Nonlinear dimension reduction for regression with nearest neighbors

Quentin Paris

Nonlinear dimension reduction for regression with nearest neighbors

Quentin PARIS

IRMAR, ENS Cachan Bretagne, CNRS, UEB
Campus de Ker Lann
Avenue Robert Schuman, 35170 Bruz, France
quentin.paris@bretagne.ens-cachan.fr

Abstract

Let (X,Y) be an $\mathscr{X}\times\mathbb{R}$ valued random variable, where $\mathscr{X}\subset\mathbb{R}^p$. We generalize to a nonlinear framework the sufficient dimension reduction approach, followed by Cadre and Dong (2010), for estimating the regression function $r(x)=\mathbb{E}(Y|X=x)$. We assume given a class $\mathscr{H}$ of functions $h:\mathscr{X}\to\mathbb{R}^p$ such that there exists $h\in\mathscr{H}$ with

$$\mathbb{E}(Y|X)=\mathbb{E}(Y|h(X)).$$

In classical sufficient dimension reduction, $\mathscr{H}$ may be considered as a particular set of matrices. Here, $\mathscr{H}$ is considered to be a general and possibly nonparametric class of functions. In this context, we define the *reduced dimension d* associated with $\mathscr{H}$ as the smallest ℓ such that there exists $h\in\mathscr{H}$ satisfying the former equality and such that $h(\mathscr{X})$ spans a subspace of dimension ℓ. Then, we construct an estimate $\hat{r}$ of r that is proved to achieve the optimal rate of convergence as if the predictor X where d-dimensional.

Index Terms — Dimension reduction, regression estimation, empirical risk minimization, nearest neighbor estimator.

AMS 2000 Classification – 62H12, 62G08.

1 Introduction

In a general setting, regression analysis deals with the problem of retrieving information about the conditional distribution of a real-valued response variable Y given an $\mathscr{X}$-valued predictor X, where $\mathscr{X}\subset\mathbb{R}^p$, and is often understood as a study of the regression function

$$r(x):=\mathbb{E}(Y|X=x). \tag{1.1}$$

It is well known that the estimation of the regression function faces the *curse of dimensionality* which, roughly speaking, means that the expected rate of convergence of a given estimate slows down as the dimension of the predictor X increases. This statement is usually understood in terms of optimal rates of convergence. For instance, if the regression function r is assumed to be Lipschitz, the optimal rate of convergence for the estimation of r is $n^{-2/(2+p)}$. For more details on optimal rates of convergence, the reader is referred to Ibragimov and Khasminskii (1981); Györfi et al. (2002); Kohler et al. (2009) or any other standard textbook on the subject.

To overcome the curse of dimensionality, many authors have considered a model which specifies that the conditional mean of Y given X depends on X only through its projection on an unknown number $d < p$ of unknown orthonormal vectors $\alpha_1, \ldots, \alpha_d \in \mathbb{R}^p$, so that

$$\mathbb{E}(Y|X) = \mathbb{E}\left(Y|\alpha_1'X, \ldots, \alpha_d'X\right) \tag{1.2}$$

(see e.g. Härdle and Stoker, 1989; Li, 1991; Cook, 1998, and the references therein). In this approach, provided d and the α_i's may be estimated, it is naturally expected that the rate of convergence of an estimate of the regression function should depend only on d since the p-dimensional predictor X may be replaced by the d-dimensional predictor $(\alpha_1'X, \ldots, \alpha_d'X)$. Many methods have been introduced in the litterature to estimate d and the α_i's, among which we mention average derivative estimation (ADE) (Härdle and Stoker, 1989), sliced inverse regression (SIR) (Li, 1991), principal Hessian directions (PHD) (Li, 1992), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), kernel dimension reduction (KSIR) (Fukumizu et al., 2009) and more recently the optimal transformation procedure (Delyon and Portier, 2013). Discussions, improvements and other relevant papers can be found in Cook and Li (2002); Fung et al. (2002); Xia et al. (2002); Cook and Ni (2005); Yin et al. (2008) and the references therein. Recently, Cadre and Dong (2010) have used these methods to prove that, in the context of this model, one could construct an estimate of the regression function which converges at the rate $n^{-2/(2+d)}$.

In the present article, we generalize the approach followed by Cadre and Dong (2010) and consider a nonlinear extension of the previous model motivated by the following observation. Assume there exists some function $h : \mathscr{X} \to \mathbb{R}^p$ such that $\dim S(h) < p$, where $S(h)$ stands for the subspace spanned by $h(\mathscr{X})$, and such that

$$\mathbb{E}(Y|X) = \mathbb{E}(Y|h(X)). \tag{1.3}$$

Then, the heuristic approach described in the previous paragraph still applies and it is expected that, provided h can be estimated, a carefully chosen estimate of r

should converge at a rate depending only on $\dim S(h)$. In particular, the fact that h is linear does not appear necessary. The general model that will be considered in this paper is therefore described by the following assumption.

Basic assumption – *We assume given a class $\mathscr{H}$ of functions $h : \mathscr{X} \to \mathbb{R}^p$ such that equation* (1.3) *holds for at least one $h \in \mathscr{H}$.*

Many models may be generated by this general formulation. In particular, when $\mathscr{H}$ is chosen to be the class of all square matrices $(\beta_1 \cdots \beta_s \, 0 \cdots 0)'$ of order p such that $s \leq p$, such that the β_i's belong to $\mathbb{R}^p$ and such that $\beta_i' \beta_j = \delta_{i,j}$, one recovers the model described by (1.2) since the matrix $(\alpha_1 \cdots \alpha_d \, 0 \cdots 0)'$ belongs to $\mathscr{H}$. Other examples may be constructed in the following way. If Φ denotes a class of functions $\phi : \mathscr{X} \to \mathbb{R}$ (such as polynomials), one may consider $\mathscr{H}$ to be the class of all functions $h(x) = (\phi_1(x), \ldots, \phi_s(x), 0, \ldots, 0)$ such that $s \leq p$ and such that the ϕ_i's belong to Φ. The estimation procedure of r presented in this article will be based on the nearest neighbors method and on the existence of a function $h \in \mathscr{H}$ which satisfies (1.3) and which minimizes $\dim S(h)$ among all functions in $\mathscr{H}$ which satisfy (1.3). The value of $\dim S(h)$ for such an optimal $h \in \mathscr{H}$ depends on $\mathscr{H}$, will be referred to as the *reduced dimension* and denoted d. It will be proved that the rate of convergence of a proper estimate of r depends only on the reduced dimension. In the existing literature, similar nonlinear generalizations of model (1.2) have been introduced (see e.g. Cook, 2007) and much effort has already been made for the estimation of an optimal h in the nonlinear context (see e.g. Wu, 2008; Wang and Yin, 2008; Yeh et al., 2009, and the references therein). As far as we know, the estimate of the reduced dimension introduced here is new as well as the method derived to estimate r.

The paper is organized as follows. In Section 2 we give a specific representation of the reduced dimension d. Section 3 is devoted to the study of an estimate of the reduced dimension d based on empirical risk minimization. In Section 4 we construct an estimate $\hat{r}$ of r, based on the nearest neighbors method, which satisfies

$$\mathbb{E}\left(r(X) - \hat{r}(X)\right)^2 = O\left(n^{-2/(2+d)}\right).$$

Section 5 is devoted to some examples. In Section 6 we present a small simulation study. Proofs of the main results are presented in Section 7 and technical results are collected in Section 8.

2 Reduced dimension

According to our basic assumption, there exists a function $h \in \mathscr{H}$ and a function $g : \mathbb{R}^p \to \mathbb{R}$ such that $r = g \circ h$. In addition to the basic assumption, we fix a class $\mathscr{G}$ and use the assumption that such a function g can only belong to $\mathscr{G}$.

Assumption (**A1**) – *For all $h \in \mathscr{H}$ satisfying* (1.3)*, any function $g : \mathbb{R}^p \to \mathbb{R}$ that verifies $r = g \circ h$ belongs to $\mathscr{G}$.*

An explicit choice for the class $\mathscr{G}$ will be specified in Section 3. The consequence of this assumption is that r belongs to the class

$$\mathscr{F} := \{g \circ h : g \in \mathscr{G},\ h \in \mathscr{H}\}.$$

Now, we proceed to giving a more tractable representation of the reduced dimension. For all $\ell \in \{1, \ldots, p\}$, let

$$\mathscr{H}_\ell := \{h \in \mathscr{H} : \dim S(h) \leq \ell\},$$

and define

$$\mathscr{F}_\ell := \{g \circ h : g \in \mathscr{G},\ h \in \mathscr{H}_\ell\}.$$

The $\mathscr{F}_\ell$'s form a nested family of models, that is $\mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \subset \mathscr{F}_p = \mathscr{F}$. The *reduced dimension* is defined by

$$d := \min\{\ell : r \in \mathscr{F}_\ell\}.$$

Denoting μ the distribution of the predictor X, we will use the following assumption.

