HAL open science

A new algorithm for estimating the effective dimension-reduction subspace

Arnak S. Dalalyan, Anatoli B. Juditsky, Vladimir Spokoiny

To cite this version:

Arnak S. Dalalyan, Anatoli B. Juditsky, Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. Journal of Machine Learning Research, 2008, 9, pp.1647-1678. hal-00128129v2

HAL Id: hal-00128129

https://hal.science/hal-00128129v2

Submitted on 18 Apr 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journal of Machine Learning Research (April 17, 2007) Submitted ; Published

A new algorithm for estimating the effective dimension-reduction subspace

Arnak S. Dalalyan DALALYAN@CCR.JUSSIEU.FR
University Paris 6
Laboratoire de Probabilités
Université Paris 6, Boîte courrier 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
Mohrenstr. 39
10117 Berlin Germany

Editor: Leslie Pack Kaelbling

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. 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.The empirical behavior of the algorithm is studied through numerical simulations.

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

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 $X^{(1)}, \ldots, X^{(d)}$ 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 $X = (X^{(1)}, \ldots, X^{(d)})$, 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

©April 17, 2007 Arnak S. Dalalyan, Anatoly Juditsky and Vladimir Spokoiny.

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 $\vartheta_1, \ldots, \vartheta_{m^*}$ and for some function $g : \mathbb{R}^{m^*} \to \mathbb{R}$, the relation $f(x) = g(\vartheta_1^\top x, \ldots, \vartheta_{m^*}^\top x)$ 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 $\vartheta_1, \ldots, \vartheta_{m^*}$ 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 assesses 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; Cook, 1998) for more details.

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

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

where $\varepsilon_1, \ldots, \varepsilon_n$ 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 $\vartheta_1, \ldots, \vartheta_{m^*}$ 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 adaption via maximum minimization (SAMM). It can be regarded as a branch of the structure-adaptive (SA) approach introduced in (Hristache et al., 2001b,a).

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, 2001), SA approach (Hristache et al., 2001a), iterative Hessian transformation (Cook and Li, 2002), minimum average variance estimation (MAVE) (Xia et al., 2002) and minimum discrepancy approach (Cook and Ni, 2005).

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 (cf. (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 are recognized to provide more accurate estimates of the EDR subspace.

These arguments, combined with the 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 or using extra-statistical arguments 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 leading to the algorithm proposed in (Hristache et al., 2001a) is to iteratively estimate L linear combinations $\beta_1, \ldots, \beta_L$ of vectors $\nabla f(X_1), \ldots, \nabla f(X_n)$ and then to recover 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 on 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, the 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 the 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 is to replace 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).

Some aspects of the application of dimension reduction techniques in bioinformatics are studied in (Antoniadis et al., 2003) and (Bura and Pfeiffer, 2003).

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

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

where $X_{ij} = X_j - X_i$, b is a *bandwidth* and $K(\cdot)$ is a univariate kernel supported on $[0,1]$. The bandwidth b should be selected so that the ball with the radius b and the center at the point of estimation X_i contains at least $d+1$ design points. For large value of d this leads to a bandwidth of order one and to a large 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 to 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 $\vartheta_1, \ldots, \vartheta_{m^*}$, 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 $\hat{\mathcal{S}}$ of $\mathcal{S}$ is available, the orthogonal projector $\widehat{\Pi}$ onto the subspace $\hat{\mathcal{S}}$ may replace Π^* in the expression (2). This rule of defining the local neighborhoods is 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 $w_{ij} = K(X_{ij}^\top (I + \rho^{-2}\widehat{\Pi}) X_{ij}/h^2)$ 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 $\widehat{\Pi}$. If we are very confident in our estimator $\widehat{\Pi}$, 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 $\nabla f(X_i)$, $i = 1, \ldots, n$. 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 $\{\psi_\ell, \ell = 1, \ldots, L\}$ be a system of functions on $\mathbb{R}^d$ satisfying the conditions $n^{-1} \sum_{i=1}^{n} \psi_\ell(X_i) \psi_{\ell'}(X_i) = \delta_{\ell,\ell'}$ for every $\ell, \ell' \in \{1, \ldots, L\}$, with $\delta_{\ell,\ell'}$ being the Kronecker symbol. Define

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

and $\mathscr{M}_L = \sum_{\ell=1}^{L} \beta_\ell \beta_\ell^\top$. By the Bessel inequality, it holds $\mathscr{M}_L \le \mathscr{M}$. 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 $\mathrm{rank}(\mathscr{M}_L) = m^*$ 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 $\mathscr{M}_L$ have a risk of order $\sqrt{L/n}$ (see (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:

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

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:

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

From now on, for two symmetric matrices A and B, $A \preceq B$ means that $B - A$ is semidefinite positive. Define $\widehat\Pi_m$ as a minimizer of the maximum of the $\hat\beta_\ell^\top (I - \Pi)\hat\beta_\ell$'s instead of their sum:

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

This is a convex optimization problem that can be effectively solved even for a large d although a closed form solution is not known. 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 $(Y_1, X_1), \ldots, (Y_n, X_n)$ from the model

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

where $\varepsilon_1, \ldots, \varepsilon_n$ 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 $X_1, \ldots, X_n$. Henceforth, without loss of generality we assume that $|X_i| < 1$ for any $i = 1, \ldots, n$, where $|v|$ denotes the Euclidian norm of the vector v.

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 $|\psi_\ell|^2 = n$. Set $k = 1$.

b) Initialize the parameters $h = h_1$, $\rho = \rho_1$ and $\widehat{\Pi}_0 = 0$.

c) Define the estimators $\widehat{\nabla f}(X_i)$ for $i = 1, \ldots, n$ by formula (3) with the current values of h, ρ and $\widehat{\Pi}$. Set

$$\hat{\beta}_\ell = \frac{1}{n} \sum_{i=1}^{n} \widehat{\nabla f}(X_i) \psi_{\ell,i}, \qquad \ell = 1, \ldots, L, \tag{7}$$

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

d) Define the new value $\widehat{\Pi}_k$ as the solution to (6).

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 $\rho < \rho_{\min}$ or $h > h_{\max}$, otherwise continue with the step c).

Let $k(n)$ be the total number of iterations. The matrix $\widehat{\Pi}_{k(n)}$ is the desired estimator of the projector Π^*. We denote by $\widehat{\Pi}_n$ the orthogonal projection onto the space spanned by the eigenvectors of $\widehat{\Pi}_{k(n)}$ corresponding to the m^* largest eigenvalues. The estimator of the EDR subspace is then the image of $\widehat{\Pi}_n$. Equivalently, $\widehat{\Pi}_n$ is the estimator of the projector onto $\mathcal{S}$.

