_{\ell,k}\right| \ge zc_0\sigma\big) \le Lze^{-(z^2-1)/2}.$$

The choice $z = \sqrt{4\log(nL)}$ leads to the desired inequality provided that $nL \ge 6$. ■

5.2 The Accuracy of the First-step Estimator

Since at the first step no information about the EDR subspace is available, we use the same bandwidth in all directions, that is the local neighborhoods are balls (and not ellipsoids) of radius h. Therefore the first step estimator $\hat{\beta}_{\ell,1}$ of the vector β_ℓ is the same as the one used in Hristache et al. (2001a).

Proposition 11 *Under assumptions (A1), (A3), (A4) and (8), for every $\ell \le L$, there exists a d-dimensional zero mean Gaussian vector $\xi^*_{\ell,1}$ so that*

$$\left| \hat{\beta}_{\ell,1} - \beta_\ell - \frac{\xi^*_{\ell,1}}{\sqrt{n}h_1} \right| \le h_1 C_g \sqrt{C_V},$$

and $\mathbf{E}|\xi^*_{\ell,1}|^2 \le d\sigma^2 C_V C_K \overline{\psi}^2$.

Proof Since P_1^* coincides by definition with the identity matrix, the arguments used in the proof of Proposition 9 apply with $S_1 = I$ and therefore $\delta_0 = \alpha_1 = 0$. More precisely, in view of (9) and $\rho_1 = 1$, we have $|\beta_{\ell,1} - \beta_\ell| \le h_1\sqrt{C_V}C_g$ for all ℓ, while in view of the relation $U_1 = I$, we have $\hat\beta_{\ell,1} - \bar\beta_{\ell,1} = \frac{1}{\sqrt{n}h_1}\xi^*_{\ell,1}$. This yields the desired result. ■

Corollary 12 *If $nL \ge 6$ and the assertions of Proposition 11 hold, then*

$$\mathbf{P}\left(\max_\ell |\hat\beta_{\ell,1} - \beta_\ell| \ge h_1 C_g\sqrt{C_V} + \frac{2\sqrt{dC_VC_K\log(nL)}\,\sigma\bar\psi}{h_1\sqrt{n}} \right) \le \frac{1}{n}.$$

Remark 13 *In order that the kernel estimator of $\nabla f(x)$ be consistent, the ball centered at x with radius h_1 should contain at least d points from $\{X_i, i = 1,\dots,n\}$. If the design is regular, this means that h_1 is at least of order $n^{-1/d}$. The optimization of the risk of $\hat\beta_{1,\ell}$ with respect to h_1 verifying $h_1 \ge n^{-1/d}$ leads to $h_1 = Const.n^{-1/(4\vee d)}$. This motivates the choice of h_1 presented in Section 3.*

5.3 Proof of Theorem 4

Recall that at the first step we use the following values of parameters: $\hat A_0 = 0$, $\rho_1 = 1$ and $h_1 = n^{-1/(d\vee 4)}$. Let us denote

$$\gamma_1 = h_1C_g\sqrt{C_V} + \frac{2\sigma\bar\psi\sqrt{2dC_VC_K\log(nL)}}{h_1\sqrt{n}}, \quad \delta_1 = 2\gamma_1\sqrt{\mu^*},$$

and introduce the event $\Omega_1 = \{\max_\ell |\hat\beta_{1,\ell} - \beta_\ell| \le \gamma_1\}$. According to Corollary 12 the probability of the event Ω_1 is at least $1 - n^{-1}$. In conjunction with Proposition 17, this implies that $\mathbf{P}(\mathrm{tr}(I - \hat A_1)\Pi^* \le \delta_1^2) \ge 1 - n^{-1}$.

Recall that for any integer $k \in [2, k(n)]$—where $k(n)$ is the total number of iterations—we use the notation $\rho_k = a_\rho\rho_{k-1}$, $h_k = a_h h_{k-1}$ and $\alpha_k = 2\delta_{k-1}^2\rho_k^{-2} + 2\delta_{k-1}\rho_k^{-1}$. Let us introduce the additional notation

$$\begin{aligned}
\gamma_k &= \frac{1}{\sqrt{n}h_k}\begin{cases}\sqrt{n}h_k\Upsilon_k + 2\sigma c_0\sqrt{\log(nL)}, & k < k(n),\\ \sqrt{n}h_k\Upsilon_k + \sigma c_0 z, & k = k(n),\end{cases}\\
\zeta_k &= 2\mu^*(\gamma_k^2\rho_k^{-2} + \sqrt{2}\,\gamma_k\rho_k^{-1}C_g),\\
\delta_k &= 2\gamma_k\sqrt{\mu^*}/\sqrt{1-\zeta_k},\\
\Omega_k &= \{\max_\ell |P_k^*(\hat\beta_{\ell,k} - \beta_\ell)| \le \gamma_k\}.
\end{aligned}$$

Combining Lemmas 24 and 25, we obtain $\mathbf{P}(\mathrm{tr}(I - \hat A_{k-1})\Pi^* > \delta_{k-1}^2) \le \mathbf{P}(\Omega_{k-1}^c)$ and therefore, using Corollary 10, we get

$$\begin{aligned}
\mathbf{P}\big(\Omega_k^c\big) &\le \mathbf{P}\Big(\max_\ell |P_k^*(\hat\beta_{\ell,k} - \beta_\ell)| > \gamma_k,\, \hat A_{k-1} \in \mathscr{P}_{k-1}\Big) + \mathbf{P}\big(\Omega_{k-1}^c\big)\\
&\le \frac{3}{n} + \mathbf{P}\big(\Omega_{k-1}^c\big), \qquad \forall\, k \le k(n) - 1.
\end{aligned}$$

Since $\mathbf{P}(\Omega_1^c) \le 1/n$, it holds $\mathbf{P}(\Omega_{k(n)-1}^c) \le (3k(n)-5)/n$ and, by virtue of Corollary 10, $\mathbf{P}(\Omega_{k(n)}^c) \le Lze^{-(z^2-1)/2} + \frac{3k(n)-5}{n}$. In conjunction with Lemma 25, this yields

$$\mathbf{P}\big(\operatorname{tr}(I-\hat{A}_{k(n)})\Pi^* > \delta_{k(n)}^2\big) \le Lze^{-(z^2-1)/2} + \frac{3k(n)-5}{n}\,. \tag{10}$$