Assumption (**A2**) – *For all $\ell \in \{1, \ldots, p\}$, the class $\mathscr{F}_\ell$ is compact in $\mathbb{L}^2(\mu)$.*

Now, let R_ℓ be the risk defined by

$$R_\ell := \inf_{f \in \mathscr{F}_\ell} \mathbb{E}(Y - f(X))^2. \tag{2.1}$$

Since the $\mathscr{F}_\ell$'s are nested, the function $\ell \in \{1, \ldots, p\} \mapsto R_\ell$ is nonincreasing. Then, using Assumption (**A2**), we deduce that

$$d = \min\left\{\ell : R_\ell = R_p\right\}. \tag{2.2}$$

Consequently, for all $0 < \delta < \Delta$, we have

$$d = \min\left\{\ell : R_\ell \leq R_p + \delta\right\}, \tag{2.3}$$

where Δ is defined by

$$\Delta := \min\left\{R_\ell - R_p : \ R_\ell > R_p\right\},$$

with the convention $\min\emptyset = +\infty$. Observe that $\Delta > 0$ and that, when $d \geq 2$, Δ corresponds to the distance from r to $\mathscr{F}_{d-1}$ in $\mathbb{L}^2(\mu)$, that is

$$\Delta = \inf_{f \in \mathscr{F}_{d-1}} \|f - r\|_\mu^2. \tag{2.4}$$

Equations (2.2) and (2.4) are proved in Appendix A. For an illustration of equation (2.3), we refer the reader to Figure 1.

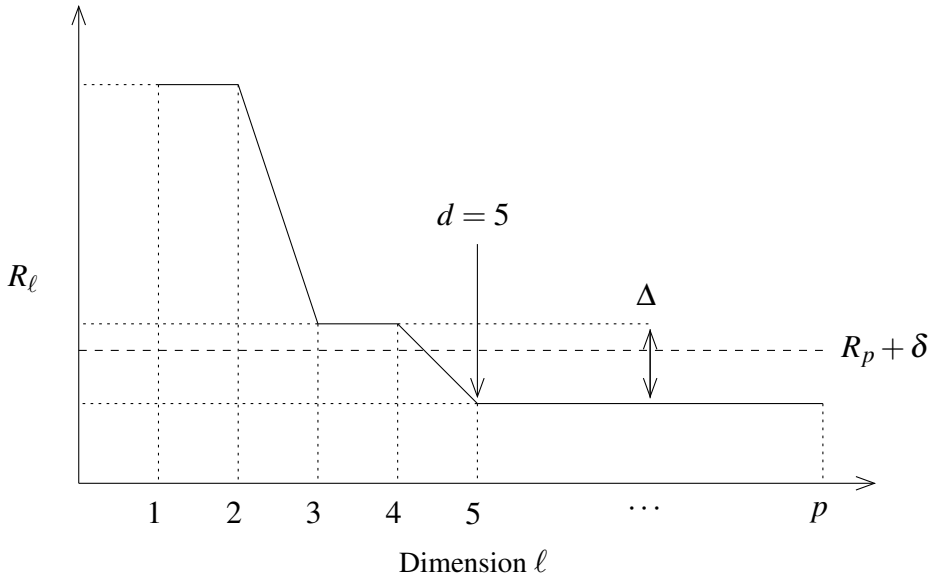

Figure 1: Illustration of Δ and δ. In this example $d = 5$, $0 < \delta < \Delta$ and the smallest ℓ for which $R_\ell \leq R_p + \delta$ is equal to the reduced dimension d as in equation (2.3). Notice that if $\delta \geq \Delta$, then the smallest ℓ for which $R_\ell \leq R_p + \delta$ is at most $d - 1$.

3 Estimation of the reduced dimension

Construction of the estimate

Consider a sample of n i.i.d. random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ with same distribution P as (X, Y) and independent from (X, Y). Let

$$\hat{R}_\ell := \inf_{f \in \mathscr{F}_\ell} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

be the empirical version of the risk R_ℓ given in equation (2.1). Our estimation procedure is inspired by the representation given by equation (2.3). For all $\delta \geq 0$, we define the estimate $\hat{d}(\delta)$ of the reduced dimension d by

$$\hat{d}(\delta) := \min\left\{\ell : \hat{R}_\ell \leq \hat{R}_p + \delta\right\}. \tag{3.1}$$

Notations and assumptions

For $\ell \in \{1,\ldots,p\}$ and $m : \mathscr{X} \to \mathbb{R}^\ell$ we denote by $\|m\|_\infty$ its supremum norm defined by

$$\|m\|_\infty = \sup_{x\in\mathscr{X}} \|m(x)\|,$$

where $\|.\|$ stands for the Euclidean norm in $\mathbb{R}^\ell$. Given a class $\mathscr{M}$ of functions $m : \mathscr{X} \to \mathbb{R}^\ell$, we recall that the ε-covering number $N(\varepsilon, \mathscr{M})$ of $\mathscr{M}$ with respect to $\|.\|_\infty$ is defined as the minimal number of $\|.\|_\infty$-balls of radius ε that are needed to cover $\mathscr{M}$.

Assumption (**A3**) – *Let $R > 0$ be fixed. The class $\mathscr{H}$ is totally bounded with respect to $\|.\|_\infty$ and every $h \in \mathscr{H}$ satisfies $\|h\|_\infty < R$.*

Our next assumption specifies that all functions in $\mathscr{G}$ have minimum regularity. For $g : \mathbb{R}^p \to \mathbb{R}$, we denote

$$\|g\|_{Lip} := \sup_u |g(u)| + \sup_{u\neq u'} \frac{|g(u) - g(u')|}{\|u - u'\|}.$$

Assumption (**A4**) – *Let $L > 0$ be fixed. All functions $g \in \mathscr{G}$ satisfy $\|g\|_{Lip} \leq L$.*

The result

Theorem 3.1 *Suppose that $|Y| \leq B$ and that assumptions* (**A1**) *to* (**A4**) *are satisfied. Then, the following statements hold.*

(*i*) *If $0 < \delta < \Delta$, we have*

$$\mathbb{P}\left(\hat{d}(\delta) < d\right) \leq 4N\left(\tfrac{\Delta-\delta}{12(B+L)}, \mathscr{F}\right)\exp\left(-\tfrac{n(\Delta-\delta)^2}{18(B+L)^4}\right) \text{ and}$$

$$\mathbb{P}\left(\hat{d}(\delta) > d\right) \leq 4N\left(\tfrac{\delta}{12(B+L)}, \mathscr{F}\right)\exp\left(-\tfrac{n\delta^2}{18(B+L)^4}\right).$$

(*ii*) *If $\delta > \Delta$, we have*

$$\mathbb{P}\left(\hat{d}(\delta) \neq d\right) \geq 1 - 4N\left(\tfrac{\delta-\Delta}{12(B+L)}, \mathscr{F}\right)\exp\left(-\tfrac{n(\delta-\Delta)^2}{18(B+L)^4}\right).$$

We can easily deduce from Theorem 3.1 that provided $0 < \delta < \Delta$, we have

$$\hat{d}(\delta) \underset{n\to+\infty}{\longrightarrow} d, \quad \text{a.s.}$$

4 Fast-rate estimation of r

Formal description of the estimate of r

For each function $h \in \mathscr{H}$, we denote r_h the regression function defined for all $u \in \mathbb{R}^p$ by

$$r_h(u) := \mathbb{E}(Y|h(X) = u). \tag{4.1}$$

Under assumption (**A1**) and by definition of the reduced dimension d, there exists $h^* \in \mathscr{H}_d$ such that

$$r = r_{h^*} \circ h^* \quad \text{and} \quad r_{h^*} \in \mathscr{G}. \tag{4.2}$$

The estimation procedure presented here is inspired by this representation of r and consists in two steps. First, for all $h \in \mathscr{H}$, we estimate r_h using the k-nearest neighbors (k-NN) method. In a second step, we estimate h^* through the minimization of an empirical criterion.

Estimate of r_h

Consider a second data set of i.i.d. copies $(X_{n+1}, Y_{n+1}), \ldots, (X_{2n}, Y_{2n})$ of (X, Y) independent from the first data set $(X_1, Y_1), \ldots, (X_n, Y_n)$ introduced in Section 3. Denote

$$\mathscr{D}_1 := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}, \qquad \mathscr{D}_2 := \{(X_{n+1}, Y_{n+1}), \ldots, (X_{2n}, Y_{2n})\},$$

and fix a real number $\delta > 0$. We describe the k-NN procedure in our context (for more information on the k-NN method, we refer the reader to Chapter 6 of the monography by Györfi et al., 2002). For all $h \in \mathscr{H}$ and all $i \in \{n+1, \ldots, 2n\}$, we let

$$X_i^h := h(X_i).$$

If $u \in \mathbb{R}^p$, we reorder the transformed data $\left(X_{n+1}^h, Y_{n+1}\right), \ldots, \left(X_{2n}^h, Y_{2n}\right)$ according to increasing values of $\left\{\|X_i^h - u\|, i = n+1, \ldots, 2n\right\}$. The reordered data sequence is denoted

$$\left(X_{(1)}^h(u), Y_{(1)}^h(u)\right), \left(X_{(2)}^h(u), Y_{(2)}^h(u)\right), \ldots, \left(X_{(n)}^h(u), Y_{(n)}^h(u)\right),$$

which means that

$$\|X_{(1)}^h(u) - u\| \leq \|X_{(2)}^h(u) - u\| \leq \cdots \leq \|X_{(n)}^h(u) - u\|.$$

In this approach, $X_{(i)}^h(u)$ is called the i-th NN of u. Note that if X_i^h and X_j^h are equidistant from u, i.e. $\|X_i^h - u\| = \|X_j^h - u\|$, then we have a tie. As usual, we

then declare X_i^h closer to u than X_j^h if $i < j$. For any $i \in \{n+1,\dots,2n\}$ and $k \in \{1,\dots,n\}$, we define

$$W_i[k](h,u) = \begin{cases} 1/k & \text{if } X_i^h \text{ is among the } k\text{-NN of } u \text{ in } \left\{X_{n+1}^h,\dots,X_{2n}^h\right\}; \\ 0 & \text{elsewhere.} \end{cases}$$