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. It also takes 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 $A^\top \cdot A$ and $\|A\|_2^2$ for the sum of squares of all elements of the matrix A.

(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 for each 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 $\bar\beta_1, \ldots, \bar\beta_{m^*} \in \mathcal{B}^*$ and constants $\mu_1, \ldots, \mu_{m^*}$ such that

$$\Pi^* \preceq \sum_{k=1}^{m^*} \mu_k \bar\beta_k \bar\beta_k^\top . \tag{8}$$

We denote $\mu^* = \mu_1 + \ldots + \mu_k$.

Remark 1 *Assumption (A2) implies that the subspace $\mathcal{S} = Im(\Pi^*)$ 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 $\Psi = (\psi_1, \ldots, \psi_L) \in \mathbb{R}^{n \times L}$, $\boldsymbol{\nabla f} = (\nabla f(X_1), \ldots, \nabla f(X_n)) \in \mathbb{R}^{d \times n}$ and write the $d \times L$ matrix $B = (\beta_1, \ldots, \beta_L)$ in the form $\boldsymbol{\nabla f} \cdot \Psi$. 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 $\tilde\beta_1, \ldots, \tilde\beta_{m^*}$ 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 $v_1, \ldots, v_{m^*}$ 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

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

Hence, inequality (8) is fulfilled with $\mu_k = 1/\lambda_{m^*}(\tilde{\mathscr{M}})$ for every $k = 1, \ldots, m^*$. ■

These arguments show that the identifiability assumption (A2) is not too stringent. 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.

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

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

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

$$\|\tilde{V}_i^{(k)}(U)^{-1}\| N_i^{(k)}(U) \le C_V, \qquad i = 1, \dots, n, \tag{9}$$

$$\sum_{i=1}^n w_{ij}^{(k)}(U)/N_i^{(k)}(U) \le C_K, \qquad j = 1, \dots, n, \tag{10}$$

$$\sum_{i=1}^n |\bar{w}_{ij}^{(k)}(U)|/N_i^{(k)}(U) \le C_{K'}, \qquad j = 1, \dots, n, \tag{11}$$

$$\sum_{j=1}^n |\bar{w}_{ij}^{(k)}(U)|/N_i^{(k)}(U) \le C_w \qquad i = 1, \dots, n, \tag{12}$$

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

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

3.3 Main result

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

$$\max_{\ell=1,\dots,L} \max_{i=1,\dots,n} |\psi_{\ell,i}| < \bar{\psi}, \tag{13}$$

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

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

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

Corollary 4 *Under the assumptions of Theorem 3, for sufficiently large n, it holds*

$$\mathbf{P}\Big(\|\widehat{\Pi}_n - \Pi^*\|_2 > Cn^{-\frac{2}{3\vee m^*}} t_n^2 + \frac{2\sqrt{2\mu^*} z c_0 \sigma}{\sqrt{n(1-\zeta_n)}}\Big) \le Lze^{-\frac{z^2-1}{2}} + \frac{3k(n)-5}{n}$$

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

Proof Easy algebra yields

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

The equality $\operatorname{tr}\Pi^* = m^*$ 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 $\|\widehat{\Pi}_n - \Pi^*\|_2^2 \le 2m^*$. ■

These results assess that for $m^* \le 4$, the estimator of $\mathcal{S}$ 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 5 *The inspection of the proof of Theorem 3 shows that the factor t_n^2 multiplying the "bias" term $n^{-2/(3\vee m^*)}$ disappears when $m^* > 3$.*

Remark 6 *The same rate of convergence remains valid in the case when the errors are not necessarily identically distributed Gaussian random variables, but have (uniformly in n) a bounded exponential moment. This can be proved along the lines of Proposition 14, see Section 5.*

Remark 7 *Note that in (A3) we implicitly assumed that the matrices $\tilde{V}_i^{(k)}$ are invertible, which may be true only if any neighborhood $E^{(k)}(X_i) = \{x : |(I+\rho_k^{-2}\Pi^*)^{-1/2}(X_i - x)| \le h_k\}$ 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 $\tilde{V}_i$ a small full-rank matrix to be sure that the resulting matrix is invertible, see Section 4.*

Remark 8 *In the case when $m = n^{1/d}$ is integer, an example of deterministic design satisfying (A3) is as follows. Choose d functions $h_k : [0,1] \to [0,\infty[$ such that $\inf_{[0,1]} h_k(x) > 0$ and $\sup_{[0,1]} h_k(x) < \infty$. Define the design points*

$$X_i = \Big(\int_{i_1/m}^{1+i_1/m} h_1(x)\,dx, \dots \int_{i_d/m}^{1+i_d/m} h_d(x)\,dx\Big),$$

where $i_1, \ldots, i_d$ range over $\{0, \ldots, m-1\}$. This definition guarantees that the number of design points lying in an ellipsoid E is asymptotically of the same order as $nVol(E)$, as $n \to \infty$. This suffices for (A3). Of course, it is unlikely to have such a design in practice, since even for small m and moderate d it leads to an unrealistically large sample size.

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".

The computer code of the procedure SAMM is distributed freely, it can be downloaded from *http://www.proba.jussieu.fr/pageperso/dalalyan/*. It requires the MATLAB packages SDPT3 and LMI.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 $\sum_{j=1}^n \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^\top w_{ij}$ in (3).

In all examples presented below the number of replications is $N = 250$. 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 $\mathfrak{S}_j$, i.e. $\mathfrak{S}_j(\mathfrak{S}_j^{-1}(k)) = k$ for every $k = 1, \ldots, n$. Define $\{\psi_\ell\}$ as the set of vectors

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

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

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)

normalized to satisfy $\sum_{i=1}^n \psi_{\ell,i}^2 = n$ for every ℓ. It is easily seen that these vectors satisfy conditions (13) 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.

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 run SAMM and MAVE 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 on $[-1, 1]$, 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 estimator based on MAVE. We plot in Figure 1 (a) the average loss normalized by the square rood 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 (a) decreases with a strong slope.