According to Lemma 24, we have $\delta_{k(n)-2} \le \rho_{k(n)-1}$, $\alpha_{k(n)-1} \le 4$ and $\zeta_{k(n)-1} \le 1/2$. Consequently, for n sufficiently large, we have

$$\delta_{k(n)-1} = \frac{2\sqrt{\mu^*}\gamma_{k(n)-1}}{\sqrt{1-\zeta_{k(n)-1}}} \le C\left(\frac{\log(Ln)}{n}\right)^{1/2} \vee n^{-2/3\vee m^*}$$

and $\alpha_{k(n)} \le 4\delta_{k(n)-1}\rho_{k(n)}^{-1} \le C[(\sqrt{\log(Ln)}(\rho_{k(n)}\sqrt{n})^{-1}) \vee n^{-1/3\vee m^*}]$. Since $h_{k(n)} = 1$ and $(n\rho_{k(n)})^{-1} \le \rho_{k(n)}^2 = n^{-2/(3\vee m^*)}$, we infer that

$$\begin{aligned}\gamma_{k(n)} &= C_g\sqrt{C_V}\,(\rho_{k(n)} + \delta_{k(n)-1})^2 + \frac{\sigma(zc_0 + c_1\alpha_{k(n)}t_n)}{\sqrt{n}} \\ &\le Ct_n^2 n^{-2/(3\vee m^*)} + \frac{c_0\sigma z}{\sqrt{n}}.\end{aligned}$$

Therefore $\zeta_n := \zeta_{k(n)} = O(\gamma_{k(n)}\rho_{k(n)}^{-1})$ tends to zero as n tends to infinity not slower than $\sqrt{\log(nL)}\,n^{-1/(6\vee m^*)}$ and the assertion of the theorem follows from (10), the definition of $\delta_{k(n)}$ and Lemma 20.

5.4 Maximal Inequality

The following result contains a well known maximal inequality for the maximum of a Gaussian process. We include its proof for the completeness of exposition. Let $\mathbb{S}_{d-1}$ denote the unit ball of $\mathbb{R}^d$.

Proposition 14 *Let r be a positive number and let Γ be a finite set. Let functions $a_{j,\gamma} : \mathbb{R}^p \to \mathbb{R}^d$ obey the conditions*

$$\begin{aligned}&\sup_{\gamma\in\Gamma}\sup_{|u-u^*|\le r}\sum_{j=1}^n |a_{j,\gamma}(u)|^2 \le \kappa_0^2, \\ &\sup_{\gamma\in\Gamma}\sup_{|u-u^*|\le r}\sup_{e\in\mathbb{S}_{d-1}}\sum_{j=1}^n \left|\frac{d}{du}\,(e^\top a_{j,\gamma}(u))\right|^2 \le \kappa_1^2\end{aligned}$$

for some $u^* \in \mathbb{R}^p$. If the ε_j's are independent $\mathcal{N}(0,\sigma^2)$-distributed random variables, then

$$\mathbf{P}\left(\sup_{\gamma\in\Gamma}\sup_{|u-u^*|\le r}\left|\sum_{j=1}^n a_{j,\gamma}(u)\,\varepsilon_j\right| > t\sigma\kappa_0 + 2\sqrt{n}\sigma\kappa_1\varepsilon\right) \le \frac{2}{n},$$

where $t = \sqrt{3\log(|\Gamma|(2r/\varepsilon)^p n)}$ and $|\Gamma|$ is the cardinality of Γ.

Proof Let B_r be the ball $\{u : |u-u^*| \le r\} \subset \mathbb{R}^p$ and $\Sigma_{r,\varepsilon}$ be an ε-net on B_r such that for any $u \in B_r$ there is an element $u_l \in \Sigma_{r,\varepsilon}$ such that $|u-u_l| \le \varepsilon$. It is easy to see that such a net with cardinality $N_{r,\varepsilon} < (2r/\varepsilon)^p$ can be constructed. For every $u \in B_r$ we denote$\eta_\gamma(u) = \sum_{j=1}^n a_{j,\gamma}(u)\varepsilon_j$. Since $\mathbf{E}(|\eta_\gamma(u)|^2) \le \sigma^2\kappa_0^2$ for any γ and for any u, we have

$$\mathbf{P}\big(|\eta_\gamma(u_l)| > t\sigma\kappa_0\big) \le \mathbf{P}\Big(|\eta_\gamma(u_l)| > t\sqrt{\mathbf{E}(|\eta_\gamma(u_l)|^2)}\Big) \le te^{-(t^2-1)/2}.$$

Thus we get

$$\mathbf{P}\Big(\sup_{\gamma\in\Gamma}\sup_{u_l\in\Sigma_{r,\varepsilon}}\big|\eta_\gamma(u_l)\big| > t\sigma\kappa_0\Big) \le \sum_{\gamma\in\Gamma}\sum_{l=1}^{N_{r,\varepsilon}}\mathbf{P}\Big(\big|\eta_\gamma(u_l)\big| > t\sigma\kappa_0\Big) \le |\Gamma|N_{r,\varepsilon}te^{-(t^2-1)/2}.$$

Hence, if $t = \sqrt{3\log(|\Gamma|N_{r,\varepsilon}n)}$, then $\mathbf{P}\Big(\sup_{\gamma\in\Gamma}\sup_{u_l\in\Sigma_{r,\varepsilon}}\big|\eta_\gamma(u_l)\big| > t\sigma\kappa_0\Big) \le 1/n$. On the other hand, for any $u, u' \in B_r$,

$$\begin{aligned}
\big|\eta_\gamma(u)-\eta_\gamma(u')\big|^2 &= \sup_{e\in\mathbb{S}_{d-1}}\big|e^\top\big(\eta_\gamma(u)-\eta_\gamma(u')\big)\big|^2 \\
&\le |u-u'|^2\cdot\sup_{u\in B_r}\sup_{e\in\mathbb{S}_{d-1}}\left|\frac{d(e^\top\eta_\gamma)}{du}(u)\right|^2 \\
&= |u-u'|^2\cdot\sup_{u\in B_r}\sup_{e\in\mathbb{S}_{d-1}}\left|\sum_{j=1}^n\frac{d(e^\top a_{j,\gamma})}{du}(u)\varepsilon_j\right|^2.
\end{aligned}$$