Observe that we have $\sum_{i=n+1}^{2n} W_i[k](h,u) = 1$. We define the estimate $\hat{r}_h[k]$ of r_h for all $u \in \mathbb{R}^p$ by

$$\hat{r}_h[k](u) = \sum_{i=n+1}^{2n} W_i[k](h,u) Y_i = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}^h(u).$$

Estimate of h^*

Now we focus on the estimation of h^* through minimization of an empirical criterion over a finite covering of $\mathscr{H}$. To this aim, for $\rho > 0$ and $\ell \in \{1,\dots,p\}$, let $\mathbf{H}_\ell(\rho)$ be a ρ-covering of $\mathscr{H}_\ell$ of minimum cardinality (recall that by Assumption (**A3**), the class $\mathscr{H}$ is totally bounded). We set $\mathbf{H}(\rho) := \cup_\ell \mathbf{H}_\ell(\rho)$ and we denote

$$\mathbf{N}(\rho) := |\mathbf{H}(\rho)|.$$

Now define

$$\hat{h}_\ell[k,\rho] := \underset{h \in \mathbf{H}_\ell(\rho)}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{r}_h[k](h(X_i))\right)^2.$$

A natural estimate of h^* defined by (4.2) is then

$$\hat{h}[k,\rho] := \hat{h}_{\hat{d}(\delta)}[k,\rho],$$

where $\hat{d}(\delta)$ is the estimate of the reduced dimension introduced by equation (3.1).

Estimate of r

Following equation (4.2), a natural estimate of r is given by

$$\hat{r}[k,\rho] := \hat{r}_{\hat{h}[k,\rho]}[k] \circ \hat{h}[k,\rho].$$

Next, we describe a data driven choice of the number k of neighbors and of the radius ρ of the covering $\mathbf{H}(\rho)$. For all $\ell \in \{1,\dots,p\}$, denote

$$\upsilon_n(\ell) := n^{-2/(2+\ell)}, \tag{4.3}$$

and let

$$\hat{k} := \left\lfloor \upsilon_n^{-1}\left(\hat{d}(\delta)\right) \right\rfloor, \qquad \hat{\rho} := \sqrt{\upsilon_n\left(\hat{d}(\delta)\right)}, \tag{4.4}$$

where $\lfloor x \rfloor$ stands for the greatest integer smaller than x. Finally, for all $x \in \mathscr{X}$, let

$$\hat{r}(x) := \hat{r}\left[\hat{k}, \hat{\rho}\right](x). \tag{4.5}$$

Results

We will use the following assumption which roughly means that r_h is close to r_{h^*} provided h is close to h^*.

Assumption (**A5**) – *Let $K > 0$ be fixed. For all $h \in \mathscr{H}$ and all $u \in \mathbb{R}^p$ we have*

$$|r_h(u) - r_{h^*}(u)| \leq K\|h - h^*\|_\infty.$$

Theorem 4.1 *Suppose that $|Y| \leq B$, that assumptions* (**A1**) *to* (**A5**) *are satisfied and that $d \geq 3$. For all $0 < \delta < \Delta$ there exists a constant $C > 0$ such that for all $n \geq 1$ we have*

$$\mathbb{E}\left(r(X) - \hat{r}(X)\right)^2 \leq Cn^{-2/(2+d)} + C\mathbf{R}_n,$$

where

$$\mathbf{R}_n := n^{-d/(2+d)} N\left(n^{-1/(2+d)}, \mathscr{H}\right) \exp\left(-\tfrac{n^{(d-2)/(2+d)}}{C}\right).$$

Remark 4.2 *When $d \leq 2$, under the additional conditions of Problem* 6.7 *in the book by Györfi et al. (2002), a slight adaptation of the proof of Theorem 4.1 enables us to derive the same convergence rate.*

In Theorem 4.1, the quadratic risk of the estimator $\hat{r}$ is bounded by a sum of two terms. The first term

$$Cn^{-2/(2+d)},$$

goes to 0 at the optimal rate of convergence associated with the class of Lipschitz functions $\mathbb{R}^d \to \mathbb{R}$. The second term

$$C\mathbf{R}_n,$$

is induced by our dimension reduction procedure. Therefore, the rate of convergence of $\hat{r}$ depends on that of term $\mathbf{R}_n$. The next assumption consists in a restriction on the complexity of $\mathscr{H}$ and is known to be satisfied by many examples as described in Section 5.

Assumption (**A6**) – *There exists $A > 0$ and $0 < s < d - 2$ such that for all $\varepsilon > 0$*

$$\log N(\varepsilon, \mathscr{H}) \leq A\varepsilon^{-s}.$$

Under this assumption, we have

$$\mathbf{R}_n \leq n^{-d/(2+d)} \exp\left(An^{s/(2+d)} - \tfrac{n^{(d-2)/(2+d)}}{C}\right) = O\left(n^{-2/(2+d)}\right),$$

which leads to the following result.

Corollary 4.3 *Under the assumptions of Theorem 4.1 and under assumption* (**A6**)*, we have*

$$\mathbb{E}\left(r(X)-\hat{r}(X)\right)^2 = O\left(n^{-2/(2+d)}\right).$$

In other words, our estimate reaches the optimal rate of convergence, should X be taking its values in $\mathbb{R}^d$.

5 Examples

In this section, we study two examples and illustrate our main assumptions in different settings. The linear case is studied first. The second example studies a case where the class $\mathscr{F}$ is nonparametric.

Linear dimension reduction

Linear dimension reduction techniques have proven effective in a large class of examples and practical situations (see e.g. Härdle and Stoker, 1989; Li, 1991; Cadre and Dong, 2010; Györfi et al., 2002). We illustrate how this topic fits into our framework.

Assume $\mathscr{X}$ is the open Euclidean ball of center the origin and radius R in $\mathbb{R}^p$. For any subspace V of $\mathbb{R}^p$, denote by $\pi_V : \mathbb{R}^p \to \mathbb{R}^p$ the orthogonal projector onto V and $P_V : \mathscr{X} \to \mathbb{R}^p$ the restriction of π_V to $\mathscr{X}$. Now consider the following model. Let $d < p$, let V_0 be a subspace of $\mathbb{R}^p$ of dimension d and suppose that for $x \in \mathscr{X}$ we have

$$r(x) := g(P_{V_0}x),\ x \in \mathscr{X}.$$

Note that this model is described by equation (1.2) when V_0 is the subspace spanned by the vectors $\alpha_1, \ldots, \alpha_d$. We assume furthermore that g is L-Lipschitz for some $L > 0$. Then, denote $\mathscr{H} := \left\{P_V : V \text{ subspace of } \mathbb{R}^p\right\}$ and let $\mathscr{G}$ be the set of all L-Lipschitz functions from $\mathbb{R}^p$ to $\mathbb{R}$. In this context, Assumption (**A4**) is satisfied. The class $\mathscr{H}$ satisfies assumption (**A3**) since for all $V \subset \mathbb{R}^p$: $\|P_V\|_\infty < R$. Therefore, $\mathscr{H}$ may be seen as a subset of the open Euclidean ball with center 0 and radius R in $\mathbb{R}^{p^2}$ and it follows that

$$\log N(\varepsilon, \mathscr{H}) \leq C \log\left(\tfrac{1}{\varepsilon}\right),$$

for a constant C depending only on p and R (see e.g. Proposition 5 in Cucker and Smale, 2001). Hence, Assumption (**A6**) is also satisfied in this case. Therefore, provided Assumptions (**A2**) and (**A5**) are also satisfied, Corollary 4.3 implies that

$$\mathbb{E}\left(r(X)-\hat{r}(X)\right)^2 = O\left(n^{-2/(2+d)}\right).$$

Smooth functions

In this example we show that one may consider classes $\mathscr{F}$ much wider than the class of projectors or even a more general parametric class. The class $\mathscr{H}$ introduced here consists in a nonparametric class of smooth functions. Fix two constants $R, \alpha > 0$. Denote by $\lfloor \alpha \rfloor$ the greatest integer strictly smaller than α and $\mathscr{C}^{\lfloor \alpha \rfloor}(\mathscr{X})$ the space of $\lfloor \alpha \rfloor$-times continuously differentiable functions $\phi : \mathscr{X} \to \mathbb{R}$. For all $\phi \in \mathscr{C}^{\lfloor \alpha \rfloor}(\mathscr{X})$ we define

$$\|\phi\|_\alpha := \max_{|s| \leq \lfloor \alpha \rfloor} \sup_x \|\partial^s \phi(x)\| + \max_{|s| = \lfloor \alpha \rfloor} \sup_{x \neq y} \frac{\|\partial^s \phi(x) - \partial^s \phi(y)\|}{\|x - y\|^{\alpha - \lfloor \alpha \rfloor}},$$

where, for all multi-index $s = (s_1, \ldots, s_p) \in \mathbb{N}^p$, we have denoted $|s| := \sum_i s_i$ and $\partial^s := \partial_1^{s_1} \cdots \partial_p^{s_p}$. Now we may define

$$\mathscr{H} := \left\{ \phi \in \mathscr{C}^{\lfloor \alpha \rfloor}(\mathscr{X}) : \|\phi\|_\alpha \leq R \right\}^{\otimes p},$$

where for any set Φ of functions from $\mathscr{X}$ to $\mathbb{R}$, $\Phi^{\otimes p}$ stands for the set of functions from $\mathscr{X}$ to $\mathbb{R}^p$ such that each coordinate function belongs to Φ. By definition of $\|.\|_\alpha$, assumption (**A3**) is satisfied.