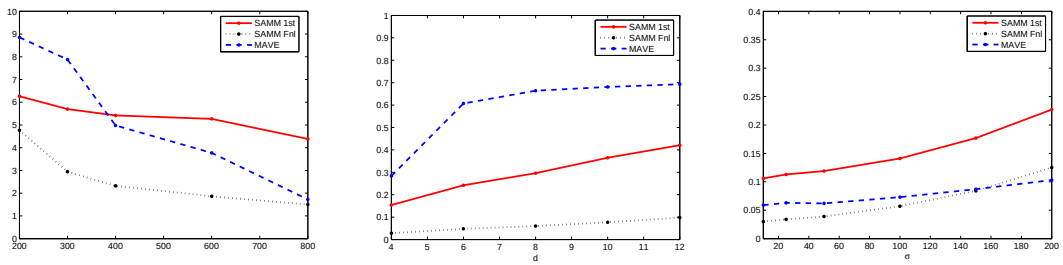
Example 2 (Double-index)

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

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



(a) **(b)** **(c)**

Figure 1: (a) Average loss multiplied by $\sqrt{n}$ versus n for the first step (full line) and the final (dotted line) estimators provided by SAMM and for the estimator based on MAVE (broken line) in Example 1, (b) (resp. (c)) Average loss versus d (resp. σ) for the first step (full line) and the final (dotted line) estimators provided by SAMM and for the estimator based on MAVE (broken line) in Example 2 (resp. Example 3).

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

$\boldsymbol{d}$	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)

where the design X is such that the coordinates $(X_i^{(j)}, j \leq d, i \leq n)$ are i.i.d. uniform on $[-40, 40]$, 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 1(b)) the quality of SAMM deteriorated linearly in d as d increased. This agrees with our theoretical results. It should be noted that in this case MAVE fails to find the EDR space.

Example 3

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

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

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.*

$\boldsymbol{\sigma}$	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)

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

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

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

Figure 1(c) 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).

Example 4 (time series)

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

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

with initial variables $T_1, \ldots, T_6$ 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

$$\begin{aligned} g(x) &= -1 + 0.6x_1 - \cos(0.5\pi x_2) + e^{-x_3^2}, \\ \vartheta_1 &= (1,0,0,2,0,0)/\sqrt{5}, \\ \vartheta_2 &= (0,0,1,0,0,2)/\sqrt{5}, \\ \vartheta_3 &= (-2,2,-2,1,-1,1)/\sqrt{15}. \end{aligned}$$

We run SAMM and MAVE procedures on the data (X_i, Y_i), $i = 1, \ldots, 250$, where $Y_i = T_{i+6}$ and $X_i = (T_i, \ldots, T_{i+5})^\top$. The results of simulations reported in Table 4 show that the qualities of SAMM and MAVE are comparable, with SAMM being slightly more performant.

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

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)

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 5.1 and 5.2. 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 3 is presented in Subsection 5.3, while some technical lemmas are postponed to Subsection 5.6.

5.1 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 $\hat{\beta}_{1,\ell}$ of the vector β_ℓ^* is the same as the one used in (Hristache et al., 2001a).

Proposition 9 *Under assumptions (A1),(A3), (A4) and (13), for every* $\ell \le L$,

$$|\hat{\beta}_{1,\ell} - \beta_\ell| \le h_1 C_g \sqrt{2C_V} + \frac{\xi_{1,\ell}}{h_1\sqrt{n}},$$

where $\xi_{1,\ell}$ *is a zero mean normal vector verifying* $\mathbf{E}|\xi_{1,\ell}|^2 \le 2d\sigma^2 C_V C_K \bar{\psi}^2$.

Proof Since at the first iteration we take $S_1 = I$, the inequality $|S_1 X_{ij}| \le h_1$ implies that $|\Pi^* X_{ij}| \le |X_{ij}| \le h_1$. Therefore the bias term $|P_1^*(E\hat{\beta}_{1,\ell} - \beta_\ell)|$ is bounded by $h_1 C_g \sqrt{C_V}$ (cf. the proof of Proposition 12).

For the stochastic term, we set $\xi_{1,\ell} = h_1\sqrt{n}(\hat{\beta}_{1,\ell} - E\hat{\beta}_{1,\ell})$. By Lemma 21, we have $\mathbf{E}|P_1^*\xi_{1,\ell}|^2 \le d\sigma^2 C_V C_K \bar{\psi}^2$. The assertion of the proposition follows now from $P_1^* = (I + \rho_1^{-2}\Pi^*)^{-1/2} \succeq I/\sqrt{2}$. ■

Corollary 10 *If $nL \geq 6$ and the assertions of Proposition 9 hold, then*

$$\mathbf{P}\bigg(\max_{\ell} |\hat{\beta}_{1,\ell} - \beta_\ell| \geq h_1 C_g \sqrt{C_V} + \frac{2\sqrt{2dC_V C_K \log(nL)}\,\sigma\bar{\psi}}{h_1\sqrt{n}}\bigg) \leq \frac{1}{n}.$$

Remark 11 *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 $\{X_i, i = 1, \ldots, n\}$. 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 $h_1 \geq n^{-1/d}$ leads to the choice $h_1 = Const.n^{-1/(4\vee d)}$.*

5.2 One step improvement

At the kth step of iteration, we have at our disposal a symmetric matrix $\Pi \in \mathcal{M}_{d\times d}$ belonging to the set

$$\mathscr{P}_\delta(\Pi^*) = \Big\{\Pi \in \mathcal{M}_{d\times d}:\ \operatorname{tr}\Pi \leq m^*, \quad 0 \preceq \Pi \preceq I, \quad \operatorname{tr}(I-\Pi)\Pi^* \leq \delta^2\Big\}$$

Thus the matrix Π is the kth step approximation of the projector Π^* onto the EDR subspace $\mathcal{S}^*$. Using this approximation, we construct the new matrix $\hat{\Pi}$ in the following way: Set $S_{\Pi,\rho} = (I + \rho^{-2}\Pi)^{1/2}$, $P_{\Pi,\rho} = S_{\Pi,\rho}^{-1}$ and define the estimator of the regression function and its gradient at the design point X_i as follows:

$$\begin{pmatrix} \hat{f}_\Pi(X_i) \\ \widehat{\nabla f}_\Pi(X_i) \end{pmatrix} = V_i(\Pi)^{-1} \sum_{j=1}^n Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} w_{ij}(\Pi),$$

where $w_{ij}(\Pi) = K\big(h^{-2}|S_{\Pi,\rho}X_{ij}|^2\big)$ and

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

To state the next result, we need some additional notation. Set $Z_{ij} = (hP_\rho^*)^{-1}X_{ij}$, $U = P_\rho^* S_{\Pi,\rho}^2 P_\rho^*$ and $U^* = I$, where $P_\rho^* = P_{\Pi^*,\rho} = (I - \Pi^*) + \rho(1+\rho^2)^{-1/2}\,\Pi^*$. In this notation, we obtain