The Cauchy-Schwarz inequality yields

$$\frac{\big|\eta_\gamma(u)-\eta_\gamma(u')\big|^2}{|u-u'|^2} \le \sup_{u\in B_r}\sup_{e\in\mathbb{S}_{d-1}}\sum_{j=1}^n\left|\frac{d(e^\top a_{j,\gamma})}{du}(u)\right|^2\sum_{j=1}^n\varepsilon_j^2 \le \kappa_1^2\sum_{j=1}^n\varepsilon_j^2.$$

Since $\mathbf{P}\big(\sum_{j=1}^n\varepsilon_j^2 > 4n\sigma^2\big)$ is certainly less than n^{-1}, we have

$$\begin{aligned}
&\mathbf{P}\Big(\sup_{\gamma\in\Gamma}\sup_{u\in B_r}\big|\eta_\gamma(u)\big| > t\sigma\kappa_0 + 2\sqrt{n}\sigma\kappa_1\varepsilon\Big) \\
&\quad\le \mathbf{P}\Big(\sup_{\gamma\in\Gamma}\sup_{u_l\in\Sigma_{r,\varepsilon}}\frac{|\eta_\gamma(u_l)|}{t\sigma\kappa_0} > 1\Big) + \mathbf{P}\Big(\sup_{\gamma\in\Gamma}\sup_{u\in B_r}\frac{|\eta_\gamma(u)-\eta_\gamma(u_l(u))|}{2\sqrt{n}\sigma\kappa_1\varepsilon} > 1\Big) \\
&\quad\le \frac{1}{n} + \mathbf{P}\Big(\sup_{u\in B_r}\kappa_1^2|u-u_l(u)|^2\sum_{j=1}^n\varepsilon_j^2 > 4n\sigma^2\kappa_1^2\varepsilon^2\Big) \le \frac{2}{n},
\end{aligned}$$

and the assertion of proposition follows. ■

5.5 Properties of the Solution to (6)

We collect below some simple facts concerning the solution to the optimization problem (6). By classical arguments, it is always possible to choose a measurable solution $\hat{A}$ to (6). This measurability will be assumed in the sequel.

In Proposition 15, the case of general m (not necessarily equal to m^*) is considered. As we explain below, this generality is useful for further developments of the method extending it to the case of unknown structural dimension m^*.

The vectors β_ℓ are assumed to belong to a m^*-dimensional subspace $\mathcal{S}$ of $\mathbb{R}^d$, but in this subsection we do not necessarily assume that β_ℓs are defined by (4). In fact, we will apply the results of this subsection to the vectors $\Pi^*\hat\beta_\ell$.

For every $A \in \mathcal{A}_{m^*}$, let us define

$$R(A) = \max_{1\le\ell\le L} \hat\beta_\ell^\top (I-A)\hat\beta_\ell, \quad \hat A_m = \operatorname*{arg\,min}_{A\in\mathcal{A}_m} R(A),$$
$$\hat{\mathcal{R}}(m) = \min_{A\in\mathcal{A}_m} \sqrt{R(A)} = \sqrt{R(\hat A_m)} = \min_{A\in\mathcal{A}_m} \max_{1\le\ell\le L} |(I-A)^{1/2}\hat\beta_\ell|.$$

We also define

$$\mathcal{R}^*(m) = \min_{A\in\mathcal{A}_m} \max_{1\le\ell\le L} |(I-A)^{1/2}\beta_\ell|$$

and denote by A_m^* a minimizer of $\max_\ell \beta_\ell^\top (I-A)\beta_\ell$ over $A \in \mathcal{A}_m$. Note also that for every $m \ge m^*$ the projector Π^* belongs to $\mathcal{A}_m$. Therefore, we have $A_m^* = \Pi^*$ and $\mathcal{R}^*(m) = 0$ for every $m \ge m^*$.

Proposition 15 *Let $\mathcal{B}^* = \{\bar\beta = \sum_\ell c_\ell\beta_\ell : \sum_\ell |c_\ell| \le 1\}$ be the convex hull of vectors $\pm\beta_\ell$. If $\max_\ell |\hat\beta_\ell - \beta_\ell| \le \varepsilon$, then*

$$\hat{\mathcal{R}}(m) \le \mathcal{R}^*(m) + \varepsilon,$$
$$\max_{\bar\beta\in\mathcal{B}^*} |(I-\hat A_m)^{1/2}\bar\beta| \le \mathcal{R}^*(m) + 2\varepsilon.$$

When $m < m^*$, we have also the lower bound $\hat{\mathcal{R}}(m) \ge (\mathcal{R}^*(m) - \varepsilon)_+$.

Proof For every $\ell \in 1,\dots,L$, we have

$$\begin{aligned}|(I-A_m^*)^{1/2}\hat\beta_\ell| &\le |(I-A_m^*)^{1/2}\beta_\ell| + |(I-A_m^*)^{1/2}(\hat\beta_\ell - \beta_\ell)| \\ &\le \mathcal{R}^*(m) + |\hat\beta_\ell - \beta_\ell| \le \mathcal{R}^*(m) + \varepsilon.\end{aligned}$$

Since $\hat A_m$ minimizes $\max_\ell |(I-A)^{1/2}\hat\beta_\ell|$ over $A \in \mathcal{A}_m$, we have

$$\max_\ell |(I-\hat A_m)^{1/2}\hat\beta_\ell| \le \max_\ell |(I-A_m^*)^{1/2}\hat\beta_\ell| \le \mathcal{R}^*(m) + \varepsilon.$$

Since $\hat A_m \in \mathcal{A}_m$, we have $0 \preceq (I-\hat A_m)^{1/2} \preceq I$ and consequently, for every ℓ,

$$\begin{aligned}|(I-\hat A_m)^{1/2}\beta_\ell| &\le |(I-\hat A_m)^{1/2}\hat\beta_\ell| + |(I-\hat A_m)^{1/2}(\beta_\ell - \hat\beta_\ell)| \\ &\le |(I-\hat A_m)^{1/2}\hat\beta_\ell| + |\beta_\ell - \hat\beta_\ell| \le \mathcal{R}^*(m) + 2\varepsilon.\end{aligned}$$

The second inequality of the proposition follows now from $|(I-\hat A_m)^{1/2}\bar\beta| \le \max_\ell |(I-\hat A_m)^{1/2}\beta_\ell|$ for every $\bar\beta \in \mathcal{B}^*$.

