.

Variance estimation for high-dimensional regression models *

Vladimir Spokoiny†

Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117 Berlin, Germany
E-mail: spokoiny@wias-berlin.de

Received ???; revised ???; accepted ???

The paper is concerned with the problem of variance estimation for a high-dimensional regression model. The results show that the accuracy $n^{-1/2}$ of variance estimation can be achieved only under some restrictions on smoothness properties of the regression function and on the dimensionality of the model. In particular, for a two times differentiable regression function, the rate $n^{-1/2}$ is achievable only for dimensionality smaller or equal to 8. For higher dimensional model, the optimal accuracy is $n^{-4/d}$ which is worse than $n^{-1/2}$. The rate optimal estimating procedure is presented.

Key Words: variance estimation,regression,high dimension

1. INTRODUCTION

In this paper, we consider the problem of estimating the error variance for the regression model

$$Y_i = f(X_i) + \varepsilon_i \tag{1}$$

where $X_1, \ldots, X_n$ are design points in the Euclidean space $\mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown regression function and $\varepsilon_1, \ldots, \varepsilon_n$ are individual random errors which are supposed independent and satisfying the conditions $\boldsymbol{E}\varepsilon_i = 0$, $\boldsymbol{E}\varepsilon_i^2 = \sigma^2$ and $\boldsymbol{E}\varepsilon_i^6 \le C_6 < \infty$ for all $i \le n$. The design $X_1, \ldots, X_n$ is assumed deterministic. Note however that the case of a random design can be considered as well, supposing $X_1, \ldots, X_n$ i.i.d. random points in $\mathbb{R}^d$ with a design density $p(x)$. Then all the result should be understood conditionally on the design.

* *1995 Mathematics Subject Classification:* 62G05; Secondary 62G20
† The research was partially supported by the DFG-RFFI Grant ???

The aim is to estimate the unknown error variance σ^2.

Wahba (1983) and Silverman (1985) proposed to use for estimating σ^2 usual nonparametric residuals obtained by removing the estimated smooth regression curve from the observations. Difference-based procedures were thoroughly discussed in Gasser et al. (1986), Siefert et al. (1993) among other. Hall et al (1990) found asymptotically optimal differences. Choosing the curve estimation with respect to extracting residual variance has been studied by Buckley et al (1988) and Hall and Marron (1990). We refer to Seifert et al (1993) for more detailed descriptions and comparison of different procedures for variance estimation. Neumann (1994) discussed fully data-driven estimate. Hall and Carroll (1989), Härdle and Tsybakov (1997), Ruppert et al (1997), Fan and Yao (1988) studied the problem of estimating the heteroscedastic conditional variance.

The majority of the mentioned results focus on the mean squared error of the variance estimation in the univariate regression model and claim the possibility to estimate σ^2 at the rate $n^{-1/2}$. Some extensions to the two-dimensional case are discussed in Hall et al. (1991) and Seifert et al. (1993). The main message of the present paper is that variance estimation with root-n rate is possible in the multivariate case as well, but only if the dimension d is not too high, more precisely, if $d \le 8$.

It is worth noting that the variance estimation is relatively rarely the target of statistical analysis. Typically it is used as a building block for further procedure like adaptive estimation (Rice, 1984; Gasser et al, 1991) of hypothesis testing (Hart, 1997), Spokoiny (1999), where some pilot variance estimation is required. This enforces to study not only the risk of estimation but also some deviation probabilities which are presented in our results.

2. THE ESTIMATE

Our approach is a multidimensional analog of the proposal from Hart (1997, p.123) which gives an unbiased estimate of the variance for a linear regression function. The idea is to construct for every design point X_i a local linear fit $\widehat{f}(X_i)$ of the unknown regression function f and then to use the *pseudo-residuals* $\widehat{e}_i = Y_i - \widehat{f}(X_i)$ for variance estimation.

The main problem comes from design sparseness and non regularity in the multidimensional situation. This makes difficult the choice of the local neighborhood for constructing the local linear fit. We propose below two approaches how this choice can be done. One utilizes a uniform bandwidth and another one allows the bandwidth to vary from point to point.

2.1. The local linear fit

First we describe the local linear fit we apply. Let $U_h(x)$ denote the ball with the center x and the radius h and $N_h(x)$ stand for the number of different design points in $U_h(x)$: $N_h(x) = \#\{X_i \in U_h(x)\}$.

Let K be the uniform kernel function $K(u) = \mathbf{1}(|u| \le 1)$. Introduce linear functions $\psi_0(x) \equiv 1$, $\psi_\ell(x) = x_\ell$, $\ell = 1, \ldots, d$ and define for every i the vector $\widehat{\theta}_h(X_i) \in \mathbb{R}^{d+1}$ via the local linear fit

$$\widehat{\theta}_h(X_i) \;=\; \operatorname*{arginf}_{\theta \in \mathbb{R}^{d+1}} \sum_{j=1}^{n} \left(Y_j - \sum_{\ell=0}^{d} \theta_\ell \psi_\ell(X_j) \right)^2 K\left(\frac{X_j - X_i}{h} \right)$$

see Katkovnik (1985), Tsybakov (1986), Fan and Gijbels (1996). This is a quadratic optimization problem with respect to the vector of coefficients

$\theta = (\theta_\ell)_{\ell=0,\dots,d}$ which can be solved explicitly. If the $(d+1)\times(d+1)$ matrix $\Psi_{i,h}$ of the form

$$\Psi_{i,h} = \left(\sum_{j=1}^{n} \psi_\ell(X_j)\, \psi_k(X_j)\, K\left(\frac{X_j - X_i}{h}\right), \ell, k = 0, \dots, d \right).$$

is non singular, then the solution exists and is unique and it is a linear combination of the observations Y_j with the deterministic coefficients depending on the design $X_1, \dots, X_n$ only. In particular, the first coefficient can be represented in the form $\widehat{\theta}_{0,h}(X_i) = \sum_{j=1}^{n} a_{ij,h} Y_j$ with some coefficients $a_{ij,h}$, $j = 1, \dots, n$. It is well known (and it is easy to check) that the such defined coefficients $a_{ij,h}$ obey the following conditions.

LEMMA 2.1. *Let the matrix $\Psi_{i,h}$ be non singular. Then the above defined coefficients $a_{ij,h}$ fulfill $a_{ij,h} = 0$ if $|X_j - X_i| \geq h$ and*

$$\begin{aligned} \sum_{j=1}^{n} a_{ij,h}\, K\left(\frac{X_j - X_i}{h}\right) &= 1, \\ \sum_{j=1}^{n} a_{ij,h}\, \psi_\ell(X_j - X_i) K\left(\frac{X_j - X_i}{h}\right) &= 0, \qquad \ell = 1, \dots, d. \end{aligned}$$

A necessary and usually sufficient condition for non singularity of the matrix $\Psi_{i,h}$ is that the ball $U_h(X_i)$ contains at least $d+1$ design points.

2.2. Procedure with a variable bandwidth

For every i, define the bandwidth h_i by the condition

$$h_i = \inf \{h : \Psi_{i,h} \text{ is non singular}\}$$

where $\Psi_{i,h}$ is the $(d+1)\times(d+1)$ matrix introduced before Lemma 2.1.