$$\begin{aligned}
\begin{pmatrix} h^{-1}\hat{f}_\Pi(X_i) \\ P_\rho^*\widehat{\nabla f}_\Pi(X_i) \end{pmatrix} &= \begin{pmatrix} h^{-1} & 0 \\ 0 & P_\rho^* \end{pmatrix} V_i(\Pi)^{-1} \sum_{j=1}^n Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} w_{ij}(\Pi) \\
&= \frac{1}{h}\begin{pmatrix} 1 & 0 \\ 0 & hP_\rho^* \end{pmatrix} V_i(\Pi)^{-1} \begin{pmatrix} 1 & 0 \\ 0 & hP_\rho^* \end{pmatrix} \sum_{j=1}^n Y_j \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} w_{ij}(\Pi) \\
&= h^{-1}\tilde{V}_i(U)^{-1} \sum_{j=1}^n Y_j \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} w_{ij}(U)
\end{aligned}$$

where $w_{ij}(U) = K\big(Z_{ij}^\top U Z_{ij}\big)$ and

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

Set $N_i(U) = \sum_j w_{ij}(U)$ and $\alpha = 2\delta^2\rho^{-2} + 2\delta\rho^{-1}$.

Proposition 12 *If (A1)-(A4) are fulfilled then there exist Gaussian vectors $\xi_1^*, \dots, \xi_L^* \in \mathbb{R}^d$ such that $\mathbf{E}[|\xi_\ell^*|^2] \le c_0^2\sigma^2$ and*

$$\mathbf{P}\Big(\sup_{\Pi,\ell}\Big|P_\rho^*(\hat\beta_{\ell,\Pi}-\beta_\ell)-\frac{\xi_\ell^*}{h\sqrt{n}}\Big| \ge \sqrt{C}_V C_g(\rho+\delta)^2 h + \frac{c_1\sigma\alpha t_n}{h\sqrt{n}}\Big) \le \frac{2}{n},$$

where the sup *is taken over $\Pi \in \mathscr{P}_\delta$, $\ell = 1,\dots,L$ and we used the notation $t_n = 5 + \sqrt{3\log(Ln) + \frac{3}{2}d^2\log n}$, $c_0 = \bar\psi\sqrt{dC_KC_V}$ and $c_1 = 30\bar\psi(C_w^2C_V^4C_K^2 + C_V^2C_{K'}^2)^{1/2}$.*

Proof Let us start with evaluating the bias term $|P_\rho^*(\mathbf{E}\hat\beta_{\ell,\Pi}-\beta_\ell)|$. According to the Cauchy-Schwarz inequality, it holds

$$\begin{aligned}
\big|P_\rho^*\big(\mathbf{E}\hat\beta_{\ell,\Pi}-\beta_\ell\big)\big|^2 &= n^{-2}\Big|\sum_{i=1}^n P_\rho^*\big(\mathbf{E}[\widehat{\nabla f}_\Pi(X_i)]-\nabla f(X_i)\big)\psi_\ell(X_i)\Big|^2\\
&\le \frac{1}{n^2}\sum_{i=1}^n\big|P_\rho^*\big(\mathbf{E}[\widehat{\nabla f}_\Pi(X_i)]-\nabla f(X_i)\big)\big|^2\sum_{i=1}^n\psi_l^2(X_i)\\
&\le \max_{i=1,\dots,n}\big|P_\rho^*\big(\mathbf{E}[\widehat{\nabla f}_\Pi(X_i)]-\nabla f(X_i)\big)\big|^2.
\end{aligned}$$

Simple computations show that

$$\begin{aligned}
&\big|P_\rho^*\big(\mathbf{E}[\widehat{\nabla f}_\Pi(X_i)]-\nabla f(X_i)\big)\big|\\
&\qquad\le \left|\mathbf{E}\begin{pmatrix} h^{-1}\hat f_\Pi(X_i)\\ P_\rho^*\widehat{\nabla f}_\Pi(X_i)\end{pmatrix} - \begin{pmatrix} h^{-1}f(X_i)\\ P_\rho^*\nabla f(X_i)\end{pmatrix}\right|\\
&\qquad= h^{-1}\left|\tilde V_i^{-1}\sum_{j=1}^n f(X_j)\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}w_{ij}(U) - \begin{pmatrix} h^{-1}f(X_i)\\ P_\rho^*\nabla f(X_i)\end{pmatrix}\right|\\
&\qquad= h^{-1}\Big|\tilde V_i^{-1}\sum_{j=1}^n r_{ij}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}w_{ij}(U)\Big| := b(X_i),
\end{aligned}$$

where $r_{ij} = f(X_j) - f(X_i) - X_{ij}^\top\nabla f(X_i)$. Define $v_j = \tilde V_i^{-1/2}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\sqrt{w_{ij}(U)}$ and $\lambda_j = h^{-1}r_{ij}\sqrt{w_{ij}(U)}$. Then

$$b(X_i) = \Big|\tilde V_i^{-1/2}\sum_{j=1}^n\lambda_jv_j\Big| \le \|\tilde V_i^{-1/2}\|\cdot|\lambda|\cdot\Big\|\sum_{j=1}^n v_jv_j^\top\Big\|^{1/2}.$$

The identity $\sum_j v_jv_j^\top = I_{d+1}$ implies

$$\begin{aligned}
b(X_i)^2 &\le \frac{1}{h^2}\ \Big\|\tilde V_i^{-1/2}\Big\|^2\cdot\sum_{j=1}^n r_{ij}^2w_{ij}(U)\\
&\le h^{-2}\max_j r_{ij}^2\ \Big\|\tilde V_i^{-1}\Big\|\cdot\sum_{j=1}^n w_{ij}(U)\\
&\le C_Vh^{-2}\max_j r_{ij}^2\,,
\end{aligned}$$

where the maximum of r_{ij} is taken over the indices j satisfying $w_{ij}(U) \neq 0$. Since the weights w_{ij} are defined via the kernel function K vanishing on the interval $[1, \infty[$, we havemax$_j\, r_{ij} = \max\{r_{ij} : |S_{\Pi,\rho} X_{ij}| \le h\}$. By Corollary 18 $|S_{\Pi,\rho} X_{ij}| \le h$ implies $|\Pi^* X_{ij}| \le (\rho+\delta)h$. Let us denote by Θ the $(d \times m^*)$ matrix having ϑ_k as kth column. Then $\Pi^* = \Theta\Theta^\top$ and therefore