Let us prove the last assertion of the proposition. According to the definition of $\mathcal{R}^*(m)$, for every matrix $A \in \mathcal{A}_m$ there exists an index $\ell(A)$ such that $|(I-A)^{1/2}\beta_{\ell(A)}| \ge \mathcal{R}^*(m)$. In particular, $|(I-\hat A_m)^{1/2}\beta_{\ell(\hat A_m)}| \ge \mathcal{R}^*(m)$ and hence $|(I-\hat A_m)^{1/2}\hat\beta_{\ell(\hat A_m)}| \ge |(I-\hat A_m)^{1/2}\beta_{\ell(\hat A_m)}| - |\hat\beta_{\ell(\hat A_m)} - \beta_{\ell(\hat A_m)}| \ge \mathcal{R}^*(m) - \varepsilon$. ■

Remark 16 *Proposition 15 can be used for estimating the structural dimension m. Indeed,* $\hat{\mathcal{R}}(m) \le \varepsilon$ *for* $m \ge m^*$ *and* $\hat{\mathcal{R}}(m) \ge (\mathcal{R}^*(m) - \varepsilon)_+$ *for* $m < m^*$. *Therefore, it is natural to search for the smallest value* $\widehat{m}$ *of m such that the function* $\hat{\mathcal{R}}(m)$ *does not significantly decrease for* $m \ge \widehat{m}$. *The rigorous application of this heuristic argument is currently under investigation.*

From now on, we assume that the structural dimension m^* is known and we use the shortened notation $\hat{A}$ instead of $\hat{A}_{m^*}$.

Proposition 17 *If the vectors* β_ℓ *satisfy* ***(A2)*** *and* $\max_\ell |\hat{\beta}_\ell - \beta_\ell| \le \varepsilon$, *then* $\operatorname{tr}(I - \hat{A})\Pi^* \le 4\varepsilon^2\mu^*$ *and* $\operatorname{tr}[(\hat{A} - \Pi^*)^2] \le 8\varepsilon^2\mu^*$.

Proof In view of the relations $\operatorname{tr}\hat{A}^2 \le \operatorname{tr}\hat{A} \le m^*$ and $\operatorname{tr}(\Pi^*)^2 = \operatorname{tr}\Pi^* = m^*$, we have

$$\operatorname{tr}(\hat{A} - \Pi^*)^2 = \operatorname{tr}(\hat{A}^2 - \Pi^*) + 2\operatorname{tr}(I - \hat{A})\Pi^* \le 2|\operatorname{tr}(I - \hat{A})\Pi^*|.$$

Note also that the equality $\operatorname{tr}(I - \hat{A})\Pi^* = \operatorname{tr}(I - \hat{A})^{1/2}\Pi^*(I - \hat{A})^{1/2}$ implies that $\operatorname{tr}(I - \hat{A})\Pi^* \ge 0$. Now condition (7) and Proposition 15 imply

$$\begin{aligned}\operatorname{tr}(I - \hat{A})\Pi^* &= \operatorname{tr}(I - \hat{A})^{1/2}\Pi^*(I - \hat{A})^{1/2} \\ &\le \sum_{k=1}^{m^*} \mu_k \operatorname{tr}(I - \hat{A})^{1/2}\bar{\beta}_k\bar{\beta}_k^\top(I - \hat{A})^{1/2} \\ &\le \sum_{k=1}^{m^*} \mu_k \bar{\beta}_k^\top(I - \hat{A})\bar{\beta}_k \le (2\varepsilon)^2 \sum_{k=1}^{m^*} \mu_k\end{aligned}$$

and the assertion follows. ■

Lemma 18 *Let* $\operatorname{tr}(I - \hat{A})\Pi^* \le \delta^2$ *for some* $\delta > 0$. *Then for any* $x \in \mathbb{R}^d$

$$|\Pi^* x| \le |\hat{A}^{1/2}x| + \delta|x|.$$

Proof In view of the triangle inequality, $|\Pi^* x| \le |\Pi^*\hat{A}^{1/2}x| + |\Pi^*(I - \hat{A}^{1/2})x|$. On the other hand,

$$|\Pi^*(I - \hat{A}^{1/2})x|^2 \le \|\Pi^*(I - \hat{A}^{1/2})\|_2^2 \cdot |x|^2 \le \operatorname{tr}[\Pi^*(I - \hat{A}^{1/2})^2\Pi^*] \cdot |x|^2.$$

For every $A \in \mathcal{A}_m$, it obviously holds $(I - A^{1/2})^2 = I - 2A^{1/2} + A \preceq I - A$, and hence, $\operatorname{tr}\Pi^*(I - A^{1/2})^2\Pi^* \le \operatorname{tr}\Pi^*(I - A)\Pi^*$. Therefore,

$$\operatorname{tr}\Pi^*(I - \hat{A}^{1/2})^2\Pi^* \le \operatorname{tr}\Pi^*(I - \hat{A})\Pi^* = \operatorname{tr}(I - \hat{A})\Pi^* \le \delta^2$$

yielding $|\Pi^* x| \le |\Pi^*\hat{A}^{1/2}x| + \delta|x| \le |\hat{A}^{1/2}x| + \delta|x|$ as required. ■

Corollary 19 *If for some* $\rho \in (0,1)$ *and for some* $x \in \mathbb{R}^d$, *we have* $|(I + \rho^{-2}\hat{A})^{1/2}x| \le h$, *then* $|\Pi^* x| \le (\rho + \sqrt{\operatorname{tr}(I - \hat{A})\Pi^*})h$.

Proof The result follows from Lemma 18 and the inequalities $|x| \le |(I+\rho^{-2}\hat{A})^{1/2}x| \le h$ and $|\hat{A}^{1/2}x| \le \rho|(I+\rho^{-2}\hat{A})^{1/2}x| \le \rho h$. ■

Lemma 20 *Let* $\mathrm{tr}(I-\hat{A})\Pi^* \le \delta^2$ *for some* $\delta \in [0,1)$ *and let* $\widehat{\Pi}_{m^*}$ *be the orthogonal projection matrix in* $\mathbb{R}^d$ *onto the subspace spanned by the eigenvectors of* $\hat{A}$ *corresponding to its largest* m^* *eigenvalues. Then* $\mathrm{tr}(I-\widehat{\Pi}_{m^*})\Pi^* \le \delta^2/(1-\delta^2)$.