Next define the local linear estimate

$$\widehat{f}(X_i) = \widehat{f}_{h_i}(X_i) = \sum_{j=1}^{n} a_{ij,h_i} Y_j$$

and *pseudo residuals* $\widehat{e}_i$

$$\widehat{e}_i = \widehat{f}(X_i) - Y_i = \sum_{j=1}^{n} c_{ij} Y_j$$

with $c_{ij} = a_{ij,h_i}$ for $j \neq i$ and $c_{ii} = a_{ii,h_i} - 1$. Finally we set

$$\begin{aligned} s_i^2 &= \sum_{j=1}^{n} c_{ij}^2, \qquad i = 1, \dots, n, \\ \widehat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} \frac{|\widehat{e}_i|^2}{s_i^2}. \end{aligned} \tag{2}$$

2.3. Procedure with a global bandwidth

Define the subset $\mathcal{X}_h$ of the set $X_1, \dots, X_n$ by

$$\mathcal{X}_h = \{X_i : \Psi_{i,h} \text{ is non singular}\}$$

and let M_h stand for the number of design points in $\mathcal{X}_h$: $M_h = \#\mathcal{X}_h$. Then, with a given $\alpha \geq 1/2$, we define the bandwidth h as the minimal value which satisfies the condition

$$M_h \geq n\alpha,$$

that is, there are at least $n\alpha$ points X_i, for which $\Psi_{i,h}$ is non singular. Next we define the local linear estimate $\widehat{f}(X_i)$ by $\widehat{f}(X_i) = \sum_{j=1}^n a_{ij,h} Y_j$ and the *pseudo residuals* $\widehat{e}_i$ by

$$\widehat{e}_i = \widehat{f}(X_i) - Y_i = \sum_{j=1}^n c_{ij} Y_j$$

with $c_{ij} = a_{ij,h}$ for $j \neq i$ and $c_{ii} = a_{ii,h} - 1$. Finally the variance estimate $\widehat{\sigma}^2$ is defined by

$$\begin{aligned} s_i^2 &= \sum_{j=1}^n c_{ij}^2, \\ \widehat{\sigma}_v^2 &= \frac{1}{M_h} \sum_{i: X_i \in \mathcal{X}_h} \frac{\widehat{e}_i^2}{s_i^2}. \end{aligned}$$

3. PROPERTIES

In this section we state some useful properties of the estimate $\widehat{\sigma}^2$ from (2). The estimate $\widehat{\sigma}_v^2$ can be studied similarly. First we present the result for the case of Gaussian errors ε_i and then we discuss the general case.

The estimate $\widehat{\sigma}^2$ assumes some smoothness of the regression function f in a small neighborhood of each design point X_i. When formulating the result, this local smoothness will be characterized by the value

$$L_i = 0.5 \sup_{u \in \mathbb{R}^d} \sup_{x \in U_{h_i}(X_i)} \frac{u^\top f''(x) u}{|u|^2}$$

where f'' denotes the $d \times d$ Hessian matrix of second derivatives of f.

THEOREM 3.1. *Let the observations $Y_1, \ldots, Y_n$ follow the regression model (1) with i.i.d. Gaussian errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and a two times differentiable regression function f. Introduce $n \times n$-matrix B with entries*

$$\beta_{ij} = \frac{1}{n} \sum_{k=1}^n s_k^{-2} c_{ki} c_{kj}, \qquad i, j = 1, \ldots, n.$$

Define the values Δ, S^2 and C_B by

$$\begin{aligned} \Delta^2 &= \frac{1}{n} \sum_{i=1}^n L_i^2 h_i^4 s_i^{-2} \left(\sum_{j \neq i} |c_{ij}| \right)^2, \\ S^2 &= 2 \operatorname{tr} B^2 = 2 \sum_{i=1}^n \sum_{j=1}^n \beta_{ij}^2, \\ C_B^2 &= \frac{n \|B\|_\infty^2}{2 \operatorname{tr} B^2} \end{aligned}$$

where $\|B\|_\infty = \sup_{u\in\mathbb{R}^n} |Bu|/|u|$. *Then for every nonnegative* γ, *the variance estimate* $\widehat{\sigma}^2$ *fulfills*

$$\boldsymbol{P}\left(\pm(\widehat{\sigma}^2 - \sigma^2) > \Delta^2 + \gamma\sigma\Delta\sqrt{2\|B\|_\infty} + \gamma\sigma^2 S\right) \leq 2e^{-\gamma^2/4} + e^{\gamma\sqrt{n}/(6C_B)}. \tag{3}$$

Remark 3. 1. The norm of the matrix B can be very roughly estimated as follows: $\|B\|_\infty^2 \leq \operatorname{tr} B^2$, which particularly implies $C_B \leq \sqrt{n/2}$.

3.1. The rate of estimation

Here we discuss some corollaries of Theorem 3.1 concerning the rate of estimation. To this end we have to bound the quantities Δ and S. This can be easily done under some additional assumptions on the design $X_1, \ldots, X_n$ and the underlying regression function f. Concerning the design, we consider here two different model assumptions widely used in applications.

RD **(Random design)** The design points $X_1, \ldots, X_n$ are i.i.d. random variables from a distribution with a density $p(x)$ which is supported on a compact set $\mathcal{X}$ and it is continuous and positive on $\mathcal{X}$.

ED **(Equispaced design)** The design points $X_1, \ldots, X_n$ form the regular grid in the unit cube $[0,1]^d$ with the step δ_n such that δ_n^{-1} is an integer number and $\delta_n^{-d} = n$.

The quantity S^2 is defined through the design only and in what follows we present some bound on S under ED or RD. The value Δ also depends on the smoothness properties of the underlying regression function f. For exposition simplicity we restrict ourselves to the class $\mathcal{F}(2, L)$ of functions with the bounded second derivative:

$$\mathcal{F}(2, L) = \{f : 0.5\|f\|_\infty \leq L\}.$$

For every $f \in \mathcal{F}(2, L)$, the values L_i defined before Theorem 3.1 are bounded by L, i.e. $L_i \leq L$.

LEMMA 3.1. *Let* $f \in \mathcal{F}(2, L)$. *Under ED, it holds*

$$\begin{aligned} \Delta^2 &\leq 2dL^2 n^{-4/d}, \\ S^2 &\leq 2N^* n^{-1}, \end{aligned}$$

where N^* *depends on* d *only.*

Next we consider the situation with a random design. In that case, the both quantities Δ and S which are defined via the design $X_1, \ldots, X_n$, are random and the result of Theorem 3.1 is stated conditionally on the design. The bounds we formulate below should be also understood in the conditional sense: they hold for a majority of design realizations (i.e. on a set of a high probability w.r.t. the design distribution).

LEMMA 3.2. *Let* $f \in \mathcal{F}(2, L)$ *and let RD hold. For every* $\mu > 0$, *there are two constants* κ *and* N^* *depending on* d *and the design density* p

only such that it holds for n large enough on the set of probability at least $1-\mu$

$$\begin{aligned} \Delta^2 &\le \kappa^2 L^2 n^{-4/d}, \\ S^2 &\le 2N^* n^{-1}. \end{aligned}$$

The inequalities $\Delta \le L\kappa n^{-2/d}$ and $S^2 \le 2N^* n^{-1}$ yield in view of (3) and Remark 3.1 the following accuracy of estimation: with a probability at least $1 - e^{-\gamma^2/4} - e^{-\gamma\sqrt{2}/6}$, it holds

$$\begin{aligned} \pm\left(\widehat{\sigma}^2 - \sigma^2\right) &\le \Delta^2 + \gamma\Delta\sigma(2S^2)^{1/4} + \gamma S\sigma^2 \\ &\le \kappa^2 L^2 n^{-4/d} + \gamma\kappa L(4N^*)^{1/4}\sigma n^{-2/d-1/4} + \gamma\sigma^2\sqrt{2N^*}n^{-1/2}. \end{aligned}$$

We observe that for $d < 8$, the first two summands in this bound are smaller in rate than the last one which is $O(n^{-1/2})$. If $d = 8$, then all three summands are of order $n^{-1/2}$ and for $d > 8$, the first term (which is of order $n^{-4/d}$) starts to dominate. Given a loss function w, define the risk of estimation

$$R(\widehat{\sigma}^2) = \begin{cases} \boldsymbol{E} w\left(n^{1/2}\sigma^{-2}(\widehat{\sigma}^2 - \sigma^2)\right), & d \le 8, \\ \boldsymbol{E} w\left(n^{4/d}\sigma^{-2}(\widehat{\sigma}^2 - \sigma^2)\right), & \text{otherwise.} \end{cases}$$