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

These estimates yield $|b(X_i)| \le \sqrt{C_V}\, C_g(\rho+\delta)^2 h$, and consequently,

$$\big|P_\rho^*\big(\mathbf{E}\hat\beta_{\ell,\Pi} - \beta_\ell\big)\big| \le \max_i b(X_i) \le \sqrt{C_V}\, C_g(\rho+\delta)^2 h. \tag{14}$$

Let us treat now the stochastic term $P_\rho^*\big(\hat\beta_{\ell,\Pi} - \beta_l^*\big)$. It can be bounded as follows

$$\big|P_\rho^*\big(\hat\beta_{\ell,\Pi} - \mathbf{E}\hat\beta_{\ell,\Pi}\big)\big| \le \Big|\sum_{j=1}^n c_{j,\ell}(U)\,\varepsilon_j\Big|,$$

where

$$c_{j,\ell}(U) = \frac{1}{hn}\sum_{i=1}^n \tilde V_i^{-1}(U)\begin{pmatrix}1\\ Z_{ij}\end{pmatrix} w_{ij}(U)\psi_\ell(X_i).$$

Let us define $\xi_\ell^* = h\sqrt{n}\, P_\rho^*(\hat\beta_{\ell,\Pi^*} - \mathbf{E}[\hat\beta_{\ell,\Pi^*}])$. In view of Lemma 21, we have $\mathbf{E}[|\xi_\ell^*|^2] \le nh^2\sigma^2 \sum_j |c_{j,\ell}(U^*)|^2 \le c_0^2\sigma^2$.

One checks that for any $\ell = 1,\dots,L$ and for any Π such that $\mathrm{tr}(I-\Pi)\Pi^* \le \delta^2$, it holds

$$\Big|P_\rho^*(\hat\beta_{\ell,\Pi} - \mathbf{E}[\hat\beta_{\ell,\Pi}]) - \frac{\xi_\ell^*}{h\sqrt{n}}\Big| \le \sup_{\|U-U^*\|_2\le\alpha}\Big|\sum_{j=1}^n \big(c_{j,\ell}(U) - c_{j,\ell}(U^*)\big)\,\varepsilon_j\Big|.$$

Set $a_{j,\ell}(U) = c_{j,\ell}(U) - c_{j,\ell}(U^*)$. Lemma 22 implies that Proposition 14 can be applied with $\kappa_0 = \frac{c_1\alpha}{h\sqrt{n}}$ and $\kappa_1 = \frac{c_1}{h\sqrt{n}}$. Setting $\epsilon = 2\alpha/\sqrt{n}$ we get that the probability of the event

$$\Big\{\sup_{U,\ell}\Big|\sum_{j=1}^n \big(c_{j,\ell}(U) - c_{j,\ell}(U^*)\big)\,\varepsilon_j\Big| \ge \frac{c_1\sigma\alpha(5+\sqrt{3\log(Ln)+3d^2\log(\sqrt{n})})}{h\sqrt{n}}\Big\}$$

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

Corollary 13 *If $nL \ge 6$ and the assumptions of Proposition 12 are fulfilled, then*

$$\mathbf{P}\Big(\sup_{\ell,\Pi}\big|P_\rho^*(\hat\beta_{\ell,\Pi} - \beta_\ell)\big| \ge \sqrt{C}_V C_g(\rho+\delta)^2 h + \frac{\sigma(zc_0 + c_1\alpha t_n)}{h\sqrt{n}}\Big) \le Lze^{-\frac{z^2-1}{2}}.$$

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

$$\Big\{\sup_{\ell,\Pi}\big|P_\rho^*(\hat\beta_{\ell,\Pi} - \beta_\ell)\big| \ge \sqrt{C}_V C_g(\rho+\delta)^2 h + \frac{\sigma(2c_0\sqrt{\log(Ln)} + c_1\alpha t_n)}{h\sqrt{n}}\Big\}$$

does not exceed $3/n$, where sup *is taken over all $\Pi \in \mathscr{P}_\delta(\Pi^*)$, $\ell = 1,\dots,L$ and c_0, c_1, t_n are defined in Proposition 12 and in Theorem 3.*

Proof In view of Lemma 7 in (Hristache et al., 2001b) and Lemma 21 , we have

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

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

5.3 Proof of Theorem 3

Recall that at the first step we use the following values of parameters: $\widehat{\Pi}_0=\mathbf{0}$, $\rho_1=1$ and $h_1=n^{-1/(d\vee 4)}$. Let us denote

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

and introduce the event $\Omega_1=\{\max_\ell|\hat{\beta}_{1,\ell}-\beta_\ell|\le\gamma_1\}$. According to Corollary 10 the probability of the event Ω_1 is at least $1-n^{-1}$. In view of Proposition 16, we get $\mathbf{P}(\mathrm{tr}(I-\widehat{\Pi}_1)\Pi^*\le\delta_1^2)\ge 1-n^{-1}$.

For any integer $k\in[2,k(n)]$ (where $k(n)$ is the total number of iterations), we define

$$\begin{aligned}
&\rho_k=a_\rho\rho_{k-1},\quad h_k=a_hh_{k-1},\quad \alpha_k=\frac{2\delta_{k-1}}{\rho_k}\Big(\frac{\delta_{k-1}}{\rho_k}+1\Big),\\
&\gamma_k=\begin{cases}C_g\sqrt{C}_V\,(\rho_k+\delta_{k-1})^2h_k+\dfrac{\sigma(2c_0\sqrt{\log(nL)}+c_1\alpha_kt_n)}{h_k\sqrt{n}}, & k<k(n),\\ C_g\sqrt{C}_V\,(\rho_k+\delta_{k-1})^2h_k+\dfrac{\sigma(zc_0+c_1\alpha_kt_n)}{h_k\sqrt{n}}, & k=k(n),\end{cases}\\
&\zeta_k=2\mu^*(\gamma_k^2\rho_k^{-2}+\sqrt{2}\,\gamma_k\rho_k^{-1}C_g),\\
&\delta_k=2\gamma_k\sqrt{\mu^*}/\sqrt{1-\zeta_k},\\
&\Omega_k=\{\max_\ell|P_k^*(\hat{\beta}_{k,\ell}-\beta_\ell)|\le\gamma_k\}.
\end{aligned}$$

Here $\hat{\beta}_{k,\ell}=\frac{1}{n}\sum_{i=1}^n\widehat{\nabla f}^{(k)}(X_i)\psi_\ell(X_i)$ with

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

and $w_{ij}^{(k)}=K\big(h_k^{-2}|(I+\rho_k^{-2}\widehat{\Pi}_{k-1})^{1/2}X_{ij}|^2\big)$.