Now consider that there exists $h \in \mathscr{H}$ such that $r = g \circ h$ for some L-Lipschitz function and let d be the reduced dimension associated to this choice of class $\mathscr{H}$. Then, provided $\mathscr{X}$ is compact, there exists a constant C depending only on α, p and the diameter of $\mathscr{X}$ such that for all $\varepsilon > 0$

$$\log N(\varepsilon, \mathscr{H}) \leq C \varepsilon^{-p/\alpha}$$

(see e.g. Theorem 2.7.1 in van der Vaart and Wellner, 1996). Hence, provided $\alpha > p/(d-2)$, assumption (**A6**) is satisfied. Therefore, provided Assumptions (**A2**) and (**A5**) are also satisfied, Corollary 4.3 implies that

$$\mathbb{E}\left(r(X) - \hat{r}(X)\right)^2 = O\left(n^{-2/(2+d)}\right).$$

6 A small simulation study

Here, we illustrate the improvement that our dimension reduction step induces compared to the classical NN approach, for a simple model. We let $p = 4$ and generate our data as follows. We let $X = (X^{(1)}, \ldots, X^{(4)})$ be a 4-dimensional vector uniformly distributed over the unit open Euclidean ball B in $\mathbb{R}^4$ and let

$$Y = X^{(1)} X^{(4)} + X^{(2)} X^{(3)} + \sigma \varepsilon.$$

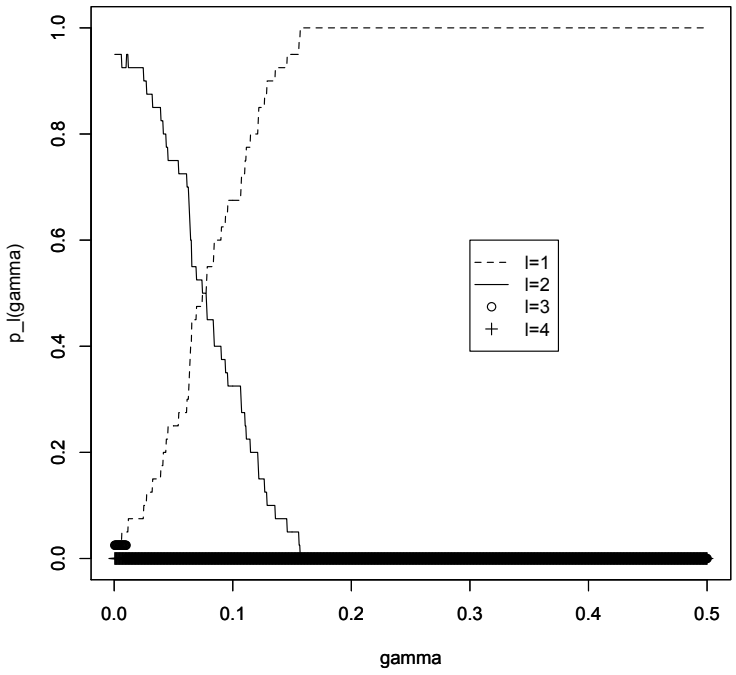

Figure 2: Plot of $p_\ell(\gamma)$ as a function of $\gamma \in (0, 0.5]$ for $\ell = 1,2,3,4$.

In this setting, $\sigma > 0$ and ε is a real random variable independent from X with standard normal distribution.

In this example, the class $\mathscr{G}$ is the class of functions g of the form $g(u_1,\ldots,u_4) = a_1u_1 + \cdots + a_4u_4$ where $a_i \in \{-10,-9.9,\ldots,9.9,10\}$. The class $\mathscr{H}$ is the class of all functions $h : x \in B \mapsto (\phi_1(x),\ldots,\phi_4(x))$ where the ϕ_i's belong to $\Phi := \{(x_1,\ldots,x_4) \in \mathbb{R}^4 \mapsto x_k x_\ell \in \mathbb{R} : \ k,\ell \in \{1,\ldots,4\}\}$. In this context, the reduced dimension is $d = 2$.

First, we study the performance of the estimate $\hat{d}(\delta)$ of d defined in (3.1) and describe an empirical procedure to select δ. We take a data set $\mathscr{D}_1 := \{(x_i, y_i) : i = 1,\ldots,1600\}$ generated by our model. We divide $\mathscr{D}_1$ into 40 sets of 40 data points and for $j = 1,\ldots,40$ we denote

$$\mathscr{S}_j := \{(x_i, y_i) : i = (j-1)n+1,\ldots,(j-1)n+40\}.$$

For all $j = 1,\ldots,40$ and all $\gamma \in \{\kappa/2000 : \kappa = 0,1,\ldots,1000\}$ we compute the estimate $\hat{d}_j(\gamma)$ of d based on the subsample $\mathscr{S}_j$ as in (3.1). Then, for $\ell = 1,2,3,4$

we plot in Figure 2 the proportion of $\hat{d}_j(\gamma)$ that are equal to ℓ, *i.e.*

$$p_\ell(\gamma) := \frac{1}{40}\sum_{j=1}^{40} \mathbf{1}\left\{\hat{d}_j(\gamma) = \ell\right\},$$

when γ ranges over $\{\kappa/2000 : \kappa = 0, 1, \ldots, 1000\}$. We select δ using the following heuristic approach. For small values of $\gamma > 0$ and when the number of data points is large enough, the probability that $\hat{d}(\gamma) = \ell$ should be close to 1 for $\ell = d$ and close to 0 for $\ell \neq d$ according to Theorem 3.1. Therefore, the value of ℓ for which $p_\ell(\gamma)$ is close to 1 for the small values of γ should correspond to the reduced dimension. Then, for this fixed value of ℓ, we select δ as the smallest maximizer γ of $p_\ell(\gamma)$. In our example, this heuristic applies successfully as we can see in Figure 2 that $p_2(\gamma) > 0.9$ for $0 < \gamma < 0.04$. Here we select $\delta = 0.01$.

For the estimation of the regression function we consider two additional independent data sets $\mathscr{D}_2 := \{(x_i, y_i) : i = 1601, ..., 3200\}$ and $\mathscr{D}_{MC} := \{(x_i, y_i) : i = 3201, ..., 4800\}$ generated by our model for $\sigma = 0.1$, 0.5 and 0.9. For each $i = 3201, ..., 4800$ we compute our estimates $\hat{r}(x_i)$ based on the subsample $\mathscr{D} := \mathscr{D}_1 \cup \mathscr{D}_2$ with our methods as in (4.5) with $\delta = 0.01$ and with the classical NN-method. Finally, we estimate $\mathbb{E}(\hat{r}(X) - \mathbb{E}(Y|X))^2$ from the Monte-Carlo approximation

$$\frac{1}{1600}\sum_{i=3201}^{4800} (\hat{r}(x_i) - \mathbb{E}(Y|X = x_i))^2 = \frac{1}{1600}\sum_{i=3201}^{4800} \left(\hat{r}(x_i) - x_i^{(1)}x_i^{(4)} - x_i^{(2)}x_i^{(3)}\right)^2,$$

where $x_i^{(j)}$ denotes the j-th coordinate of x_i. The obtained results are given in Table 1 and the variance of each experiment is given in parenthesis. As expected from our main theorem, our method performs better than the classical 4-dimensional NN method.

	Our meth.	4-dim NN-meth.
$\sigma = 0.1$	$0.027(8.4e-05)$	$0.039(1.2e-04)$
$\sigma = 0.5$	$0.216(7.3e-03)$	$0.336(8.4e-03)$
$\sigma = 0.9$	$0.720(4.7e-02)$	$0.898(4.9e-02)$

Table 1: Estimated mean squared error. The variance of the MC approximation is given in parenthesis.