The above considerations lead to the following

THEOREM 3.2. *Let $\widehat{\sigma}^2$ be the variance estimate from (2). Let the quantities Δ, $\|B\|_\infty$ and S defined in Theorem 3.1 and depending on n, the design $X_1, \dots, X_n$ and on the smoothness properties of the regression function f, satisfy the conditions*

$$\begin{aligned} \Delta &\le D\sigma^2 n^{-2/d}, \\ S^2 &\le 2N^* n^{-1} \end{aligned}$$

with some fixed constants B, N^*. Then for every continuously differentiable loss function w which obeys the conditions $w(0) = 0$, $w(x) = w(-x)$, $w'(x) \ge 0$ for $x > 0$ and $\int w'(x)e^{-\alpha x}\,\mathrm{d}x < \infty$ for every $\alpha > 0$, the corresponding risk $R(\widehat{\sigma}^2)$ remains bounded by some constant $C = C(B, N^*, w)$ depending on D, N^* and the function w only:

$$R(\widehat{\sigma}^2) \le C(D, N^*, w).$$

3.2. Non-Gaussian case

Here we discard the assumption that the errors ε_i are normally distributed. Instead we assume that that they are independent identically distributed with 6 finite moments.

THEOREM 3.3. *Let the errors ε_i from (1) be i.i.d. random variables with $\boldsymbol{E}\varepsilon_i = 0$, $\boldsymbol{E}\varepsilon_i^2 = \sigma^2$, $\boldsymbol{E}(\varepsilon_i^2 - \sigma^2)^2 \le C_4^2\sigma^4$ and $\boldsymbol{E}|\varepsilon_i^2 - \sigma^2|^3 \le C_6\sigma^6$ for all i. Let also value C_A be such that*

$$\frac{n \max\limits_{i=1,\dots,n} \sum\limits_{j=1}^n \beta_{ij}^2}{\sum\limits_{i=1}^n \sum\limits_{j=1}^n \beta_{ij}^2} \le C_A, \qquad \frac{n \max\limits_{i=1,\dots,n} \beta_{ii}^2}{\sum\limits_{i=1}^n \beta_{ii}^2} \le C_A$$

where the coefficients β_{ij} *are defined in Theorem 3.1. Then there exists an absolute constant* C *such that for every* $\gamma \geq 0$ *and every* δ *with* $0 < \delta \leq 1$

$$\boldsymbol{P}\left(\pm(\widehat{\sigma}^2 - \sigma^2) > \Delta^2 + 2\Delta S^{1/2}\sigma + (\gamma+\delta)S\sigma^2 + \gamma S''\sigma^2\right) \leq 2e^{-\gamma^2/4} + e^{-\gamma\sqrt{n}/(6C_A)} + Cn^{-1/2}\delta^{-3}$$

where Δ *and* S *are defined in Theorem 3.1,* $|S''|^2 = \sum_{i=1}^{n} \beta_{ii}^2$ *and the constant* C *depends on* C_4, C_6 *and* C_A *only.*

This result clearly implies an analog of Theorem 3.2 for non-Gaussian errors under the conditions of Theorem 3.3.

3.3. Rate optimality

Here we show that the critical dimension $d = 8$ appears not only for our particular estimator. Actually, no estimator achieves the rate $n^{-1/2}$ for $d > 8$ uniformly over any class of smooth functions with the smoothness degree 2.

To simplify the construction, we suppose hereafter that $n^{1/d}$ is an integer number, and $X_1, \ldots, X_n$ form the regular grid in the unit cube $[0,1]^d$. Define the following Sobolev type class $\mathcal{F}_n(2, L)$:

$$\mathcal{F}_n(2,L) = \left\{ f : \frac{1}{n}\sum_{i=1}^{n} \sup_{x:|x-X_i|\leq n^{-1/d}} \|f''(x)\|^2 \leq L^2 \right\}.$$

Let $\boldsymbol{P}_{f,\sigma^2}$ denote the measure on the observation space which corresponds to a regression function f and the variance σ^2 and let $\boldsymbol{E}_{f,\sigma^2}$ denote the expectation w.r.t. $\boldsymbol{P}_{f,\sigma^2}$.

THEOREM 3.4. *Let* $X_1, \ldots, X_n$ *be the equispaced design in the unit cube* $[0,1]^d$ *and the the observations* $Y_1, \ldots, Y_n$ *be generated from the regression model (1) with i.i.d. Gaussian errors* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. *For* $d \geq 8$, *sufficiently large* L *and for every continuous bounded loss function* w,

$$\lim_{n\to\infty} \inf_{\widetilde{\sigma}_n^2} \sup_{f\in\mathcal{F}_n(2,L)} \sup_{\sigma^2\in\Sigma_n} \boldsymbol{E}_{f,\sigma^2}\, w\left(n^{4/d}(\widetilde{\sigma}_n^2 - \sigma^2)\right) = r > 0$$

where the infimum is taken over the set of all possible estimates of the parameter σ^2 *and* Σ_n *is the three points set of the form* $\Sigma_n = \{1, 1 + n^{-4/d}, 1 + 2n^{-4/d}\}$.

Due to this result, even if the unknown variance is valued in a three-point set Σ_n, a consistent variance estimation is impossible and the risk of estimation is of order $n^{-1/d}$.

4. PROOFS

In this section we present the proofs of Theorem 3.1 through 3.4.

4.1. Proof of Theorem 3.1

Define

$$f_{h_i}(X_i) = \sum_{j=1}^{n} a_{ij,h_i} f(X_j)$$

so that

$$\sum_{j=1}^{n} c_{ij} f(X_j) = \sum_{j=1}^{n} a_{ij,h_i} f(X_j) - f(X_i) = f_{h_i}(X_i) - f(X_i).$$

The model equation (1) implies for every $i \le n$

$$\widehat{e}_i = \sum_{j=1}^{n} c_{ij} Y_j = \sum_{j=1}^{n} c_{ij} f(X_j) + \sum_{j=1}^{n} c_{ij} \varepsilon_j$$

which leads to the following representation for the estimate $\widehat{\sigma}^2$:

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{|\widehat{e}_i|^2}{s_i^2} = \sum_{i=1}^{n} (b_i + \xi_i)^2 = |b + \xi|^2$$

where

$$\begin{aligned} b_i &= n^{-1/2} s_i^{-1} \{ f_{h_i}(X_i) - f(X_i) \}, \\ \xi_i &= n^{-1/2} s_i^{-1} \sum_{j=1}^{n} c_{ij} \varepsilon_j = \sum_{j=1}^{n} \alpha_{ij} \varepsilon_j \end{aligned}$$

with $\alpha_{ij} = n^{-1/2} s_i^{-1} c_{ij}$.