Combining Lemmas 23 and 24, we obtain $\mathbf{P}(\mathrm{tr}(I-\widehat{\Pi}_{k-1})\Pi^*>\delta_{k-1}^2)\le\mathbf{P}(\Omega_{k-1}^c)$ and therefore, using Corollary 13, we get

$$\begin{aligned}
\mathbf{P}\big(\Omega_k^c\big)&\le\mathbf{P}\Big(\max_\ell|P_k^*(\hat{\beta}_{k,\ell}-\beta_\ell)|>\gamma_k,\ \mathrm{tr}(I-\widehat{\Pi}_{k-1})\Pi^*\le\delta_{k-1}^2\Big)+\mathbf{P}\big(\Omega_{k-1}^c\big)\\
&\le\mathbf{P}\Big(\sup_{\Pi\in\mathscr{P}_{m^*,\delta_{k-1}}}\max_\ell|P_k^*(\hat{\beta}_{k,\ell}-\beta_\ell)|>\gamma_k\Big)+\mathbf{P}\big(\Omega_{k-1}^c\big)\\
&\le\frac{3}{n}+\mathbf{P}\big(\Omega_{k-1}^c\big),\qquad k\le k(n)-1.
\end{aligned}$$

Since $\mathbf{P}(\Omega_1^c) \le 1/n$, it holds $\mathbf{P}(\Omega_{k(n)-1}^c) \le (3k(n)-5)/n$ and $\mathbf{P}(\Omega_{k(n)}^c) \le Lze^{-(z^2-1)/2} + \frac{3k(n)-5}{n}$. Lemma 24 implies that

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

According to Lemma 23, 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 the definition of $\delta_{k(n)}$ and Lemma 19 (see below).

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.

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*

$$\sup_{\gamma\in\Gamma}\sup_{|u-u^*|\le r}\sum_{j=1}^n |a_{j,\gamma}(u)|^2 \le \kappa_0^2, \tag{15}$$

$$\sup_{\gamma\in\Gamma}\sup_{|u-u^*|\le r}\sup_{\boldsymbol{e}\in S_{d-1}}\sum_{j=1}^n \left|\frac{d}{du}\,(\boldsymbol{e}^\top a_{j,\gamma}(u))\right|^2 \le \kappa_1^2. \tag{16}$$

If the ε_j's are independent $\mathcal{N}(0,\sigma^2)$-distributed random variables, then

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

where $t = \sqrt{3\log(|\Gamma|(2r/\epsilon)^p n)}$.

Proof Let B_r be the ball $\{u : |u - u^*| \le r\} \subset \mathbb{R}^p$ and $\Sigma_{r,\epsilon}$ be the ϵ-net on B_r such that for any $u \in B_r$ there is an element $u_l \in \Sigma_{r,\epsilon}$ such that $|u - u_l| \le \epsilon$. It is easy to see that such a net with cardinality $N_{r,\epsilon} < (2r/\epsilon)^p$ 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

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

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

$$\begin{aligned}\big|\eta_\gamma(u) - \eta_\gamma(u')\big|^2 &= \sup_{\boldsymbol{e}\in S_{d-1}} \big|\boldsymbol{e}^\top\big(\eta_\gamma(u) - \eta_\gamma(u')\big)\big|^2\\ &\le |u-u'|^2 \cdot \sup_{u\in B_r}\sup_{\boldsymbol{e}\in S_{d-1}} \left|\frac{d(\boldsymbol{e}^\top\eta_\gamma)}{du}\,(u)\right|^2\\ &= |u-u'|^2 \cdot \sup_{u\in B_r}\sup_{\boldsymbol{e}\in S_{d-1}} \left|\sum_{j=1}^n \frac{d(\boldsymbol{e}^\top a_{j,\gamma})}{du}\,(u)\,\varepsilon_j\right|^2\\ &\le |u-u'|^2 \cdot \sup_{u\in B_r}\sup_{\boldsymbol{e}\in S_{d-1}} \sum_{j=1}^n \left|\frac{d(\boldsymbol{e}^\top a_{j,\gamma})}{du}\,(u)\right|^2 \sum_{j=1}^n \varepsilon_j^2\\ &\le \kappa_1^2|u-u'|^2\sum_{j=1}^n \varepsilon_j^2.\end{aligned}$$

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

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

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 $\widehat{\Pi}$ 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 assume that β_ℓs are defined by (4). In fact, we will apply the results of this subsection to the vectors $\Pi^*\hat{\beta}_\ell$.

Denote

$$R(\Pi) = \max_\ell \hat{\beta}_\ell^\top (I - \Pi)\hat{\beta}_\ell,$$
$$\hat{\mathcal{R}}(m) = \min_{\Pi \in \mathcal{A}_m} \sqrt{R(\Pi)} = \sqrt{R(\widehat{\Pi}_m)}.$$

We also define

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

and denote by Π_m^* a minimizer of $\max_\ell \beta_\ell^\top (I - \Pi)\beta_\ell$ over $\Pi \in \mathcal{A}_m$. Since for $m \geq m^*$ the projector Π^* is in $\mathcal{A}_m$, we have $\Pi_m^* = \Pi^*$ and $\mathcal{R}^*(m) = 0$.

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

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

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

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

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

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

$$\max_\ell |(I - \widehat{\Pi}_m)^{1/2}\hat{\beta}_\ell| \leq \max_\ell |(I - \Pi_m^*)^{1/2}\hat{\beta}_\ell| \leq \mathcal{R}^*(m) + \varepsilon.$$

Denote $A = (I - \widehat{\Pi}_m)^{1/2}$. From definition $0 \preceq A \preceq I$. Therefore, for every ℓ

$$|A\beta_\ell| \leq |A\hat{\beta}_\ell| + |A(\beta_\ell - \hat{\beta}_\ell)| \leq |A\hat{\beta}_\ell| + |\beta_\ell - \hat{\beta}_\ell| \leq \mathcal{R}^*(m) + 2\varepsilon.$$

The second inequality of the proposition follows now from $|A\bar{\beta}| \leq \max_\ell |A\beta_\ell|$ for every $\bar{\beta} \in \mathcal{B}^*$.