7 Proofs

7.1 Proof of Theorem 3.1

Proof of (i) – Since the function $\ell \mapsto \hat{R}_\ell$ is non increasing, one has for all integer $q \in \{1,\dots,p\}$ and every $\delta \geq 0$

$$\min\left\{\ell = 1,\dots,p : \hat{R}_\ell - \hat{R}_p \leq \delta\right\} \leq q \quad \Leftrightarrow \quad \hat{R}_q - \hat{R}_p \leq \delta.$$

Now assume $0 < \delta < \Delta$. Using $R_{d-1} - R_p = \Delta$ and the equivalence above, we have

$$\begin{aligned}
\mathbb{P}\left(\hat{d}(\delta) < d\right) &= \mathbb{P}\left(\min\left\{\ell = 1,\dots,p : \hat{R}_\ell - \hat{R}_p \leq \delta\right\} \leq d-1\right) \\
&= \mathbb{P}\left(\hat{R}_{d-1} - \hat{R}_p \leq \delta\right) \\
&= \mathbb{P}\left(\left(\hat{R}_{d-1} - R_{d-1}\right) + \Delta + \left(R_p - \hat{R}_p\right) \leq \delta\right) \\
&\leq \mathbb{P}\left(|\hat{R}_{d-1} - R_{d-1}| \geq \tfrac{\Delta-\delta}{2}\right) + \mathbb{P}\left(|\hat{R}_p - R_p| \geq \tfrac{\Delta-\delta}{2}\right) \\
&\leq 4N\left(\tfrac{\Delta-\delta}{12(B+L)}, \mathscr{F}\right) \exp\left(-\tfrac{n(\Delta-\delta)^2}{18(B+L)^4}\right),
\end{aligned}$$

where the last inequality follows from Lemma 8.1. Next, using $R_d = R_p$, we obtain similarly that

$$\begin{aligned}
\mathbb{P}\left(\hat{d}(\delta) > d\right) &= \mathbb{P}\left(\min\left\{\ell = 1,\dots,p : \hat{R}_\ell - \hat{R}_p \leq \delta\right\} > d\right) \\
&= \mathbb{P}\left(\hat{R}_d - \hat{R}_p > \delta\right) \\
&= \mathbb{P}\left(\left(\hat{R}_d - R_d\right) + \left(R_p - \hat{R}_p\right) > \delta\right) \\
&\leq \mathbb{P}\left(|\hat{R}_p - R_p| \geq \tfrac{\delta}{2}\right) + \mathbb{P}\left(|\hat{R}_d - R_d| \geq \tfrac{\delta}{2}\right) \\
&\leq 4N\left(\tfrac{\delta}{12(B+L)}, \mathscr{F}\right) \exp\left(-\tfrac{n\delta^2}{18(B+L)^4}\right).
\end{aligned}$$

Proof of (ii) – Now assume $\delta > \Delta$. We have

$$\begin{aligned}
\mathbb{P}\left(\hat{d}(\delta) \neq d\right) &\geq \mathbb{P}\left(\hat{d}(\delta) < d\right) \\
&= \mathbb{P}\left(\min\left\{\ell = 1,\dots,p : \hat{R}_\ell - \hat{R}_p \leq \delta\right\} \leq d-1\right) \\
&= \mathbb{P}\left(\hat{R}_{d-1} - \hat{R}_p \leq \delta\right) \\
&= \mathbb{P}\left(\left(\hat{R}_{d-1} - R_{d-1}\right) + \Delta + \left(R_p - \hat{R}_p\right) \leq \delta\right) \\
&= 1 - \mathbb{P}\left(\left(\hat{R}_{d-1} - R_{d-1}\right) + \left(R_p - \hat{R}_p\right) > \delta - \Delta\right) \\
&\geq 1 - \mathbb{P}\left(|\hat{R}_{d-1} - R_{d-1}| \geq \tfrac{\delta-\Delta}{2}\right) - \mathbb{P}\left(|\hat{R}_p - R_p| \geq \tfrac{\delta-\Delta}{2}\right) \\
&\geq 1 - 4N\left(\tfrac{\delta-\Delta}{12(B+L)}, \mathscr{F}\right) \exp\left(-\tfrac{n(\delta-\Delta)^2}{18(B+L)^4}\right),
\end{aligned}$$

by Lemma 8.1 once again.□

7.2 Proof of Theorem 4.1

First, we give an intermediate result.

Proposition 7.1 *Suppose that $|Y| \leq B$, that assumptions* (**A1**) *to* (**A5**) *are satisfied and that $d \geq 3$. Then for all $0 < \delta < \Delta$, for all $k \in \{1,\ldots,n\}$ and for all $\rho > 0$ we have*

$$\mathbb{E}\left(r(X) - \hat{r}[k,\rho](X)\right)^2 \leq C_1 \left\{ \frac{1}{k} + \left(\frac{k}{n}\right)^{2/d} \right\} + C_1 \left\{ \mathbf{D}_n(\delta) + \mathbf{T}_n(\rho) + \frac{1}{\sqrt{n}} \right\},$$

where

$$\begin{aligned} \mathbf{D}_n(\delta) &:= N\left(\tfrac{\Delta-\delta}{12(B+L)}, \mathscr{F}\right) \exp\left(-\tfrac{n(\Delta-\delta)^2}{18(B+L)^4}\right), \\ \mathbf{T}_n(\rho) &:= \rho^2 + \upsilon_n(d) + \tfrac{p}{n\upsilon_n(d)} N(\rho, \mathscr{H}) \exp\left(-\tfrac{n\upsilon_n^2(d)}{C_1}\right), \end{aligned}$$

and where

$$C_1 := \max\left\{1;\ 8B^4;\ 8B^2\sqrt{2\pi};\ 4(B+L)^2;\ 10(K^2+L^2)\ ;\ 96L^2R^2 4^{1/d}\left(\tfrac{d-2}{2}\right)^{4/d}\right\}.$$

Proof of Proposition 7.1 – Fix $k \in \{1,\ldots,n\}$, $\rho > 0$ and a function $h \in \mathbf{H}_d(\rho)$ such that $\|h - h^*\|_\infty \leq \rho$. We have

$$\begin{aligned} \mathbb{E}\left(r(X) - \hat{r}[k,\rho](X)\right)^2 &= \mathbb{E}\Big[\left(r(X) - \hat{r}[k,\rho](X)\right)^2 \mathbf{1}\left\{\hat{d}(\delta) < d\right\} \Big] \\ &\qquad + \mathbb{E}\Big[\left(r(X) - \hat{r}[k,\rho](X)\right)^2 \mathbf{1}\left\{\hat{d}(\delta) \geq d\right\} \Big] \\ &=: T_1 + T_2. \end{aligned}$$

By assumption, we have $|Y| \leq B$. Hence, from the construction of $\hat{r}[k,\rho]$, we have $|\hat{r}[k,\rho](X)| \leq B$, and so

$$\left(r(X) - \hat{r}[k,\rho](X)\right)^2 \leq (B+L)^2.$$

Therefore

$$T_1 \leq (B+L)^2 \mathbb{P}\Big(\hat{d}(\delta) < d\Big).$$

Using Theorem 3.1, we deduce that

$$T_1 \leq 4(B+L)^2 N\left(\tfrac{\Delta-\delta}{12(B+L)}, \mathscr{F}\right) \exp\left(-\tfrac{n(\Delta-\delta)^2}{18(B+L)^4}\right).$$

Next, we have

$$\begin{aligned} T_2 &= \mathbb{E}\Big[(r(X) - \hat{r}[k,\rho](X))^2 \mathbf{1}\{\hat{d}(\delta) \geq d\} \Big] \\ &= \mathbb{E}\Big[\mathbb{E}\Big[(r(X) - \hat{r}[k,\rho](X))^2 \,|\mathscr{D}\Big] \mathbf{1}\{\hat{d}(\delta) \geq d\} \Big] \\ &= \mathbb{E}\Big[\Big\{ \mathbb{E}\Big[(Y - \hat{r}[k,\rho](X))^2 \,|\mathscr{D}\Big] - \mathbb{E}\Big[(Y - r(X))^2 \Big] \Big\} \mathbf{1}\{\hat{d}(\delta) \geq d\} \Big] \\ &=: \mathbb{E}\Big[(I_1 + I_2 + I_3) \mathbf{1}\{\hat{d}(\delta) \geq d\} \Big], \end{aligned}$$

where I_1, I_2 and I_3 are defined by

$$\begin{aligned} I_1 &= \left\{ \mathbb{E}\left[(Y - \hat{r}[k,\rho](X))^2 \,|\mathscr{D}\right] - \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{r}[k,\rho](X_i))^2 \right\}, \\ I_2 &= \left\{ \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{r}[k,\rho](X_i))^2 - \mathbb{E}\left[(Y - \hat{r}_h[k](h(X)))^2 \,|\mathscr{D}\right] \right\}, \\ I_3 &= \left\{ \mathbb{E}\left[(Y - \hat{r}_h[k](h(X)))^2 \,|\mathscr{D}\right] - \mathbb{E}\,(Y - r(X))^2 \right\}. \end{aligned}$$

By taking $a = \upsilon_n(d)$ in Lemma 8.2, we obtain

$$\mathbb{E}\Big[I_1 \mathbf{1}\{\hat{d}(\delta) \geq d\} \Big] \;\leq\; \mathbb{E}\Big[\big| I_1 \big| \Big] \leq\; C' \left\{ \upsilon_n(d) + \frac{\mathbf{N}(\rho)}{n\upsilon_n(d)} \exp\left(-\tfrac{n\upsilon_n^2(d)}{C'} \right) \right\},$$

where we recall that $\upsilon_n(d)$ has been defined in equation (4.3) and where $C' = \max\{1;\, 8B^4\}$. From the construction of $\mathbf{H}(\rho)$, we deduce that

$$\mathbb{E}\Big[I_1 \mathbf{1}\{\hat{d}(\delta) \geq d\} \Big] \leq C' \left\{ \upsilon_n(d) + \frac{pN(\rho, \mathscr{H})}{n\upsilon_n(d)} \exp\left(-\tfrac{n\upsilon_n^2(d)}{C'} \right) \right\}. \tag{7.1}$$

Now on the event $\{\hat{d}(\delta) \geq d\}$, we have the inclusion $\mathbf{H}_d(\rho) \subset \mathbf{H}_{\hat{d}(\delta)}(\rho)$, and therefore

$$\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{r}[k,\rho](X_i))^2 \leq \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{r}_h[k](h(X_i)))^2 .$$