The smoothness assumption on the function f implies for every j with $|X_j - X_i| \le h_i$

$$|f(X_j) - f(X_i) - f'(X_i)(X_j - X_i)| \le L_i h_i^2.$$

The properties $\sum_{j=1}^{n} c_{ij} = 0$ and $\sum_{j=1}^{n} c_{ij}(X_j - X_i) = 0$ provide

$$\begin{aligned} &|f_{h_i}(X_i) - f(X_i)| \\ &\quad = \left| \sum_{j=1}^{n} c_{ij} f(X_j) - f(X_i) \sum_{j=1}^{n} c_{ij} - f'(X_i) \sum_{j=1}^{n} c_{ij} (X_j - X_i) \right| \\ &\quad = \left| \sum_{j=1}^{n} c_{ij} \{ f(X_j) - f(X_i) - f'(X_i)(X_j - X_i) \} \right| \\ &\quad \le L_i \, h_i^2 \sum_{j \ne i} |c_{ij}|. \end{aligned}$$

Therefore

$$|b|^2 = \sum_{i=1}^{n} b_i^2 \le \frac{1}{n} \sum_{i=1}^{n} L_i^2 h_i^4 s_i^{-2} \left(\sum_{j \ne i} |c_{ij}| \right)^2 = \Delta^2. \tag{4}$$

We now apply the following general statement, see Lemma 4.3 below. Let A be a $n \times n$-matrix with entries α_{ij}, $\lambda_A = \|A^\top A\|_\infty$ and $S^2 = 2 \operatorname{tr}(A^\top A)^2$. Then for every positive $\gamma > 0$

$$\begin{aligned} \boldsymbol{P} \left(\pm \left(|b + \xi|^2 - |b|^2 - \sigma^2 \operatorname{tr}(A^\top A) \right) > \gamma \sigma |b| (2\lambda_A)^{1/2} + \gamma \sigma^2 S \right) \\ \le 2 e^{-\gamma^2/4} + e^{-\gamma S / (6 \lambda_A)}. \end{aligned}$$

Since $\sum_{j=1}^n \alpha_{ij}^2 = 1/n$, then clearly

$$\mathrm{tr}(A^\top A) = \sum_{i=1}^n \beta_{ii} = \sum_{i=1}^n \sum_{j=1}^n \alpha_{ij}^2 = 1.$$

This implies the required assertion in view of (4).

4.2. Proof of Lemmas 3.1 and 3.2

Recall that each bandwidth h_i is defined as the smallest radius h providing a non degenerated linear fit in the ball $U_h(X_i)$. This implies that the number $N_h(X_i)$ of design points in the ball $U_h(X_i)$ is at least $d+1$. Define $\overline{N} = \max_i N_{h_i}(X_i) - 1$. It is straightforward to see that under RD, $\boldsymbol{P}\left(\overline{N} = d\right) = 1$, and under ED, it holds $\overline{N} = 2d$.

Further, let $\overline{h} = \left(n^{-1}\sum_{i=1}^n h_i^4\right)^{1/4}$. Under ED, one clearly has $h_i = n^{-1/d}$ for all i, so that $\overline{h} = n^{-1/d}$. Under RD, the following result can be proved:

LEMMA 4.1. *Under RD, for every small positive number μ_1, there exists a positive constant $\kappa \geq 1$ depending on d and the design density $p(x)$ only such that*

$$\boldsymbol{P}\left(\overline{h} > \kappa n^{-1/d}\right) \leq \mu_1.$$

The idea of the proof is that a ball $U_h(X_i)$ contains under RD in mean about $C_d h^d p(X_i)$ design points with a fixed constant C_d. Therefore, if $h > \kappa n^{-1/d}$ with $\kappa^d C_d p(X_i) > 2d$ for all or almost all $i \leq n$, then the majority of the balls $U_h(X_i)$ contain at least $d+1$ design points. We omit the details.

Now we bound Δ under ED or RD. Since $s_i^2 = \sum_j c_{ij}^2$, the Cauchy-Schwarz inequality implies

$$\left(\sum_{j\neq i} |c_{ij}|\right)^2 \leq (N_{h_i}(X_i) - 1)\sum_{j\neq i} c_{ij}^2 \leq (N_{h_i}(X_i) - 1)\, s_i^2$$

and hence, if $f \in \mathcal{F}(2, L)$, then $L_i \leq L$ for all i and

$$\Delta^2 \leq \frac{1}{n}\sum_{i=1}^n L_i^2 h_i^4 \,(N_{h_i}(X_i) - 1) \leq \frac{L^2\overline{N}}{n}\sum_{i=1}^n h_i^4 = L^2\overline{h}^4\overline{N} \leq \kappa^2 L^2 n^{-4/d}.$$

Under ED, this inequality applies with $\kappa = 1$. Under RD κ from Lemma 4.1 should be used and the bound holds with a probability at least $1 - \mu_1$.

Next we consider S. Define

$$\begin{aligned} N_i &= \#\{X_j : |X_j - X_i| < h_i + h_j\}, \qquad i = 1, \dots, n, \\ N^* &= \frac{1}{n}\sum_{i=1}^n N_i\,. \end{aligned}$$

One can easily show that under ED the value N^* is bounded by a constant depending on d only. Under RD, a similar bound can be obtained outside

a random set of a small probability μ_2 and the constant N^* would also depend on the design density, cf. Lemma 4.1.

We now intend to show that $S^2 \le 2N^* n^{-1}$. Obviously $\|B\|_\infty = \|A^\top A\|_\infty = \|AA^\top\|_\infty$ and $S^2 = \mathrm{tr}(A^\top A)^2 = \mathrm{tr}(AA^\top)^2$.

The entries $r_{ij} = \sum_{k=1}^n \alpha_{ik}\alpha_{jk}$ of the matrix $AA^\top$ satisfy the conditions $r_{ii} = \sum_{k=1}^n \alpha_{ik}^2 = n^{-1}$ and $r_{ij} \le n^{-1}$. Moreover, if $|X_i - X_j| > h_i + h_j$, then two local linear fits in X_i and in X_j are defined over non overlapping neighborhoods and therefore $r_{ij} = 0$. This implies for every $i \le n$

$$\sum_{j=1}^n r_{ij}^2 \le N_i n^{-2}$$

and hence,

$$S^2 = 2\sum_{i=1}^n \sum_{j=1}^n \beta_{ij}^2 \le 2n^{-2} \sum_{i=1}^n N_i = \frac{2N^*}{n}.$$

4.3. Proof of Theorem 3.2

This result is an easy corollary of Theorem 3.1. Indeed, application of this result and Remark 3.1 with $d \le 8$ and varying γ yields

$$\boldsymbol{P}\left(n^{1/2}\sigma^{-2}(\widehat{\sigma}^2 - \sigma^2) > \Delta^2\sigma^{-2}n^{1/2} + \gamma K\right) \le 2e^{-\gamma^2/4} + e^{-\gamma c}.$$

where $K = \sqrt{n}\,\Delta\sigma^{-1}\sqrt{2\lambda_A} + \sqrt{n}\,S$ and $c = S/(6\lambda_A)$. The conditions of the theorem yield for $d \le 8$ in view of Remark 3.1

$$\Delta^2\sigma^{-2}n^{1/2} \le D, \qquad K \le (2D)^{1/2}(4N^*)^{1/4} + (2N^*)^{1/2}, \qquad c \ge \sqrt{2}/6.$$

Therefore

$$\begin{aligned}
R(\widehat{\sigma}^2) &= \boldsymbol{E}w\left(\sqrt{n}\sigma^{-2}(\widehat{\sigma}^2 - \sigma^2)\right) \\
&\le -\int_0^\infty w(x)\,\mathrm{d}\boldsymbol{P}\left(\sqrt{n}\sigma^{-2}|\widehat{\sigma}^2 - \sigma^2| > x\right) \\
&\le 2w(D) + K\int_0^\infty w'(D + K\gamma)\boldsymbol{P}\left(\sqrt{n}\sigma^{-2}|\widehat{\sigma}^2 - \sigma^2| > D + K\gamma\right)\,\mathrm{d}\gamma \\
&\le 2w(D) + 2K\int_0^\infty w'(D + K\gamma)\left(e^{-\gamma^2/4} + e^{-c\gamma}\right)\,\mathrm{d}\gamma
\end{aligned}$$

and the assertion follows. The case of $d > 8$ can be treated similarly.