Proof Let $\hat{\lambda}_j$ and $\hat{\vartheta}_j$, $j=1,\dots,d$ be respectively the eigenvalues and the eigenvectors of $\hat{A}$. Assume that $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \ldots \ge \hat{\lambda}_d$. Then $\hat{A} = \sum_{j=1}^d \hat{\lambda}_j \hat{\vartheta}_j \hat{\vartheta}_j^\top$ and $\widehat{\Pi}_{m^*} = \sum_{j=1}^{m^*} \hat{\vartheta}_j \hat{\vartheta}_j^\top$. Moreover, $\sum_{j=1}^d \hat{\vartheta}_j \hat{\vartheta}_j^\top = I$ since $\{\hat{\vartheta}_1,\dots,\hat{\vartheta}_d\}$ is an orthonormal basis of $\mathbb{R}^d$. This implies that

$$\begin{aligned}
\mathrm{tr}[\hat{A}\Pi^*] &\le \sum_{j\le m^*} \hat{\lambda}_j \,\mathrm{tr}[\hat{\vartheta}_j\hat{\vartheta}_j^\top\Pi^*] + \hat{\lambda}_{m^*} \sum_{j>m^*} \mathrm{tr}[\hat{\vartheta}_j\hat{\vartheta}_j^\top\Pi^*] \\
&= \sum_{j\le m^*} (\hat{\lambda}_j - \hat{\lambda}_{m^*})\,\mathrm{tr}[\hat{\vartheta}_j\hat{\vartheta}_j^\top\Pi^*] + \hat{\lambda}_{m^*}\,\mathrm{tr}\Big[\sum_{j=1}^d \hat{\vartheta}_j\hat{\vartheta}_j^\top\Pi^*\Big] \\
&= \sum_{j\le m^*} (\hat{\lambda}_j - \hat{\lambda}_{m^*})\,\mathrm{tr}[\hat{\vartheta}_j\hat{\vartheta}_j^\top\Pi^*] + m^*\hat{\lambda}_{m^*}.
\end{aligned}$$

Since $\mathrm{tr}[\hat{\vartheta}_j\hat{\vartheta}_j^\top\Pi^*] = |\Pi^*\hat{\vartheta}_j|^2 \le 1$, we get $\mathrm{tr}[\hat{A}\Pi^*] \le \sum_{j\le m^*}\hat{\lambda}_j$. Taking into account the relations $\sum_{j\le d}\hat{\lambda}_j \le m^*$, $\mathrm{tr}\,\Pi^* = m^*$ and $(1-\hat{\lambda}_{m^*+1})(I-\widehat{\Pi}_{m^*}) \preceq I-\hat{A}$, we get $\hat{\lambda}_{m^*+1} \le m^* - \sum_{j\le m^*}\hat{\lambda}_j \le \mathrm{tr}[(I-\hat{A})\Pi^*] \le \delta^2$ and therefore $I-\widehat{\Pi}_{m^*} \preceq (1-\delta^2)^{-1}(I-\hat{A})$. Consequently, $\mathrm{tr}[(I-\widehat{\Pi}_{m^*})\Pi^*] \le (1-\delta^2)^{-1}\mathrm{tr}[(I-\hat{A})\Pi^*] \le \delta^2/(1-\delta^2)$. ■

5.6 Technical Lemmas

This subsection contains five technical results. The first three lemmas have been used in the proof of Proposition 9, whereas the two last lemmas have been used in the proof of Theorem 4.

Lemma 21 *For every* $\rho \in (0,1]$ *and for every* $A \in \mathcal{A}_{m^*}$ *we have*

$$\|P_\rho^*(I+\rho^{-2}A)P_\rho^* - I\|_2 \le 2\delta_A^2\rho^{-2} + 2\delta_A\rho^{-1},$$

where $P_\rho^* = (I+\rho^{-2}\Pi^*)^{-1/2}$ *and* $\delta_A^2 = \mathrm{tr}[(I-A)\Pi^*]$.

Proof The inequality $P_\rho^* \preceq (I-\Pi^*) + \rho\Pi^*$ implies that

$$\begin{aligned}
\rho^2\big\|P_\rho^*(I+\rho^{-2}A)P_\rho^* - I\big\|_2 &= \big\|P_\rho^*(A-\Pi^*)P_\rho^*\big\|_2 \\
&\le \rho^2\big\|\Pi^*(A-\Pi^*)\Pi^*\big\|_2 + \big\|(I-\Pi^*)(A-\Pi^*)(I-\Pi^*)\big\|_2 \\
&\quad + 2\rho\big\|\Pi^*(A-\Pi^*)(I-\Pi^*)\big\|_2.
\end{aligned}$$

Since $\|B\|_2^2 = \mathrm{tr}\,BB^\top \le (\mathrm{tr}(BB^\top)^{1/2})^2$ for any matrix B, it holds

$$\begin{aligned}
\big\|\Pi^*(A-\Pi^*)\Pi^*\big\|_2 &= \big\|\Pi^*(I-A)\Pi^*\big\|_2 \\
&\le \mathrm{tr}\,\Pi^*(I-A)\Pi^* = \mathrm{tr}(I-A)\Pi^* = \delta_A^2.
\end{aligned}$$

By similar arguments one checks that

$$\begin{aligned}\left\|(I-\Pi^*)(A-\Pi^*)(I-\Pi^*)\right\|_2 &= \left\|(I-\Pi^*)A(I-\Pi^*)\right\|_2 \le \mathrm{tr}(I-\Pi^*)A \\ &= \mathrm{tr}A - \mathrm{tr}\Pi^* + \mathrm{tr}\Pi^*(I-A) \le \delta_A^2,\end{aligned}$$

and

$$\begin{aligned}\left\|\Pi^*(A-\Pi^*)(I-\Pi^*)\right\|_2 &\le \left\|\Pi^*(A-\Pi^*)\right\|_2 = \left\|\Pi^*(I-A)\right\|_2 \\ &\le \left\|\Pi^*(I-A)^{1/2}\right\|_2 = (\mathrm{tr}\Pi^*(I-A)\Pi^*)^{1/2} \\ &= (\mathrm{tr}(I-A)\Pi^*)^{1/2} = \delta_A.\end{aligned}$$