We conclude that on the event $\{\hat{d}(\delta) \geq d\}$ we have

$$\begin{aligned} I_2 \;\leq\; J_2 \;&:=\; \left\{ \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{r}_h[k](h(X_i)))^2 - \mathbb{E}\left[(Y - \hat{r}_h[k](h(X)))^2 \,|\mathscr{D}\right] \right\} \\ &=\; \left\{ \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{r}_h[k](h(X_i)))^2 - \mathbb{E}\left[(Y - \hat{r}_h[k](h(X)))^2 \,|\mathscr{D}_2\right] \right\}, \end{aligned}$$

where in the last inequality we have used the fact that for all $h \in \mathscr{H}$, $\hat{r}_h[k]$ is constructed on $\mathscr{D}_2 = \{(X_i, Y_i); i = n+1, \ldots, 2n\}$. Conditionally to $\mathscr{D}_2$, the variables

$$(Y - \hat{r}_h[k](h(X)))^2 \quad \text{and} \quad (Y_i - \hat{r}_h[k](h(X_i)))^2,\ i = 1, \ldots, n,$$

are i.i.d. and bounded by $4B^2$. Hence, for all $\varepsilon > 0$, Hoeffding's inequality yields

$$\mathbb{P}\Big(|J_2| > \varepsilon \big| \mathscr{D}_2\Big) \leq 2\exp\left(-\tfrac{n\varepsilon^2}{8B^4}\right).$$

Therefore

$$\begin{aligned}
\mathbb{E}\Big[I_2 \mathbf{1}\left\{\hat{d}(\delta) \geq d\right\}\Big] &\leq \mathbb{E}\Big[|J_2| \mathbf{1}\left\{\hat{d}(\delta) \geq d\right\}\Big] \\
&\leq \mathbb{E}\Big[\big|J_2\big|\Big] \\
&= \int_0^{+\infty} \mathbb{P}\Big(|J_2| > \varepsilon\Big) \mathrm{d}\varepsilon \\
&= \int_0^{+\infty} \mathbb{E}\left[\mathbb{P}\Big(|J_2| > \varepsilon | \mathscr{D}_2\Big)\right] \mathrm{d}\varepsilon \\
&\leq 2\int_0^{+\infty} \exp\left(-\tfrac{n\varepsilon^2}{8B^4}\right) \mathrm{d}\varepsilon \\
&= \frac{4B^2}{\sqrt{n}} \int_0^{+\infty} \exp\left(-\tfrac{\tau^2}{2}\right) \mathrm{d}\tau \\
&= \frac{2B^2\sqrt{2\pi}}{\sqrt{n}} \\
&\leq \frac{C''}{\sqrt{n}}, \qquad (7.2)
\end{aligned}$$

where $C'' := 8B^2\sqrt{2\pi}$. Finally, we have

$$\begin{aligned}
\mathbb{E}\Big[I_3 \mathbf{1}\left\{\hat{d}(\delta) \geq d\right\}\Big] &= \mathbb{E}\Big[\mathbb{E}\left[(r(X) - \hat{r}_h[k](h(X)))^2 | \mathscr{D}\right] \mathbf{1}\left\{\hat{d}(\delta) \geq d\right\}\Big] \\
&\leq \mathbb{E}\Big[(r(X) - \hat{r}_h[k](h(X)))^2\Big] \\
&\leq C''' \left\{\frac{1}{k} + \left(\frac{k}{n}\right)^{2/d} + \rho^2\right\}, \qquad (7.3)
\end{aligned}$$

where the last inequality follows from Lemma 8.3 an where constant C''' can be taken equal to $\max\Big\{2(B+L)^2;\ 10(K^2+L^2)\ ;\ 96L^2R^2 4^{1/d}\left(\frac{d-2}{2}\right)^{4/d}\Big\}$. Combining (7.1), (7.2) and (7.3) and denoting

$$C_1 := \max\Big\{1;\ 8B^4;\ 8B^2\sqrt{2\pi};\ 4(B+L)^2;\ 10(K^2+L^2)\ ;\ 96L^2R^2 4^{1/d}\left(\tfrac{d-2}{2}\right)^{4/d}\Big\},$$

we conclude that

$$T_2 \leq C_1 \left\{ \frac{1}{k} + \left(\frac{k}{n}\right)^{2/d} \right\} + C_1 \left\{ \mathbf{T}_n(\rho) + \frac{1}{\sqrt{n}} \right\},$$

where

$$\mathbf{T}_n(\rho) := \rho^2 + \upsilon_n(d) + \frac{pN(\rho, \mathscr{H})}{n\upsilon_n(d)} \exp\left(-\tfrac{n\upsilon_n^2(d)}{C_1}\right).$$

As a result, we obtain

$$\begin{aligned}
\mathbb{E}\Big[(r(X) - \hat{r}[k,\rho](X))^2 \Big] &\leq T_1 + T_2 \\
&\leq C_1 \left\{ \frac{1}{k} + \left(\frac{k}{n}\right)^{2/d} \right\} + C_1 \left\{ \mathbf{D}_n(\delta) + \mathbf{T}_n(\rho) + \frac{1}{\sqrt{n}} \right\},
\end{aligned}$$

where

$$\mathbf{D}_n(\delta) := N\left(\tfrac{\Delta-\delta}{12(B+L)}, \mathscr{F}\right) \exp\left(-\tfrac{n(\Delta-\delta)^2}{18(B+L)^4}\right),$$

as desired. □

Proof of Theorem 4.1 – By assumption we have $|Y| \leq B$. Hence, from the construction of $\hat{r}$, we have $|\hat{r}(X)| \leq B$, and so

$$(r(X) - \hat{r}(X))^2 \leq (B+L)^2.$$

Therefore

$$\begin{aligned}
\mathbb{E}\Big[(r(X) - \hat{r}(X))^2 \Big] &= \mathbb{E}\Big[(r(X) - \hat{r}(X))^2 \mathbf{1}\left\{\hat{d}(\delta) \neq d\right\} \Big] \\
&\quad + \mathbb{E}\Big[(r(X) - \hat{r}(X))^2 \mathbf{1}\left\{\hat{d}(\delta) = d\right\} \Big] \\
&= \mathbb{E}\Big[(r(X) - \hat{r}(X))^2 \mathbf{1}\left\{\hat{d}(\delta) \neq d\right\} \Big] \\
&\quad + \mathbb{E}\Big[\Big(r(X) - \hat{r}\left[\lfloor \upsilon_n^{-1}(d) \rfloor, \upsilon_n^{1/2}(d) \right](X) \Big)^2 \mathbf{1}\left\{\hat{d}(\delta) = d\right\} \Big] \\
&\leq (B+L)^2 \mathbb{P}\Big(\hat{d}(\delta) \neq d\Big) \\
&\quad + \mathbb{E}\Big[\Big(r(X) - \hat{r}\left[\lfloor \upsilon_n^{-1}(d) \rfloor, \upsilon_n^{1/2}(d) \right](X) \Big)^2 \Big], \\
&=: U_1 + U_2. \qquad (7.4)
\end{aligned}$$

According to Theorem 3.1 we have

$$\begin{aligned}
\mathbb{P}\Big(\hat{d}(\delta) \neq d\Big) &\leq N\left(\tfrac{\Delta-\delta}{12(B+L)}, \mathscr{F}\right) \exp\left(-\tfrac{n(\Delta-\delta)^2}{18(B+L)^4}\right) \\
&\quad + N\left(\tfrac{\delta}{12(B+L)}, \mathscr{F}\right) \exp\left(-\tfrac{n\delta^2}{18(B+L)^4}\right) \\
&=: \mathbf{A}_n(\delta).
\end{aligned}$$

Hence

$$U_1 \leq (B+L)^2 \mathbf{A}_n(\delta). \tag{7.5}$$

Now, by Proposition 7.1 applied with $k = \lfloor \upsilon_n^{-1}(d) \rfloor$ and $\rho = \upsilon_n^{1/2}(d)$, we have

$$\begin{aligned} U_2 &\leq C_1 \left\{ \frac{1}{\lfloor \upsilon_n^{-1}(d) \rfloor} + \left(\frac{\lfloor \upsilon_n^{-1}(d) \rfloor}{n} \right)^{2/d} \right\} \\ &+ C_1 \left\{ \mathbf{D}_n(\delta) + \mathbf{T}_n(\upsilon_n^{1/2}(d)) + \frac{1}{\sqrt{n}} \right\}, \end{aligned}$$

where C_1, $\mathbf{D}_n(\delta)$ and $\mathbf{T}_n(\rho)$ have been defined in Proposition 7.1. Since $x-1 < \lfloor x \rfloor \leq x$, for all positive number x, we deduce that

$$\begin{aligned} &\frac{1}{\lfloor \upsilon_n^{-1}(d) \rfloor} + \left(\frac{\lfloor \upsilon_n^{-1}(d) \rfloor}{n} \right)^{2/d} \\ &\leq \frac{1}{\upsilon_n^{-1}(d)-1} + \left(\frac{\upsilon_n^{-1}(d)}{n} \right)^{2/d} \\ &= \left(\frac{1}{\upsilon_n^{-1}(d)-1} - \frac{1}{\upsilon_n^{-1}(d)} \right) + \frac{1}{\upsilon_n^{-1}(d)} + \left(\frac{\upsilon_n^{-1}(d)}{n} \right)^{2/d} \\ &= \frac{1}{(\upsilon_n^{-1}(d)-1)\upsilon_n^{-1}(d)} + 2\upsilon_n(d) \\ &\leq 3\upsilon_n(d). \end{aligned} \tag{7.6}$$