4.4. Proof of Theorem 3.3

Let the matrix A with the entries α_{ij} be defined in the proof of Theorem 3.1 and $B = A^\top A$. The difference $\widehat{\sigma}^2 - \sigma^2$ can be represented in the form (see again the proof of Theorem 3.1)

$$\begin{aligned}
\widehat{\sigma}^2 - \sigma^2 &= |b|^2 + 2b^\top A\varepsilon + \varepsilon^\top B\varepsilon - \sigma^2\,\mathrm{tr}\,B \\
&= |b|^2 + 2b^\top A\varepsilon + \sum_{i=1}\beta_{ii}(\varepsilon_i^2 - \sigma^2) + \sum_{i=1}^n\sum_{j\ne i}\beta_{ij}\varepsilon_i\varepsilon_j \\
&= |b|^2 + Q_2 + Q_3 + Q_4.
\end{aligned}$$

We now estimate separately each term in this expression. Note first that $|b|^2 \le \Delta^2$, see the proof of Theorem 3.1.

Let $\widetilde{\varepsilon}_1, \dots, \widetilde{\varepsilon}_n$ be a sequence of i.i.d. random variables from the normal law $\mathcal{N}(0, \sigma^2)$. Define the sums $\widetilde{Q}_2, \widetilde{Q}_3, \widetilde{Q}_4$ similarly to Q_2, Q_3, Q_4 with $\widetilde{\varepsilon}_i$'s in place of ε_i's. The idea is to show that the distribution of every Q_k only weakly depends on the particular distribution of ε_i's and therefore, the bounds for $\widetilde{Q}_k$ are valid for Q_k as well (in some asymptotic sense if n is large enough), $k = 2, 3, 4$.

First we estimate the sum $Q_2 = 2b^\top A\varepsilon$. Note that $\boldsymbol{E}Q_2 = 0$ and

$$\boldsymbol{E}Q_2^2 = \sigma^2 |A^T b|^2 = \sigma^2 b^\top A A^\top b \le \|AA^\top\|_\infty |b|^2 \le \|B\|_\infty \Delta^2.$$

By the Cauchy-Schwarz inequality

$$\boldsymbol{P}\left(|b^\top A\varepsilon| > \Delta S^{1/2}\sigma\right) \le \frac{\boldsymbol{E}Q_2^2}{\Delta^2 S\sigma^2} \le \frac{\|B\|_\infty}{S}$$

and by the conditions of the theorem, $n\|B\|^2/S^2 \le C_A^2$, so that

$$\boldsymbol{P}\left(|Q_2| > 2\Delta S^{1/2}\sigma\right) \le 4C_A n^{-1/2}.$$

Next, it holds for Q_3

$$\boldsymbol{E}Q_3^2 = \boldsymbol{E}\left(\sum_{i=1}^n \beta_{ii}(\varepsilon_i^2 - \sigma^2)\right)^2 = C_4^2\sigma^4 \sum_{i=1}^n \beta_{ii}^2$$

and the Berry-Essen inequality, see Petrov (1975), applied to Q_3 yields with $S'' = \sigma^{-2}\sqrt{\boldsymbol{E}Q_3^2}$

$$\begin{aligned}\boldsymbol{P}\left(Q_3 > xS''\sigma^2\right) &\le \boldsymbol{P}\left(\widetilde{Q}_3 > xS''\sigma^2\right) + \rho\delta^{-3}\frac{1}{S''^3\sigma^6}\sum_{i=1}^n \boldsymbol{E}\left|\beta_{ii}(\varepsilon_i^2 - \sigma^2)\right|^3 \\ &\le \boldsymbol{P}\left(\widetilde{Q}_3 > (x-\delta)S''\sigma^2\right) + C_6\rho\delta^{-3}(S'')^{-3}\sum_{i=1}^n |\beta_{ii}|^3.\end{aligned}$$

The conditions of the theorem provide

$$\sum_{i=1}^n |\beta_{ii}|^3 \le \max_{i=1,\dots,n}\beta_{ii}\sum_{i=1}^n |\beta_{ii}|^2 \le C_A^2 S''^3 n^{-1/2}$$

and hence

$$\boldsymbol{P}\left(Q_3 > xS''\sigma^2\right) \le \boldsymbol{P}\left(\widetilde{Q}_3 > xS''\sigma^2\right) + C_6\rho\delta^{-3}C_A^2 n^{-1/2}.$$

In addition, the use of Lemma 4.3 yields for every γ

$$\boldsymbol{P}\left(\widetilde{Q}_3 > \gamma S''\sigma^2\right) \le e^{-\gamma^2/4} + e^{-\gamma\sqrt{n}/(6C_A)}.$$

For estimating Q_4, we apply the following general result from Spokoiny (1999, Corollary 6.2). Let $U = (u_{ij}, i, j = 1, \dots, n)$ be a $n \times n$ symmetric matrix with $u_{ii} = 0$ for all i. By $U(\varepsilon_1, \dots, \varepsilon_n)$ we denote the corresponding quadratic form of i.i.d. random variables $\varepsilon_1, \dots \varepsilon_n$, that is,

$$U(\varepsilon_1, \dots, \varepsilon_n) = \sum_{i=1}^n \sum_{j \ne i}^n u_{ij}\varepsilon_i\varepsilon_j.$$

Let also $\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_n$ be a sequence of independent Gaussian r.v.'s with $\boldsymbol{E}\widetilde{\varepsilon}_i = 0$ and $\boldsymbol{E}\widetilde{\varepsilon}_i^2 = \sigma^2$, $i = 1, \ldots, n$. Define another quadratic form

$$U(\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_n) = \sum_{i=1}^{n} \sum_{j \neq i} u_{ij} \widetilde{\varepsilon}_i \widetilde{\varepsilon}_j.$$

Clearly $\boldsymbol{E}U(\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_n) = 0$ and $\boldsymbol{E}|U(\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_n)|^2 = \boldsymbol{E}|U(\varepsilon_1, \ldots, \varepsilon_n)|^2$.

PROPOSITION 4.1. *Let $\boldsymbol{E}\varepsilon_i^4 \leq C_4 \sigma^4$ for some fixed constant $C_4 \geq 3$. Let, for a symmetric matrix U with $u_{ii} = 0$ for $i = 1, \ldots, n$, and for a normalizing constant G, the value C_U be defined by*

$$C_U = \max_{i=1,\ldots,n} n G^{-2} \sigma^4 \sum_{j=1}^{n} u_{ij}^2.$$

Then, for each $\delta > 0$ and every x

$$\begin{aligned} \boldsymbol{P}\left(G^{-1} U(\varepsilon_1, \ldots, \varepsilon_n) > x\right) &\leq \boldsymbol{P}\left(G^{-1} U(\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_n) > x - \delta\right) \\ &\quad + \rho (C_4 C_U)^{3/2} n^{-1/2} \delta^{-3} \end{aligned}$$

with an absolute constant ρ.

We now apply this result to Q_4 with $u_{ij} = \beta_{ij}$, $i \neq j$ and

$$G = \sigma^2 \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij}^2 \right)^{1/2}.$$

Since

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij}^2 = \operatorname{tr}(A A^\top)^2 = \operatorname{tr} B^2 = S^2$$

we derive

$$\boldsymbol{P}\left(Q_4 > (\gamma + \delta)\sigma^2 S\right) \leq \boldsymbol{P}\left(\widetilde{Q}_4 > \gamma \sigma^2 S\right) + \rho (C_4 C_A)^{3/2} n^{-1/2} \delta^{-3}.$$

The bound from Lemma 4.3 applied to $\widetilde{Q}_4$ provides for every γ

$$\boldsymbol{P}\left(\widetilde{Q}_4 > \gamma \sigma^2 S\right) \leq e^{-\gamma^2/4} + e^{-\gamma \sqrt{n}/(6 C_A)}.$$

Summing up everything, what we have got so far, leads to the bound

$$\begin{aligned} &\boldsymbol{P}\left(\pm(\widehat{\sigma}^2 - \sigma^2) > \Delta^2 + 2\Delta S^{1/2} \sigma + (\gamma + \delta) S \sigma^2 + \gamma S'' \sigma^2\right) \\ &\leq \boldsymbol{P}\left(|Q_2| > 2\Delta S^{1/2} \sigma\right) + \boldsymbol{P}\left(\pm Q_3 > \gamma S'' \sigma^2\right) + \boldsymbol{P}\left(\pm Q_4 > (\gamma + \delta) S \sigma^2\right) \\ &\leq 2 e^{-\gamma^2/4} + e^{-\gamma \sqrt{n}/(6 C_A)} + C n^{-1/2} \delta^{-3} \end{aligned}$$

where C depends on C_4, C_6 and C_A only.