To prove the last assertion, remark that according to the definition of $\mathcal{R}^*(m)$, for every matrix $\Pi \in \mathcal{A}_m$ there exists an index ℓ such that $|(I-\Pi)^{1/2}\beta_\ell| \geq \mathcal{R}^*(m)$. In particular, $|(I-\widehat{\Pi}_m)^{1/2}\beta_\ell| \geq \mathcal{R}^*(m)$ for some ℓ and hence $|(I-\widehat{\Pi}_m)^{1/2}\hat{\beta}_\ell| \geq |(I-\widehat{\Pi}_m)^{1/2}\beta_\ell| - |\hat{\beta}_\ell - \beta_\ell| \geq \mathcal{R}^*(m) - \varepsilon$. ■

Proposition 15 can be used for estimating the structural dimension m. Indeed, $\hat{\mathcal{R}}(m) \leq \varepsilon$ for $m \geq m^*$ and the results mean that $\hat{\mathcal{R}}(m) \geq (\mathcal{R}^*(m) - \varepsilon)_+$ 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 $m \geq \widehat{m}$.

From now on, we assume that the structural dimension m^* is known and write $\widehat{\Pi}$ instead of $\widehat{\Pi}_{m^*}$.

Proposition 16 *If the vectors β_ℓ satisfy **(A2)** and $\max_\ell |\hat{\beta}_\ell - \beta_\ell| \leq \varepsilon$, then $\operatorname{tr}(I-\widehat{\Pi})\Pi^* \leq 4\varepsilon^2\mu^*$ and $\operatorname{tr}[(\widehat{\Pi}-\Pi^*)^2] \leq 8\varepsilon^2\mu^*$.*

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

$$\operatorname{tr}(\widehat{\Pi}-\Pi^*)^2 = \operatorname{tr}(\widehat{\Pi}^2 - \Pi^*) + 2\operatorname{tr}(I-\widehat{\Pi})\Pi^* \leq 2|\operatorname{tr}(I-\widehat{\Pi})\Pi^*|.$$

Note also that the equality $\operatorname{tr}(I-\widehat{\Pi})\Pi^* = \operatorname{tr}(I-\widehat{\Pi})^{1/2}\Pi^*(I-\widehat{\Pi})^{1/2}$ implies that $\operatorname{tr}(I-\widehat{\Pi})\Pi^* \geq 0$. Now condition (8) and Proposition 15 imply

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

and the assertion follows. ■

Lemma 17 *Let $\operatorname{tr}(I-\widehat{\Pi})\Pi^* \leq \delta^2$ for some $\delta < 1$. Then for any $x \in \mathbb{R}^d$*

$$|\Pi^* x| \leq |\widehat{\Pi}^{1/2}x| + \delta|x|.$$

Proof Denote $\hat{A} = \widehat{\Pi}^{1/2}$. It obviously holds $|\Pi^* x| \leq |\Pi^*\hat{A}x| + |\Pi^*(I-\hat{A})x|$ and

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

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

$$\operatorname{tr}\Pi^*(I-\hat{A})^2\Pi^* \leq \operatorname{tr}\Pi^*(I-\widehat{\Pi})\Pi^* = \operatorname{tr}(I-\widehat{\Pi})\Pi^* \leq \delta^2$$

yielding $|\Pi^* x| \leq |\Pi^*\hat{A}x| + \delta|x| \leq |\hat{A}x| + \delta|x|$ as required. ■

Corollary 18 *Let $\rho \in (0,1)$, and $\hat{S}_\rho = (I + \rho^{-2}\widehat{\Pi})^{1/2}$. If $\mathrm{tr}(I - \widehat{\Pi})\Pi^* \le \delta^2$, then for any $x \in \mathbb{R}^d$, the condition $|\hat{S}_\rho x| \le h$ implies $|\Pi^* x| \le (\rho + \delta)h$.*

Proof The result follows from Lemma 17 and the obvious inequalities $|x| \le |\hat{S}_\rho x| \le h$ and $|\widehat{\Pi}^{1/2} x| \le \rho |\hat{S}_\rho x| \le \rho h$. ■

Lemma 19 *Let $\mathrm{tr}(I - \widehat{\Pi})\Pi^* \le \delta^2$ 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 $\widehat{\Pi}$ 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 $\hat{\vartheta}_j$, $j = 1, \ldots, d$ be respectively the eigenvalues and the eigenvectors of $\widehat{\Pi}$. Assume that $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \ldots \ge \hat{\lambda}_d$. Then $\widehat{\Pi} = \sum_{j=1}^d \hat{\lambda}_j \hat{\vartheta}_j \hat{\vartheta}_j^\top$ 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 $\{\hat{\vartheta}_1, \ldots, \hat{\vartheta}_d\}$ is an orthonormal basis of $\mathbb{R}^d$, Therefore, on the one hand,

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

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

5.6 Technical lemmas

This subsection contains five technical results. The first three lemmas have been used in the proof of Proposition 12, whereas the two last lemmas have been used in the proof of Theorem 3.

Lemma 20 *If $\rho \le 1$, then $\|U - U^*\|_2 \le \alpha$.*

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

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

Since $\|A\|_2^2 = \operatorname{tr} AA^\top \le (\operatorname{tr}(AA^\top)^{1/2})^2$ for any matrix A, it holds

$$\begin{aligned}\left\|\Pi^*(\Pi-\Pi^*)\Pi^*\right\|_2 &= \left\|\Pi^*(I-\Pi)\Pi^*\right\|_2 \\ &\le \operatorname{tr}\,\Pi^*(I-\Pi)\Pi^* = \operatorname{tr}(I-\Pi)\Pi^* \le \delta^2.\end{aligned}$$

By similar arguments one checks that

$$\begin{aligned}\left\|(I-\Pi^*)(\Pi-\Pi^*)(I-\Pi^*)\right\|_2 &= \left\|(I-\Pi^*)\Pi(I-\Pi^*)\right\|_2 \le \operatorname{tr}(I-\Pi^*)\Pi \\ &= \operatorname{tr}\Pi - \operatorname{tr}\Pi^* + \operatorname{tr}\Pi^*(I-\Pi) \\ &\le m^* - m^* + \delta^2, \\ \left\|\Pi^*(\Pi-\Pi^*)(I-\Pi^*)\right\|_2 &\le \left\|\Pi^*(\Pi-\Pi^*)\right\|_2 = \left\|\Pi^*(I-\Pi)\right\|_2 \\ &\le \left\|\Pi^*(I-\Pi)^{1/2}\right\|_2 \le (\operatorname{tr}\Pi^*(I-\Pi)\Pi^*)^{1/2} \\ &= (\operatorname{tr}(I-\Pi)\Pi^*)^{1/2} \le \delta.\end{aligned}$$

Thus we get $\left\|U-U^*\right\|_2 \le \delta^2(1+\rho^{-2}) + 2\delta\rho^{-1}$. The assumption $\rho \le 1$ yields the assertion of the lemma. ■