From the definitions, it is clear that

$$\mathbf{D}_n(\delta) \leq \mathbf{A}_n(\delta). \tag{7.7}$$

Next, we have

$$\begin{aligned} \mathbf{T}_n(\upsilon_n^{1/2}(d)) &= 2\upsilon_n(d) + \frac{pN\left(\upsilon_n^{1/2}(d), \mathscr{H}\right)}{n\upsilon_n(d)} \exp\left(-\tfrac{n\upsilon_n^2(d)}{C_1}\right) \\ &=: 2\upsilon_n(d) + \mathbf{B}_n. \end{aligned} \tag{7.8}$$

From inequalities (7.6), (7.7) and (7.8), we deduce that

$$U_2 \leq 5C_1\upsilon_n(d) + C_1 \left\{ \mathbf{A}_n(\delta) + \mathbf{B}_n + \frac{1}{\sqrt{n}} \right\}. \tag{7.9}$$

As a consequence, we conclude from (7.4), (7.5) and (7.9) that

$$\begin{aligned}\mathbb{E}\Big[\,(r(X)-\hat{r}(X))^2\,\Big] &\leq U_1+U_2\\ &\leq (B+L)^2\mathbf{A}_n(\delta)+5C_1\upsilon_n(d)+C_1\left\{\mathbf{A}_n(\delta)+\mathbf{B}_n+\frac{1}{\sqrt{n}}\right\}\\ &\leq 5C_1\upsilon_n(d)+C_1\left\{2\mathbf{A}_n(\delta)+\mathbf{B}_n+\frac{1}{\sqrt{n}}\right\},\end{aligned}$$

since $(B+L)^2 \leq C_1$. To conclude the proof, we need only to observe that we have

$$\mathbf{A}_n(\delta) = O\left(n^{-2/(2+d)}\right),$$

and that since $d \geq 3$, we have

$$\frac{1}{\sqrt{n}} = O\left(n^{-2/(2+d)}\right).$$

□

8 Technical results

Lemma 8.1 *Suppose $|Y| \leq B$ and suppose assumptions* (**A1**) *to* (**A4**) *hold. Then, for all $\ell \in \{1,\ldots,p\}$ and for all $\varepsilon > 0$, we have*

$$\mathbb{P}\left(|\hat{R}_\ell - R_\ell| \geq \varepsilon\right) \leq 2N\left(\tfrac{\varepsilon}{6(B+L)}, \mathscr{F}_\ell\right)\exp\left(-\tfrac{2n\varepsilon^2}{9(B+L)^4}\right).$$

PROOF – First, we have

$$\begin{aligned}\mathbb{P}\left(|\hat{R}_\ell - R_\ell| \geq \varepsilon\right) &= \mathbb{P}\left(\Big|\inf_{f\in\mathscr{F}_\ell}\frac{1}{n}\sum_{i=1}^n (Y_i-f(X_i))^2 - \inf_{f\in\mathscr{F}_\ell}\mathbb{E}(Y-f(X))^2\Big| \geq \varepsilon\right)\\ &\leq \mathbb{P}\left(\sup_{f\in\mathscr{F}_\ell}\Big|\frac{1}{n}\sum_{i=1}^n (Y_i-f(X_i))^2 - \mathbb{E}(Y-f(X))^2\Big| \geq \varepsilon\right).\end{aligned}$$

For all $f \in \mathscr{F}_\ell$ we denote $m_f : (x,y) \in \mathscr{X}\times[-B,B] \mapsto (y-f(x))^2$. Then, for all $f \in \mathscr{F}_\ell$, we have $|m_f(X,Y)| \leq (B+L)^2$. Therefore, according to Lemma 9.1 in Györfi et al. (2002) we have

$$\mathbb{P}\Big(|\hat{R}_\ell - R_\ell| \geq \varepsilon\Big) \leq 2N\left(\tfrac{\varepsilon}{3}, \{m_f :\ f\in\mathscr{F}_\ell\}\right)\exp\left(-\tfrac{2n\varepsilon^2}{9(B+L)^4}\right). \qquad (8.1)$$

Then, for any two functions $f, f' \in \mathscr{F}_\ell$, we have for all $(x,y) \in \mathscr{X} \times [-B,B]$

$$\begin{aligned} |m_f(x,y) - m_{f'}(x,y)| &= |(y-f(x))^2 - (y-f'(x))^2| \\ &= |(2y - f(x) - f'(x))(f(x) - f'(x))| \\ &\leq 2(B+L)|f(x) - f'(x)|. \end{aligned}$$

Hence, for all $\varepsilon > 0$, we have

$$N\left(\varepsilon, \left\{m_f : \ f \in \mathscr{F}_\ell\right\}\right) \leq N\left(\tfrac{\varepsilon}{2(B+L)}, \mathscr{F}_\ell\right). \tag{8.2}$$

Combining (8.1) and (8.2) yields the expected result. □

Lemma 8.2 *Suppose that assumptions* (**A1**) *to* (**A3**) *hold. Then, for all* $k \in \{1,\ldots,n\}$*, all* $\rho > 0$ *and all* $a > 0$*, we have*

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}[k,\rho](X_i))^2 - \mathbb{E}\Big[(Y - \hat{r}[k,\rho](X))^2 \,\Big|\mathscr{D}\Big]\right|\right] \leq C\left\{a + \frac{\mathbf{N}(\rho)}{na}\exp\left(-\tfrac{na^2}{C}\right)\right\},$$

where $C := \max\{1;\ 8B^4\}$.

Proof – Fix $k \in \{1,\ldots,n\}$, $\rho > 0$ and $a > 0$. Denote $\mathbf{H}(\rho) = \left\{h_j : j = 1,\ldots,\mathbf{N}(\rho)\right\}$. Since (X,Y) is independent from $\mathscr{D}$, since $\hat{r}_h[k] \circ h$ depends only on the subsample $\mathscr{D}_2 = \{(X_i,Y_i); i = n+1,\ldots,2n\}$ for all $h \in \mathscr{H}$ and since $\hat{h}[k,\rho]$ takes its values in $\mathbf{H}(\rho)$, we have

$$\begin{aligned} \mathbb{E}\Big[(Y - \hat{r}[k,\rho](X))^2 \,\Big|\mathscr{D}\Big] &= \mathbb{E}\left[\left(Y - \sum_{j=1}^{\mathbf{N}(\rho)} \hat{r}_{h_j}[k](h_j(X))\mathbf{1}\left\{\hat{h}[k,\rho] = h_j\right\}\right)^2 \,\middle|\, \mathscr{D}\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{\mathbf{N}(\rho)} \left(Y - \hat{r}_{h_j}[k](h_j(X))\right)^2 \mathbf{1}\left\{\hat{h}[k,\rho] = h_j\right\} \,\middle|\, \mathscr{D}\right] \\ &= \sum_{j=1}^{\mathbf{N}(\rho)} \mathbb{E}\Big[\left(Y - \hat{r}_{h_j}[k](h_j(X))\right)^2 |\mathscr{D}\Big] \mathbf{1}\left\{\hat{h}[k,\rho] = h_j\right\} \\ &= \sum_{j=1}^{\mathbf{N}(\rho)} \mathbb{E}\Big[\left(Y - \hat{r}_{h_j}[k](h_j(X))\right)^2 |\mathscr{D}_2\Big] \mathbf{1}\left\{\hat{h}[k,\rho] = h_j\right\}. \end{aligned}$$

Therefore, denoting $E_j := \left\{\hat{h}[k,\rho] = h_j\right\}$, we obtain

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}[k,\rho](X_i))^2 - \mathbb{E}\Big[(Y - \hat{r}[k,\rho](X))^2 |\mathscr{D}\Big] \\ &\quad = \sum_{j=1}^{\mathbf{N}(\rho)} \left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{r}_{h_j}[k](h_j(X_i))\right)^2 - \mathbb{E}\left[\left(Y - \hat{r}_{h_j}[k](h_j(X))\right)^2 |\mathscr{D}_2\right]\right\} \mathbf{1}\left\{E_j\right\}. \end{aligned}$$

Now for all $i \in \{1,\dots,n\}$ and all $j \in \{1,\dots,\mathbf{N}(\rho)\}$, let

$$Z_{i,j} := \left(Y_i - \hat{r}_{h_j}[k](h_j(X_i))\right)^2 - \mathbb{E}\Big[\left(Y - \hat{r}_{h_j}[k](h_j(X))\right)^2 | \mathscr{D}_2\Big].$$

Using the fact that the events E_j are pairwise disjoint, we deduce that for all $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}[k,\rho](X_i))^2 - \mathbb{E}\Big[(Y - \hat{r}[k,\rho](X))^2 | \mathscr{D}\Big]\Big| \geq \varepsilon\right) \\
&= \quad \mathbb{P}\left(\Big|\sum_{j=1}^{\mathbf{N}(\rho)}\left\{\frac{1}{n}\sum_{i=1}^{n} Z_{i,j}\right\}\mathbf{1}\left\{E_j\right\}\Big| \geq \varepsilon\right) \\
&= \quad \mathbb{P}\left(\max_{j=1,\dots,\mathbf{N}(\rho)}\Big|\frac{1}{n}\sum_{i=1}^{n} Z_{i,j}\Big| \geq \varepsilon\right) \\
&\leq \quad \mathbf{N}(\rho)\max_{j=1,\dots,\mathbf{N}(\rho)}\mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^{n} Z_{i,j}\Big| \geq \varepsilon\right),
\end{aligned}$$

where the last inequality follows from the union bound. Now, conditionally to $\mathscr{D}_2$, and for all $j \in \{1,\dots,\mathbf{N}(\rho)\}$, the variables

$$\left(Y - \hat{r}_{h_j}[k](h_j(X))\right)^2 \quad \text{and} \quad \left(Y_i - \hat{r}_{h_j}[k](h_j(X_i))\right)^2,\ i \in \{1,\dots,n\}$$

are i.i.d. an bounded by $4B^2$. By Hoeffding's inequality, it follows that for all $j \in \{1,\dots,\mathbf{N}(\rho)\}$

$$\mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^{n} Z_{i,j}\Big| \geq \varepsilon \Big| \mathscr{D}_2\right) \leq 2\exp\left(-\tfrac{n\varepsilon^2}{8B^4}\right).$$