4.5. Proof of Theorem 3.4

The idea of the proof is as follows. We first change the minimax statement for a Bayes one. For a prior measure π on the set $\mathcal{F}$, define the

corresponding marginal measure $\boldsymbol{P}_{\pi,\sigma^2}$ by

$$\boldsymbol{P}_{\pi,\sigma^2}(A) = \int \boldsymbol{P}_{f,\sigma^2}(A)\,\pi(\mathrm{d}f).$$

We intend to show that there exists a sequence of random functions f_n with prior distributions π_n satisfying $\pi_n(\mathcal{F}_n(2,L)) \to 1$ and such that

$$\boldsymbol{E}_{\pi_n,\sigma^2} w\left(n^{4/d}(\widetilde{\sigma}_n^2 - \sigma^2)\right) = r > 0$$

for n large enough. For the latter, it suffices to show that the measures $\boldsymbol{P}_{\pi_n,\sigma_0^2}$ with $\sigma_0^2 = 1$ and $\boldsymbol{P}_{\pi_n,\sigma_n^2}$ with $\sigma_n^2 = \sigma_0^2 + n^{-4/d}$ are not asymptotically separable.

The priors π_n are selected on the base of the following consideration. We define the values of random functions f_n either identically zero or i.i.d. normally distributed at each design point X_i. If d is sufficiently large and if the variance of this distribution is small enough, then this random function will be with a large probability in the class $\mathcal{F}_n(2,L)$. Then clearly this random function f_n introduce some additional noise in the observations Y_i and we cannot distinguish whether this noise comes from the errors ε_i only (this would be the case when $f_n \equiv 0$) or there is some contribution from the random regression function f_n. More precisely, let $\xi_1, \ldots, \xi_n$ be i.i.d. standard Gaussian r.v.'s and $\delta_n = n^{-2/d}$. We will show that there exist random functions g_n with $g_n(X_i) = \delta_n \xi_i$ and with $\boldsymbol{P}(g_n \in \mathcal{F}_n(2,L)) \to 1$ as $n \to \infty$ for $d \geq 8$. The random functions f_n are constructed as follows. With probability $1/2$, we set $f_n = 0$ and with probability $1/2$, the function f_n coincides with g_n. Then, for $\sigma = \sigma_0$ the marginal distribution of the observations $Y_i = f(X_i) + \sigma\varepsilon_i$ is with probability $1/2$ i.i.d. from $\mathcal{N}(0,\sigma_0^2)$ and with probability $1/2$ i.i.d. from $\mathcal{N}(0,\sigma_n^2)$. Similarly, for $\sigma = \sigma_n$, the marginal distribution of the observations Y_i corresponds with probability $1/2$ an i.i.d. sample from $\mathcal{N}(0,\sigma_n^2)$ and with probability $1/2$ an i.i.d. sample from $\mathcal{N}(0,\sigma_n^2 + n^{-4/d})$. Hence, with a positive probability, these two marginal distributions coincide and therefore any estimate has a non-vanishing risk.

Now we present a formal description. Let $h = n^{-1/d}$. Define for every grid point X_i a function ϕ_i of the form

$$\phi_i(x) = \prod_{\ell=1}^{d} Q\left(\frac{x_\ell - X_{i,\ell}}{h}\right)$$

where Q is a smooth symmetric nonnegative function supported on $[-1,1]$. Clearly all functions ϕ_i have non-overlapping supports and for every i

$$\begin{aligned} |\phi_i(x)| &\leq 1, \\ \left|\frac{\partial \phi_i(x)}{\partial x_\ell}\right| &\leq \frac{\|Q'\|}{h}, \\ \left|\frac{\partial^2 \phi_i(x)}{\partial x_\ell\,\partial x_k}\right| &\leq \frac{\max\{\|Q'\|^2, \|Q''\|\}}{h^2} \end{aligned}$$

so that

$$\|\phi_i''(x)\| \leq \frac{C_Q}{h^2} \tag{5}$$

with $C_Q = \sqrt{d}\max\{\|Q'\|^2, \|Q''\|\}$.

Let also $\{\xi_i,\, i = 1, \dots, n\}$ be a collection of independent standard Gaussian random variables. Define the random function g_n of the form

$$g_n(x) = \delta_n \sum_{i=1}^{n} \xi_i \phi_i(x).$$

Finally, for an independent of g_n Bernoulli random variable ν_n with $\boldsymbol{P}(\nu_n = 0) = \boldsymbol{P}(\nu_n = 1) = 1/2$, define

$$f_n = \nu_n g_n.$$

The property (5) provides for every $i \le n$

$$\sup_{x:|x-X_i|\le n^{-1/d}} \|g_n''(x)\|^2 \le C_Q h^{-4} \delta_n^2 \max_{j:X_j \in U_h(X_i)} \xi_j^2 \le C_Q \sum_{j:X_j \in U_h(X_i)} \xi_j^2$$

and hence, using $N_h(X_i) \le 2d+1$

$$\frac{1}{n} \sum_{i=1}^{n} \sup_{x:|x-X_i|\le n^{-1/d}} \|g_n''(x)\|^2 \le \frac{C_Q}{n} (2d+1) \sum_{i=1}^{n} \xi_i^2$$

so that, for $L^2 > (2d+1)C_Q$, by the law of large numbers,

$$\boldsymbol{P}\left(\frac{1}{n} \sum_{i=1}^{n} \sup_{x:|x-X_i|\le n^{-1/d}} \|g_n''(x)\|^2 > L^2 \right) \to 0, \qquad n \to \infty.$$

This means that the random functions g_n belong to $\mathcal{F}_n(2, L)$ with a probability close to 1 if $L^2 > (2d+1)C_Q$ and clearly the same holds for the f_n's.

Let now $\boldsymbol{P}_\sigma^{(n)}$ denote the product measure in $\mathbb{R}^n$ corresponding to the model $Y_i = \sigma\varepsilon_i$ with i.i.d. standard normal errors ε_i. Then clearly

$$\boldsymbol{P}_{f_n,\sigma_0^2} = \left(\boldsymbol{P}_{\sigma_0}^{(n)} + \boldsymbol{P}_{\sigma_n}^{(n)} \right)/2,$$
$$\boldsymbol{P}_{f_n,\sigma_n^2} = \left(\boldsymbol{P}_{\sigma_n}^{(n)} + \boldsymbol{P}_{s_n}^{(n)} \right)/2$$

with $s_n^2 = \sigma_n^2 + n^{-4/d} = \sigma_0^2 + 2n^{-4/d}$. Next we show that all three sequences of measures $(\boldsymbol{P}_{\sigma_0}^{(n)})$, $(\boldsymbol{P}_{\sigma_n}^{(n)})$ and $(\boldsymbol{P}_{s_n}^{(n)})$ are pairwise asymptotically singular, if $d > 8$. Then the required assertion follows from the next general result.