Lemma 21 *If ψ_ℓs and U satisfy **(A3)** and (13), then*

$$\sum_{j=1}^n |c_{j,\ell}(U)|^2 \le \frac{dC_K C_V \bar{\psi}^2}{h^2 n}.$$

Proof Simple computations yield

$$\sum_{j=1}^n \left|\tilde{V}_i^{-1}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\right|^2 w_{ij} = \operatorname{tr}(\tilde{V}_i^{-1}) \le \frac{dC_V}{N_i}. \tag{17}$$

Hence, we have

$$\begin{aligned}\sum_{j=1}^n |c_{j,\ell}|^2 &= \frac{1}{h^2n^2}\sum_{j=1}^n \left|\sum_{i=1}^n \tilde{V}_i^{-1}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix} w_{ij}\,\psi_\ell(X_i)\right|^2 \\ &\le \frac{\bar{\psi}^2}{h^2n^2}\sum_{j=1}^n\left(\sum_{i=1}^n \frac{w_{ij}}{N_i}\right)\left(\sum_{i=1}^n \left|\tilde{V}_i^{-1}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\right|^2 N_i w_{ij}\right) \\ &\le \frac{C_K\bar{\psi}^2}{h^2n^2}\sum_{j=1}^n\sum_{i=1}^n \left|\tilde{V}_i^{-1}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\right|^2 N_i w_{ij}.\end{aligned}$$

Interchanging the order of summation and using inequality (17) we get the desired result. ■

Lemma 22 *If **(A3)** and (13) are fulfilled, then, for any $j = 1, \ldots, n$,*

$$\sup_U \sup_{\boldsymbol{e}\in S_{d-1}} \left|\frac{d}{dU}(\boldsymbol{e}^\top c_{j,\ell})(U)\right|^2 \le C\left(\frac{C_w^2 C_V^4 C_K^2 \bar{\psi}^2}{n^2h^2} + \frac{C_V^2 C_{K'}^2 \bar{\psi}^2}{n^2h^2}\right),$$

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

Proof We have

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

One checks that $\|d\,w_{ij}(U)/dU\|_2=|w'_{ij}(U)|\cdot|Z_{ij}|^2\le 5|w'_{ij}(U)|$, where we used the notation $w'_{ij}(U)=K'(Z_{ij}^\top UZ_{ij})$ and the inequality

$$\begin{aligned}h^2|Z_{ij}|^2=|S_\rho^*X_{ij}|^2&=|(I-\Pi^*)X_{ij}|^2+2\rho^{-2}|\Pi^*X_{ij}|^2\\ &\le h^2+2(\delta/\rho+1)^2h^2\le 5h^2,\end{aligned}$$

which follows from Lemma 17. We get

$$\Delta_2\le\frac{50\bar\psi^2}{n^2h^2}\left(\sum_{i=1}^n\left|\tilde V_i^{-1}(U)\begin{pmatrix}1\\Z_{ij}\end{pmatrix}w'_{ij}(U)\right|\right)^2\le\frac{C\bar\psi^2C_V^2C_{K'}^2}{n^2h^2}.$$

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

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

Simple computations show that

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

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

$$\frac{da_1^\top\tilde V_i^{-1}a_2}{dU}\ (U)=\sum_{j=1}^n a_1^\top\tilde V_i^{-1}(U)\begin{pmatrix}1\\Z_{ij}\end{pmatrix}\begin{pmatrix}1\\Z_{ij}\end{pmatrix}^\top\tilde V_i^{-1}(U)a_2\ w'_{ij}(U)Z_{ij}Z_{ij}^\top.$$

This relation combined with the estimate $|Z_{ij}|\le 5$ for all i,j such that $w_{ij}\ne 0$, implies the norm estimate

$$\begin{aligned}\left\|\frac{da_1^\top\tilde V_i^{-1}a_2}{dU}\ (U)\right\|_2&\le 25\sum_{j=1}^n\left|a_1^\top\tilde V_i^{-1}(U)\begin{pmatrix}1\\Z_{ij}\end{pmatrix}\begin{pmatrix}1\\Z_{ij}\end{pmatrix}^\top\tilde V_i^{-1}(U)a_2\ w'_{ij}(U)\right|\\ &\le 150|a_1|\,|a_2|\sum_{j=1}^n\left\|\tilde V_i^{-1}(U)\right\|^2|w'_{ij}(U)|\\ &\le 150C_wC_V^2|a_1|\,|a_2|N_i(U)^{-1}.\end{aligned}$$

It provides the following estimate of the term Δ_1:

$$\Delta_1 \leq C\,\frac{C_w^2 C_V^4 C_K^2 \bar{\psi}^2}{n^2 h^2},$$

and the assertion of the lemma follows. ■

Lemma 23 *There exists an integer $n_0 \geq 0$ such that, as soon as $n \geq n_0$, $\delta_{k-1} \leq \rho_k$, $\alpha_k \leq 4$ and $\zeta_k \leq 1/2$ for all $k \in \{2, \ldots, k(n)\}$.*

Proof In view of the inequalities $C_0 n^{-1/(d\vee 4)} = \rho_1 h_1$ and $\rho_{k(n)} h_{k(n)} \geq C_2 n^{-1/3}$, 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 an induction on k. Since $s_n \to 0$ as $n \to \infty$ and $\gamma_1 \leq s_n$, the inequality $\delta_1 = 2\gamma_1\sqrt{\mu^*} \leq 1/\sqrt{2} = \rho_1/\sqrt{2}$ is true for sufficiently large values of n. Let us prove the implication

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

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

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

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

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

Lemma 24 *If $k > 2$ and $\zeta_{k-1} < 1$ then $\Omega_{k-1} \subset \{\mathrm{tr}(I - \widehat{\Pi}_{k-1})\Pi^* \leq \delta_{k-1}^2\}$.*

Proof Let us denote $\tilde{\beta}_\ell = \Pi^* \hat{\beta}_{k-1,\ell}$, then $\tilde{\beta}_\ell \in \mathcal{S}^*$ and under Ω_{k-1} we have

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

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 $\bar{\beta}_i = \sum_\ell c_\ell \beta_\ell$. 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

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

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 $\Pi^* \preceq (1 - \zeta_{k-1})^{-1} \tilde{B}$ and, in view of Proposition 16 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.

References

A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc. Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, 19:563–570, 2003.

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 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, 2001.

E. Bura and R. M. Pfeiffer. Graphical methods for class prediction using dimension reduction techniques on dna microarray data. *Bioinformatics*, 19:1252–1258, 2003.

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 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.

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.

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.