This leads to the inequality $\|P_\rho^*(I+\rho^{-2}A)P_\rho^* - I\|_2 \le \delta_A^2(1+\rho^{-2}) + 2\delta_A\rho^{-1}$, which, in view of the condition $\rho \le 1$, yields the assertion of the lemma. ■

Lemma 22 *If ψ_ℓs and U satisfy (8) and **(A3)**, then $\sum_{j=1}^n |c_{j,\ell}(U)|^2 \le dC_KC_V\overline{\psi}^2/(nh_k^2)$.*

Proof Simple computations yield

$$\sum_{j=1}^n \left| E_1V_i^{(k)}(U)^{-1}\begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}\right|^2 w_{ij}^{(k)}(U) = \mathrm{tr}(E_1V_i^{(k)}(U)^{-1}E_1) \le \frac{dC_V}{N_i^{(k)}(U)}. \tag{11}$$

Hence, we have

$$\begin{aligned}\sum_{j=1}^n |c_{j,\ell}|^2 &= \frac{1}{n^2h_k^2}\sum_{j=1}^n \left|\sum_{i=1}^n E_1V_i^{(k)}(U_k)^{-1}\begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix} w_{ij}^{(k)}(U)\,\psi_{\ell,i}\right|^2 \\ &\le \frac{\overline{\psi}^2}{n^2h_k^2}\sum_{j=1}^n \left(\sum_{i=1}^n \frac{w_{ij}^{(k)}(U)}{N_i^{(k)}(U)}\right)\left(\sum_{i=1}^n \left|E_1V_i^{(k)}(U)^{-1}\begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}\right|^2 N_i^{(k)}(U)w_{ij}^{(k)}(U)\right) \\ &\le \frac{C_K\overline{\psi}^2}{n^2h_k^2}\sum_{j=1}^n\sum_{i=1}^n \left|E_1V_i^{(k)}(U)^{-1}\begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}\right|^2 N_i^{(k)}(U)w_{ij}^{(k)}(U).\end{aligned}$$

Interchanging the order of summation and using inequality (11) we get the desired result. ■

Lemma 23 *If **(A3)** and (8) are fulfilled, then, for any $e \in \mathbb{S}_{d-1}$, we have*

$$\sup_{U:\|U-I\|_2\le 1/2}\max_{j=1,\dots,n}\left\|\frac{d}{dU}(e^\top c_{j,\ell})(U)\right\|_2^2 \le \frac{24C_w^2C_V^4C_K^2\overline{\psi}^2}{n^2h_k^2} + \frac{216C_V^2C_{K'}^2\overline{\psi}^2}{n^2h_k^2},$$

where $\frac{d}{dU}(e^\top c_{j,\ell})(U)$ is the $d\times d$ matrix with entries $\frac{\partial e^\top c_{j,\ell}(U)}{\partial U_{pq}}$.

Proof In order to ease the notation, we will remove the superscripts (k) in this proof. Thus, we will write V_i, w_{ij} and Z_{ij} instead of $V_i^{(k)}$, $w_{ij}^{(k)}$ and $Z_{ij}^{(k)}$. By definition of $c_{j,\ell}$ we have

$$\begin{aligned}\left\|\frac{d}{dU}\,(e^\top c_{j,\ell})(U)\right\|_2^2 &\le 2\left\|\frac{1}{nh_k}\sum_{i=1}^n\left[\frac{d}{dU}\,\tilde e^\top V_i^{-1}(U)\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\right]w_{ij}(U)\psi_{\ell,i}\right\|_2^2\\ &\quad+2\left\|\frac{1}{nh_k}\sum_{i=1}^n\tilde e^\top V_i^{-1}(U)\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\frac{dw_{ij}(U)}{dU}\,\psi_{\ell,i}\right\|_2^2\\ &=\Delta_1+\Delta_2,\end{aligned}$$

where $\tilde e = E_1^\top e$ satisfies $|\tilde e|\le|e|=1$. One checks that $d\,w_{ij}(U)/dU = \bar w_{ij}(U)Z_{ij}Z_{ij}^\top$, where we used the notation $\bar w_{ij}(U)=K'(Z_{ij}^\top UZ_{ij})$. On the one hand, $|\bar w_{ij}(U)|\cdot|Z_{ij}|^2=0$ if $Z_{ij}^\top UZ_{ij}>1$. On the other hand, the inequality $\|I-U\|_2\le 1/2$ implies that

$$|Z_{ij}|^2 \le Z_{ij}^\top UZ_{ij}+|Z_{ij}^\top(I-U)Z_{ij}| \le Z_{ij}^\top UZ_{ij}+|Z_{ij}|^2\|I-U\|_2 \le Z_{ij}^\top UZ_{ij}+|Z_{ij}|^2/2.$$

Therefore $|Z_{ij}|^2\le 2$ for all Z_{ij} verifying $Z_{ij}^\top UZ_{ij}\le 1$. Hence, $\|d\,w_{ij}(U)/dU\|_2=|\bar w_{ij}(U)|\cdot|Z_{ij}|^2\le 2|\bar w_{ij}(U)|$ and we get

$$\Delta_2 \le \frac{8\overline{\psi}^2}{n^2h_k^2}\left(\sum_{i=1}^n\left|V_i^{-1}(U)\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\bar w_{ij}(U)\right|\right)^2 \le \frac{24\overline{\psi}^2C_V^2C_{K'}^2}{n^2h_k^2}\,.$$

In order to estimate the term Δ_1, remark that the differentiation (with respect to U_{pq}) of the identity $V_i^{-1}(U)V_i(U)=I_{d+1}$ yields

$$\frac{\partial V_i^{-1}}{\partial U_{pq}}\,(U) = -V_i^{-1}(U)\,\frac{\partial V_i}{\partial U_{pq}}\,(U)V_i^{-1}(U).$$

Simple computations show that

$$\begin{aligned}\frac{\partial V_i}{\partial U_{pq}}\,(U) &= \sum_{j=1}^n\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}^\top\frac{\partial}{\partial U_{pq}}w_{ij}(U)\\ &= \sum_{j=1}^n\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}^\top\bar w_{ij}(U)(Z_{ij})_p(Z_{ij})_q.\end{aligned}$$