LEMMA 4.2. *Let three sequences $P_j^{(n)}$, $j = 0, 1, 2,$, of probability measures be pairwise asymptotically singular, that is,*

$$Z_{k,j}^{(n)} = \frac{\mathrm{d}P_k^{(n)}}{\mathrm{d}P_j^{(n)}} \xrightarrow{P_j^{(n)}} 0, \qquad n \to \infty, \qquad k \ne j.$$

Then for any continuous bounded function $u(x)$, it holds

$$H_n = \frac{1}{2} \int u\left(\frac{\mathrm{d}P_0^{(n)} + \mathrm{d}P_1^{(n)}}{\mathrm{d}P_1^{(n)} + \mathrm{d}P_2^{(n)}} \right) \mathrm{d}\left(P_1^{(n)} + P_2^{(n)} \right) \to \frac{u(0) + u(1)}{2}, \qquad (6)$$

that is, the likelihood $\frac{\mathrm{d}P_0^{(n)} + \mathrm{d}P_1^{(n)}}{\mathrm{d}P_1^{(n)} + \mathrm{d}P_2^{(n)}}$ converges weakly to the Bernoulli distribution with parameter 1/2.

Proof. One obviously has

$$2H_n = \int u\left(\frac{Z_{0,1}^{(n)}+1}{Z_{2,1}^{(n)}+1}\right) \mathrm{d}P_1^{(n)} + \int u\left(\frac{Z_{0,2}^{(n)}+Z_{1,2}^{(n)}}{Z_{1,2}^{(n)}+1}\right) \mathrm{d}P_2^{(n)} \to u(1)+u(0)$$

as required. ∎

It remains to check (6) for the sequences $\boldsymbol{P}_\sigma^{(n)}$ with $\sigma \in \{\sigma_0, \sigma_n, s_n\}$. We consider the derivative $Z_{0,1}^{(n)} = \mathrm{d}\boldsymbol{P}_{\sigma_0}^{(n)} / \mathrm{d}\boldsymbol{P}_{\sigma_n}^{(n)}$, the other cases can be treated similarly.

The definition $\sigma_n^2 = \sigma_0^2 + \delta_n^2 = 1 + \delta_n^2$ clearly yields

$$\begin{aligned} L_{0,1}^{(n)} &:= \log \frac{\mathrm{d}\boldsymbol{P}_{\sigma_0}^{(n)}}{\mathrm{d}\boldsymbol{P}_{\sigma_n}^{(n)}} \\ &= n\log(\sigma_n/\sigma_0) - \sum_{i=1}^n \frac{Y_i^2}{2\sigma_0^2} + \sum_{i=1}^n \frac{Y_i^2}{2\sigma_n^2} \\ &= \frac{n}{2}\log\frac{\sigma_0^2+\delta_n^2}{\sigma_0^2} - \sum_{i=1}^n \frac{Y_i^2\delta_n^2}{2\sigma_0^2\sigma_n^2}. \end{aligned}$$

Under the measure $\boldsymbol{P}_{\sigma_n}^{(n)}$, it holds $Y_i = \sigma_n\zeta_i$ with i.i.d. standard normal r.v.'s ζ_i. Therefore

$$\begin{aligned} L_{0,1}^{(n)} &= \frac{n}{2}\log\left(1+\delta_n^2\right) - \frac{\delta_n^2}{2}\sum_{i=1}^n \zeta_i^2 \\ &= \frac{n}{2}\log\left(1+\delta_n^2\right) - \frac{n\delta_n^2}{2} - \frac{\sqrt{n}\delta_n^2}{2}\frac{1}{\sqrt{n}}\sum_{i=1}^n(\zeta_i^2-1) \\ &= \frac{n}{2}\log\left(1+\delta_n^2\right) - \frac{n\delta_n^2}{2} - \frac{\sqrt{n}\delta_n^2}{2}\eta_n \\ &= \frac{\sqrt{n}\delta_n^2}{2}\left(r_n - \eta_n\right) \end{aligned}$$

where the random variables $\eta_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n(\zeta_i^2-1)$ are asymptotically standard normal and

$$\begin{aligned} r_n &= \frac{\sqrt{n}}{\delta_n^2}\log\left(1+\delta_n^2\right) - \sqrt{n} \le -\sqrt{n}\left\{\frac{\delta_n^2}{2} - \frac{\delta_n^4}{3}\right\} \\ &= -\sqrt{n}\left\{\frac{n^{-4/d}}{2} - \frac{n^{-8/d}}{3}\right\} \to -\infty \end{aligned}$$

if $d > 8$. Since also $\sqrt{n}\delta_n^2 = n^{1/2-4/d} \to \infty$, this implies $L_{0,1}^{(n)} \to -\infty$ and hence $Z_{0,1}^{(n)} = \exp L_{0,1}^{(n)} \to 0$ as required.

4.6. Large deviation probability for Gaussian quadratic forms

LEMMA 4.3. *Let $A = (a_{ij},\, i,j = 1,\dots,n)$ be a $n\times n$-matrix. Define the values S_A and λ_A by:*

$$\begin{aligned} S_A^2 &= 2\,\mathrm{tr}(A^\top A)^2 = 2\,\mathrm{tr}(AA^\top)^2, \\ \lambda_A &= \|A^\top A\|_\infty = \|AA^\top\|_\infty. \end{aligned}$$

If $\varepsilon_1, \dots, \xi_n$ *are i.i.d. normal* $\mathcal{N}(0, \sigma^2)$ *r.v.'s, and* $b = (b_1, \dots, b_n)^\top$ *is a deterministic vector in* $\mathbb{R}^d$ *then the quadratic form*

$$Q = \sum_{i=1}^{n} \left(b_i + \sum_{j=1}^{n} a_{ij} \varepsilon_j \right)^2$$

fulfills for every $z \geq 0$ *the condition*

$$\boldsymbol{P}\left(\pm \left(Q - |b|^2 - \sigma^2 \operatorname{tr}(A^\top A) \right) > z\sigma |b| (2\lambda_A)^{1/2} + z\sigma^2 S_A \right) \leq 2e^{-z^2/4} + e^{-zS_A/(6\lambda_A)}.$$

Proof. The standardization by σ^2 allows to reduce the general case to the situation with $\sigma^2 = 1$, which is supposed in what follows. With vector notation the studied expression can be represented as

$$\begin{aligned} Q - |b|^2 - \operatorname{tr}(A^\top A) &= (b + A\varepsilon)^\top (b + A\varepsilon) - \operatorname{tr}(A^\top A) - |b|^2 \\ &= 2b^\top A\varepsilon + \varepsilon^\top A^\top A\varepsilon - \operatorname{tr}(A^\top A) \end{aligned}$$

where ε denotes the vector $(\varepsilon_1, \dots, \varepsilon_n)^\top$. The latter expression can be decomposed into linear and quadratic parts:

$$Q - \operatorname{tr}(A^\top A) - |b|^2 = 2b^\top A\varepsilon + \varepsilon^\top A^\top A\varepsilon - \operatorname{tr}(A^\top A) = Q_1 + Q_2 \qquad (7)$$

with

$$\begin{aligned} Q_1 &= 2b^\top A\varepsilon, \\ Q_2 &= \varepsilon^\top A^\top A\varepsilon - \operatorname{tr}(A^\top A). \end{aligned}$$

The term Q_1 is a linear combination of the r.v.'s ε_i and hence it is a Gaussian r.v. with zero mean and the variance

$$\boldsymbol{E} Q_1^2 = 4\boldsymbol{E} b^\top A\varepsilon\varepsilon^\top A^\top b = 4b^\top AA^\top b \leq 4\lambda_A |b|^2.$$

(Here we have used that $\boldsymbol{E}\varepsilon\varepsilon^\top = \mathbf{1}_n$.) Therefore,

$$\boldsymbol{P}\left(\pm Q_1 > z(2\lambda_A)^{1/2} |b| \right) \leq \exp\left\{ -\frac{z^2 2\lambda_A |b|^2}{2\boldsymbol{E} Q_1^2} \right\} \leq e^{-z^2/4}. \qquad (8)$$