Therefore

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{r}[k,\rho](X_i))^2-\mathbb{E}\Big[(Y-\hat{r}[k,\rho](X))^2\,\Big|\mathscr{D}\Big]\right|\right]\\
&\quad=\quad\int_0^{+\infty}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{r}[k,\rho](X_i))^2-\mathbb{E}\Big[(Y-\hat{r}[k,\rho](X))^2\,\Big|\mathscr{D}\Big]\right|\geq\varepsilon\right)\mathrm{d}\varepsilon\\
&\quad\leq\quad a+\int_a^{+\infty}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{r}[k,\rho](X_i))^2-\mathbb{E}\Big[(Y-\hat{r}[k,\rho](X))^2\,\Big|\mathscr{D}\Big]\right|\geq\varepsilon\right)\mathrm{d}\varepsilon\\
&\quad\leq\quad a+\mathbf{N}(\rho)\int_a^{+\infty}\max_{j=1,\ldots,\mathbf{N}(\rho)}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}Z_{i,j}\right|\geq\varepsilon\right)\mathrm{d}\varepsilon\\
&\quad=\quad a+\mathbf{N}(\rho)\int_a^{+\infty}\max_{j=1,\ldots,\mathbf{N}(\rho)}\mathbb{E}\left[\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}Z_{i,j}\right|\geq\varepsilon\Big|\mathscr{D}_2\right)\right]\mathrm{d}\varepsilon\\
&\quad\leq\quad a+2\mathbf{N}(\rho)\int_a^{+\infty}\exp\left(-\tfrac{n\varepsilon^2}{8B^4}\right)\mathrm{d}\varepsilon.
\end{aligned}$$

Now, using the fact that for all $x>0$

$$\int_x^{+\infty}\exp\left(-\tfrac{\tau^2}{2}\right)\mathrm{d}\tau\leq\frac{1}{x}\exp\left(-\tfrac{x^2}{2}\right),$$

we obtain

$$\begin{aligned}
\int_a^{+\infty}\exp\left(-\tfrac{n\varepsilon^2}{8B^4}\right)\mathrm{d}\varepsilon\quad&=\quad\frac{2B^2}{\sqrt{n}}\int_{\frac{a\sqrt{n}}{2B^2}}^{+\infty}\exp\left(-\tfrac{\tau^2}{2}\right)\mathrm{d}\tau\\
&\leq\quad\frac{4B^4}{an}\exp\left(-\tfrac{na^2}{8B^4}\right).
\end{aligned}$$

This leads to

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{r}[k,\rho](X_i))^2-\mathbb{E}\Big[(Y-\hat{r}[k,\rho](X))^2\,\Big|\mathscr{D}\Big]\right|\right]\\
&\quad\leq\quad a+\frac{8B^4\mathbf{N}(\rho)}{an}\exp\left(-\tfrac{na^2}{8B^4}\right)\\
&\quad\leq\quad C\left\{a+\frac{\mathbf{N}(\rho)}{an}\exp\left(-\tfrac{na^2}{C}\right)\right\},
\end{aligned}$$

with $C:=\max\{1;\ 8B^4\}$, as desired. □

Lemma 8.3 *Suppose that* $|Y| \leq B$ *and that assumptions* (**A1**) *to* (**A5**) *are satisfied. Then, for all* $k \in \{1,\ldots,n\}$*, for all* $\rho > 0$ *and for all* $h \in \mathbf{H}_d(\rho)$ *satisfying* $\|h - h^*\|_\infty \leq \rho$*, we have*

$$\mathbb{E}\Big[\,(r(X) - \hat{r}_h[k](h(X)))^2\,\Big] \leq C\left\{\frac{1}{k} + \left(\frac{k}{n}\right)^{2/d}\right\} + C\rho^2,$$

where $C := \max\Big\{2(B+L)^2;\ 10(K^2+L^2)\ ;\ 96L^2R^2 4^{1/d}\left(\frac{d-2}{2}\right)^{4/d}\Big\}$.

The proof of Lemma 8.3 is very similar to that of Theorem 6.2 in Györfi et al. (2002) or Theorem 2.2 in Cadre and Dong (2010). Therefore, it has been reported in the supplementary material.

A Reduced dimension d and parameter Δ

In this appendix, we prove equations (2.2) and (2.4). First, observe that since $\mathscr{F}_\ell$ is compact in $\mathbb{L}^2(\mu)$ and since $r \in \mathscr{F}$, we have

$$\begin{aligned}
r \in \mathscr{F}_\ell &\Leftrightarrow \inf_{f\in\mathscr{F}_\ell} \mathbb{E}\,(f(X) - r(X))^2 = 0 \\
&\Leftrightarrow \inf_{f\in\mathscr{F}_\ell} \mathbb{E}\,(Y - f(X))^2 - \mathbb{E}\,(Y - r(X))^2 = 0 \\
&\Leftrightarrow \inf_{f\in\mathscr{F}_\ell} \mathbb{E}\,(Y - f(X))^2 - \inf_{f\in\mathscr{F}} \mathbb{E}\,(Y - f(X))^2 = 0 \\
&\Leftrightarrow R_\ell = R_p.
\end{aligned}$$

Therefore, since the function $\ell \in \{1,\ldots,p\} \mapsto R_\ell$ is non-increasing, we deduce that

$$\begin{aligned}
d &:= \min\Big\{\ell : r \in \mathscr{F}_\ell\Big\} \\
&= \min\Big\{\ell : R_\ell = R_p\Big\},
\end{aligned}$$

which proves equation (2.2). Using (2.2) and the fact that $r \in \mathscr{F}$ we obtain that

$$\begin{aligned}
\Delta &= \min\Big\{\ R_\ell - R_p :\ R_\ell > R_p\ \Big\} \\
&= R_{d-1} - R_p \\
&= \inf_{f\in\mathscr{F}_{d-1}} \mathbb{E}\,(Y - f(X))^2 - \inf_{f\in\mathscr{F}} \mathbb{E}\,(Y - f(X))^2 \\
&= \inf_{f\in\mathscr{F}_{d-1}} \mathbb{E}\,(Y - f(X))^2 - \mathbb{E}\,(Y - r(X))^2 \\
&= \inf_{f\in\mathscr{F}_{d-1}} \mathbb{E}\,(f(X) - r(X))^2 \\
&= \inf_{f\in\mathscr{F}_{d-1}} \|f - r\|_\mu^2,
\end{aligned}$$

which proves (2.4).

Acknowledgments – The author is indebted to Benoit Cadre and Bruno Pelletier for their time and help.

Supplementary material – Supplement to Nonlinear dimension reduction for regression with nearest neighbors.

References

B. Cadre and Q. Dong. Dimension reduction in regression estimation with nearest neighbor. *Electronic Journal of Statistics*, 4:436–460, 2010.

R.D. Cook. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York, 1998.

R.D. Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22:1–26, 2007.

R.D. Cook and B. Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30:455–474, 2002.

R.D. Cook and L. Ni. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *Journal of the American Statistical Association*, 100:410–428, 2005.

R.D. Cook and S. Weisberg. Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–342, 1991.

B. Delyon and F. Portier. Optimal transformation: a new approach for covering the central subspace (to appear). *Journal of Multivariate Analysis*, 2013.

K. Fukumizu, F. Bach, and M. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37:1871–1905, 2009.

W.K. Fung, X. He, L. Liu, and P. Shi. Dimension reduction based on canonical correlation. *Statistica Sinica*, 12:1093–1113, 2002.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

W. Härdle and T.M. Stoker. Investigating smooth multiple regression by the method of average derivative. *Journal of the American Statistical Association*, 84:986–995, 1989.

I.A. Ibragimov and R.Z. Khasminskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.

M. Kohler, A. Krzyżak, and H. Walk. Optimal global rates of convergence in nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 123:1286–1296, 2009.

K.C. Li. Sliced inverse regression for dimension reduction (with discussions). *Journal of the American Statistical Association*, 86:316–342, 1991.

K.C. Li. On principal hessian directions for data visualization and dimension reduction: another application of steins lemma. *Journal of the American Statistical Association*, 87:1025–1039, 1992.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

Q. Wang and X. Yin. A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave. *Computational Statistics and Data Analysis*, 52:4512–4520, 2008.

H.M. Wu. Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics*, 17:590–610, 2008.

Y. Xia, H. Tong, W.K. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society (B)*, 64:1–28, 2002.

Y.-R. Yeh, S.-Y. Huang, and Y.-Y. Lee. Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering*, 21:1590–1603, 2009.

X. Yin, B. Li, and R.D. Cook. Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99:1733–1757, 2008.