Hence, for any $a_1,a_2\in\mathbb{R}^{d+1}$,

$$\frac{da_1^\top V_i^{-1}(U)a_2}{dU} = \sum_{j=1}^n a_1^\top V_i^{-1}(U)\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}^\top V_i^{-1}(U)a_2\;\bar w_{ij}(U)Z_{ij}Z_{ij}^\top.$$

This relation, combined with the estimate $|Z_{ij}|^2\le 2$ for all i,j such that $\bar w_{ij}\ne 0$, implies the norm estimate

$$\begin{aligned}\left\|\frac{da_1^\top V_i^{-1}(U)a_2}{dU}\right\|_2 &\le 2\sum_{j=1}^n\left|a_1^\top V_i^{-1}(U)\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}^\top V_i^{-1}(U)a_2\;\bar w_{ij}(U)\right|\\ &\le 6|a_1|\,|a_2|\sum_{j=1}^n\left\|V_i^{-1}(U)\right\|^2|\bar w_{ij}(U)|\\ &\le 6C_wC_V^2|a_1|\,|a_2|N_i(U)^{-1}.\end{aligned}$$

This yields $\Delta_1 \le 216 C_w^2 C_V^4 C_K^2 \overline{\psi}^2/(nh_k)^2$ and the assertion of the lemma follows. ■

Note that under the assumptions of Lemma 23, for some $\tilde{U}$ satisfying $\|\tilde{U}-I\|_2 \le \|U-I\|_2$, it holds

$$\begin{aligned}
|c_{j,\ell}(U)-c_{j,\ell}(I)| &= \sup_{e\in\mathbb{S}_{d-1}} |e^\top (c_{j,\ell}(U)-c_{j,\ell}(I))| \\
&= \sup_{e\in\mathbb{S}_{d-1}} |\mathbf{vec}\Big[\frac{d\, e^\top c_{j,\ell}}{dU}(\tilde{U})\Big]^\top \mathbf{vec}(U-I)| \\
&\le \sup_{e\in\mathbb{S}_{d-1}} \Big\|\frac{d\, e^\top c_{j,\ell}}{dU}(\tilde{U})\Big\|_2 \|U-I\|_2 \\
&\le \frac{\sqrt{216}\,\overline{\psi}}{nh_k}\,(C_w^2 C_V^4 C_K^2 + C_V^2 C_{K'}^2)^{1/2}\|U-I\|_2, \qquad (12)
\end{aligned}$$

where $\mathbf{vec}(\cdot)$ is a matrix operator that stacks the matrix's columns one by one. In other terms, for every $d\times d$ matrix M, $\mathbf{vec}(M) = (m_{\bullet,1}^\top,\dots,m_{\bullet,d}^\top)^\top$ where $m_{\bullet,j}$ stands for the j^{th} column of M.

Lemma 24 *There exists an integer $n_0 \ge 0$ such that, for every $n \ge n_0$ and for all $k \in \{2,\dots,k(n)\}$, we have $\delta_{k-1} \le \rho_k$, $\alpha_k \le 4$ and $\zeta_k \le 1/2$.*

Proof In view of the relations $C_0 n^{-1/(d\vee 4)} = \rho_1 h_1$ and $\rho_{k(n)} h_{k(n)} \ge C_2 n^{-1/3}$, the sequence

$$s_n = 4\sqrt{C_V} C_g h_1 + \frac{4\sigma(c_0\sqrt{\log(Ln)}+c_1 t_n)}{\sqrt{n}\,\rho_{k(n)} h_{k(n)}}$$

tends to zero as $n\to\infty$.

We do now induction on k. Since $s_n \to 0$ as $n\to\infty$ and $\gamma_1 \le s_n$, the inequality $\delta_1 = 2\gamma_1\sqrt{\mu^*} \le 1/\sqrt{2} = \rho_1/\sqrt{2}$ is true for sufficiently large values of n. Let us prove the implication

$$\delta_{k-1} \le \rho_{k-1}/\sqrt{2} \quad\Longrightarrow\quad \begin{cases} \zeta_k \le 1/2, \\ \delta_k \le \rho_k/\sqrt{2}. \end{cases}$$

Since $1/\sqrt{2} \le e^{-1/6}$, the inequality $\delta_{k-1} \le \rho_k/\sqrt{2}$ entails that $\delta_{k-1} \le \rho_k$ and therefore $\alpha_k \le 4$. By our choice of a_h and a_ρ, we have $\rho_1 h_1 \ge \rho_k h_k \ge \rho_{k(n)} h_{k(n)}$. Therefore,

$$\begin{aligned}
\frac{\gamma_k}{\rho_k} &\le 4\sqrt{C_V} C_g \rho_k h_k + \frac{4\sigma(c_0\sqrt{\log(Ln)}+c_1 t_n)}{\sqrt{n}\,\rho_k h_k} \\
&\le 4\sqrt{C_V} C_g h_1 + \frac{4\sigma(c_0\sqrt{\log(Ln)}+c_1 t_n)}{\sqrt{n}\rho_{k(n)} h_{k(n)}} = s_n.
\end{aligned}$$

Thus, for n large enough, $\zeta_k \le 1/2$ and $\gamma_k \le \rho_k/4$. This implies that $\delta_k = 2\gamma_k(1-\zeta_k)^{-1/2} \le \rho_k/\sqrt{2}$.

By induction we infer that $\delta_{k-1} \le \rho_{k-1}/\sqrt{2} \le \rho_k$ and $\zeta_k \le 1/2$ for any $k = 2,\dots,k(n)-1$. This completes the proof of the lemma. ■

Lemma 25 *If $k > 2$ and $\zeta_{k-1} < 1$ then $\Omega_{k-1} \subset \{\mathrm{tr}(I - \hat{A}_{k-1})\Pi^* \le \delta_{k-1}^2\}$.*

Proof Let us denote by $\tilde{\beta}_\ell$ the vector $\Pi^* \hat{\beta}_{\ell,k-1}$, which clearly belongs to $\mathcal{S}$. It holds

$$|P_{k-1}^*(\hat{\beta}_{\ell,k-1} - \beta_\ell)| \le \gamma_{k-1} \quad \Longrightarrow \quad \begin{cases} |\hat{\beta}_{\ell,k-1} - \tilde{\beta}_\ell| \le \gamma_{k-1}, \\ |\tilde{\beta}_\ell - \beta_\ell| \le \sqrt{2}\gamma_{k-1}/\rho_{k-1}. \end{cases}$$