Next we intend to show that

$$\boldsymbol{P}\left(\pm Q_2 > zS_A \right) \leq e^{-z^2/4} + e^{-zS_A/(6\lambda_A)}.$$

The symmetric matrix $A^\top A$ can be decomposed as

$$A^\top A = U^\top \Lambda U,$$

with an orthonormal matrix U (i.e. $U^\top U = \mathbf{1}_n$), and a diagonal matrix Λ, $\Lambda = \operatorname{diag}\{\lambda_1, \dots, \lambda_n\}$. It holds

$$\begin{aligned} \operatorname{tr} A^\top A &= \operatorname{tr} \Lambda = \sum_{i=1}^{n} \lambda_i, \\ S_A^2 &= 2 \operatorname{tr}(A^\top A)^2 = 2 \operatorname{tr} \Lambda^2 = 2 \sum_{i=1}^{n} \lambda_i^2, \\ \lambda_A &= \max\{|\lambda_1|, \dots, |\lambda_n|\}. \end{aligned}$$

Therefore

$$Q_2 = \widetilde{\varepsilon}^\top \Lambda \widetilde{\varepsilon} - \operatorname{tr}\Lambda = \sum_{i=1}^n \lambda_i (\widetilde{\varepsilon}_i^2 - 1),$$

where $\widetilde{\varepsilon} = U\varepsilon$ is also a standard Gaussian vector in $\mathbb{R}^n$. We apply the exponential Tschebyscheff-inequality: for every $\mu \geq 0$

$$\boldsymbol{P}\left(Q_2 > a\right) \leq e^{-\mu a} \boldsymbol{E} e^{\mu Q_2}.$$

This yields

$$\begin{aligned} P(z) &:= \boldsymbol{P}\left(\sum_{i=1}^n \lambda_i (\widetilde{\varepsilon}_i^2 - 1) > z S_A \right) \\ &\leq \exp\left\{-\mu\, z\, S_A\right\} \boldsymbol{E} \exp\left\{ \mu \sum_{i=1}^n \lambda_i (\widetilde{\varepsilon}_i^2 - 1) \right\} \\ &= \exp\left\{ -\mu\, z\, S_A - \mu \sum_{i=1}^n \lambda_i \right\} \boldsymbol{E} \prod_{i=1}^n \exp\left\{\mu\, \lambda_i\, \widetilde{\varepsilon}_i^2\right\}. \end{aligned}$$

Since $\widetilde{\varepsilon}$ are independent standard normal, we obtain

$$\begin{aligned} P(z) &\leq \exp\left\{ -\mu\, z\, S_A - \mu \sum_{i=1}^n \lambda_i \right\} \prod_{i=1}^n \boldsymbol{E} \exp\left\{\mu\, \lambda_i\, \widetilde{\varepsilon}_i^2\right\} \\ &= \exp\left\{ -\mu\, z S_A - \sum_{i=1}^n \left[\mu\lambda_i + \frac{1}{2}\log(1 - 2\mu\, \lambda_i) \right] \right\} \end{aligned} \tag{9}$$

provided that $2\mu\lambda_i < 1$ for all i.

Now we apply the following simple inequality:

$$-\log(1-u) \leq u + u^2, \qquad \forall u \leq 2/3.$$

This yields with any $\mu \leq 1/(3\lambda_A)$ and all i:

$$-\mu\lambda_i - \frac{1}{2}\log(1 - 2\mu\, \lambda_i) \leq 2\mu^2 \lambda_i^2$$

and

$$\begin{aligned} -\mu z S_A - \sum_{i=1}^n \left(\mu\lambda_i + \frac{1}{2}\log(1 - 2\mu\, \lambda_i) \right) &\leq -\mu z S_A - \sum_{i=1}^n 2\mu^2 \lambda_i^2 \\ &= -\mu z S_A + \mu^2 S_A^2. \end{aligned} \tag{10}$$

If $z \leq \frac{2S_A}{3\lambda_A}$, then we select $\mu = \frac{z}{2S_A}$. With this choice the condition $\mu \leq 1/(3\lambda_A)$ is fulfilled and

$$-\mu z S_A + \mu^2 S_A^2 = -z^2/4.$$

For $z > \frac{2S_A}{3\lambda_A}$ we set $\mu = 1/(3\lambda_A)$, so that

$$-\mu z S_A + \mu^2 S_A^2 = -\frac{z S_A}{3\lambda_A} + \frac{S_A^2}{(3\lambda_A)^2} = -\frac{z S_A}{3\lambda_A}\left(z - \frac{S_A}{3\lambda_A} \right) \leq -\frac{z S_A}{6\lambda_A}.$$

It now follows from (9) and (10)

$$P(z) \le e^{-z^2/4} + e^{-zS_A/(6\lambda_A)}$$

as desired. Similarly one can bound the probability

$$P'(z) = \boldsymbol{P}\left(\sum_{i=1}^{n} \lambda_i(\tilde{\varepsilon}_i^2 - 1) < -zS_A\right)$$

and the assertion follows in view of (7) and (8). ▌

REFERENCES

1. Buckley, M.J. and Eagleson, G.K. and Silverman, B.W. (1988). The estimation of residual variance in nonparametric regression. Biometrika **75**, No.2, 189–199.
2. Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall, London.
3. Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85** no. 3 645–660.
4. Gasser, T. and Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.
5. Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing. *J. Am. Statist. Assoc.* **86** 643–652.
6. Hall, P. and Carroll, R.J. (1989). Variance function estimation in regression: the effect of estimation of the mean. *J. R. Stat. Soc.* **B 51** 3–14.
7. Hall, P., Kay, J.W. and Titterington, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528.
8. Hall, P., Kay, J.W. and Titterington, D.M. (1991). On estimation of noise variance in two-dimensional signal processing. *Adv. Appl. Probab.* **23** 476–495.
9. Hall, P. and Marron, J.S. (1990). On variance estimation in nonparametric regression. Biometrika **77**, No.2, 415–419.
10. Härdle, W. and Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometr.* **81** 233–242.
11. Hart, J. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests* New York, Berlin, Heidelberg: Springer.
12. Katkovnik, V. Ja. (1985). *Nonparametric Identification and Data Smoothing: Local Approximation Approach*. Nauka, Moscow (in Russian).
13. Müller, H.G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15** 610–625.
14. Müller, H.G. and Stadtmüller, U. (1993). On variance function estimation with quadratic forms. *J. Statist. Plan. Inf.* **35** 213–231.
15. Neumann, M.H. (1994). Fully data-driven nonparametric variance estimation. *Statistics* **25** 189–212.
16. Petrov, V.V. (1975). *Sums of Independent Random Variables*. Springer, New York.
17. Rice, J., (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics* **12** 1215–1230.
18. Ruppert, D. and Wand, M.P. , Holst, U. and Hössjer, O. (1997). Local polynomial variance function estimation. *Technometrics* **39** 262–273.
19. Seifert, B., Gasser, T. and Wolf, A. (1993). Nonparametric estimation of residual variance revisited. *Biometrika* **80** no. 2, 373–383.
20. Silverman, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. R. Statist. Soc.* **B 47** 1–52.
21. Spokoiny, V. (1999). Data driven testing the fit of linear models. Preprint **472**, Weierstrass Institute, Berlin. `http://www.wias-berlin.de`.
22. Tsybakov, A. (1986). Robust reconstruction of functions by the local approximation. *Prob. Inf. Transm.*, **22**, 133–146.
23. Wahba. G. (1983). Bayesian "confidence interval" for the cross-validated smoothing spline. *J. R. Statist. Soc.* **B 45** 133–150.