Set $B = \sum_{i=1}^{m^*} \mu_i \bar{\beta}_i \bar{\beta}_i^\top$ and $\tilde{B} = \sum_{i=1}^{m^*} \mu_i \bar{\tilde{\beta}}_i \bar{\tilde{\beta}}_i^\top$, where $\bar{\tilde{\beta}}_i = \sum_\ell c_\ell \tilde{\beta}_\ell$ if $\bar{\beta}_i = \sum_\ell c_\ell \beta_\ell$, see assumption (A2). Since $\sum_\ell |c_\ell| \le 1$, we have $|\bar{\beta}_i| \le \max_\ell |\beta_\ell| \le \|\nabla f\|_\infty$ and $|\bar{\beta}_i - \bar{\tilde{\beta}}_i| \le \max_\ell |\beta_\ell - \tilde{\beta}_\ell|$. Therefore

$$\begin{aligned} \|B - \tilde{B}\| &\le \sum_{i=1}^{m^*} \mu_i \|\bar{\beta}_i \bar{\beta}_i^\top - \bar{\tilde{\beta}}_i \bar{\tilde{\beta}}_i^\top\| \le \mu^* \max_i \|\bar{\beta}_i \bar{\beta}_i^\top - \bar{\tilde{\beta}}_i \bar{\tilde{\beta}}_i^\top\| \\ &\le \mu^* \max_i \Big(|\bar{\beta}_i - \bar{\tilde{\beta}}_i|^2 + 2|\bar{\beta}_i| \cdot |\bar{\beta}_i - \bar{\tilde{\beta}}_i| \Big) \\ &\le \mu^* \big(2\gamma_{k-1}^2 \rho_{k-1}^{-2} + 2\sqrt{2}\gamma_{k-1}\rho_{k-1}^{-1} \max_\ell |\beta_\ell| \big) = \zeta_{k-1} \end{aligned}$$

and hence, for every unit vector $v \in \mathcal{S}$, $v^\top \tilde{B} v \ge \big(v^\top B v - \big| v^\top B v - v^\top \tilde{B} v \big| \big) \ge v^\top B v - \|B - \tilde{B}\| \ge 1 - \zeta_{k-1}$. This inequality implies that $\Pi^* \preceq (1 - \zeta_{k-1})^{-1} \tilde{B}$. Thus the vectors $\tilde{\beta}_\ell$ satisfy assumption (A2) with μ^* replaced by $\mu^*/(1 - \zeta_{k-1})$. Applying Proposition 17 to these vectors we obtain the assertion of the lemma. ■

Acknowledgments

Much of this work has been carried out when the first author was visiting the Weierstrass Institute for Applied Analysis and Stochastics. The financial support from the institute and the hospitality of Professor Spokoiny are gratefully acknowledged.

The authors are grateful to the referees for their constructive comments, which have greatly improved the paper.

References

P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Princeton University Press, Springer, New York, 1998.

E. Bura. Using linear smoothers to assess the structural dimension of regressions. *Statistica Sinica*, 13(1):143–162, 2003.

E. Bura and R. D. Cook. Estimating the structural dimension of regressions via parametric inverse regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(2):393–410, 2001a.

E. Bura and R. D. Cook. Extending sliced inverse regression: The weighted chi-squared test. *J. Amer. Statist. Assoc.*, 96(455):996–1003, 2001b.

K. S. Chan, M. C. Li, and H. Tong. Partially linear reduced-rank regression. *Technical report, available at* `www.stat.uiowa.edu/techrep/tr328.pdf`, 2004.

R. D. Cook. *Regression graphics. Ideas for studying regressions through graphics*. Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, Inc., New York, 1998.

R. D. Cook and B. Li. Dimension reduction for conditional mean in regression. *Ann. Statist.*, 30(2): 455–474, 2002.

R. D. Cook and B. Li. Determining the dimension of iterative hessian transformation. *Ann. Statist.*, 32(6):2501–2531, 2004.

R. D. Cook and L. Ni. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Amer. Statist. Assoc.*, 100(470):410–428, 2005.

R. D. Cook and L. Ni. Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika*, 93(1):65–74, 2006.

R. D. Cook and S. Weisberg. *Applied Regression Including Computing and Graphics*. Hoboken NJ: John Wiley, 1999.

R. D. Cook and S. Weisberg. Discussion of "sliced inverse regression for dimension reduction" by K. C. Li. *J. Amer. Statist. Assoc.*, 86(414):328–332, 1991.

M. Delecroix, M. Hristache, and V. Patilea. On semiparametric *m*-estimation in single-index regression. *J. Statist. Plann. Inference*, 136(3):730–769, 2006.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability, 66, Chapman & Hall, London, 1996.

M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Ann. Statist.*, 29(6):1537–1566, 2001a.

M. Hristache, A. Juditsky, and V Spokoiny. Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3):595–623, 2001b.

K.C. Li. On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma. *J. Amer. Statist. Assoc.*, 87(420):1025–1039, 1992.

K.C. Li. Sliced inverse regression for dimension reduction. with discussion and a rejoinder by the author. *J. Amer. Statist. Assoc.*, 86(414):316–342, 1991.

K.C. Li and N. Duan. Regression analysis under link violation. *Ann. Statist.*, 17(3):1009–1052, 1989.

L. Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.

L. Ni, R. D. Cook, and C.-L. Tsai. A note on shrinkage sliced inverse regression. *Biometrika*, 92 (1):242–247, 2005.

A. Samarov, V. Spokoiny, and C. Vial. Component identification and estimation in nonlinear high-dimensional regression models by structural adaptation. *J. Amer. Statist. Assoc.*, 100(470):429–445, 2005.

H. Wang, L. Ni, and C.-L. Tsai. Improving dimension reduction via contour- projection. *Statistica Sinica*, 18:299–311, 2008.

Y. Xia. A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.*, 35(6):2654–2690, 2007.

Y. Xia, H. Tong, W. K. Li, and L. X. Zhu. An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):363–410, 2002.

X. Yin and R. D. Cook. Direction estimation in single-index regressions. *Biometrika*, 92(2):371–384, 2005.

X. Yin and R. D. Cook. Dimension reduction via marginal high moments in regression. *Statist. Probab. Lett.*, 76(4):393–400, 2006.