Electronic Journal of Statistics
Vol. 6 (2012) 861–907
ISSN: 1935-7524
DOI: 10.1214/12-EJS693

Spatial adaptation in heteroscedastic regression: Propagation approach*

Nora Serdyukova

Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Goldschmidtstr. 7, 37077 Göttingen, Germany
e-mail: nora.serdyukova@mathematik.uni-goettingen.de

Abstract: The paper concerns the problem of pointwise adaptive estimation in regression when the noise is heteroscedastic and incorrectly known. The use of the local approximation method, which includes the local polynomial smoothing as a particular case, leads to a finite family of estimators corresponding to different degrees of smoothing. Data-driven choice of localization degree in this case can be understood as the problem of selection from this family. This task can be performed by a suggested in Katkovnik and Spokoiny (2008) FLL technique based on Lepski's method. An important issue with this type of procedures – the choice of certain tuning parameters – was addressed in Spokoiny and Vial (2009). The authors called their approach to the parameter calibration "propagation". In the present paper the propagation approach is developed and justified for the heteroscedastic case in presence of the noise misspecification. Our analysis shows that the adaptive procedure allows a misspecification of the covariance matrix with a relative error of order $(\log n)^{-1}$, where n is the sample size.

AMS 2000 subject classifications: Primary 62G05; secondary 62G08.
Keywords and phrases: Adaptive estimation, heteroscedastic data, nonparametric regression, Lepski's method, minimax rate of convergence, model misspecification, nonparametric regression, oracle inequalities, propagation.

Received May 2011.

Contents

*Funding of the DFG FOR 916 and ANR-07-BLAN-0234 is acknowledged.
†The author wishes to thank the Associate Editor and unknown referee for many fruitful questions and comments that greatly improved the paper, as well as her supervisor Professor Vladimir Spokoiny for introduction to the astonishing world of adaptive estimation.

1. Introduction

Consider a regression model

$$\boldsymbol{Y} = \boldsymbol{f} + \Sigma_0^{1/2}\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n) \tag{1.1}$$

with response vector $\boldsymbol{Y} \in \mathbb{R}^n$ and unknown diagonal covariance matrix $\Sigma_0 = \operatorname{diag}(\sigma_{0,1}^2, \ldots, \sigma_{0,n}^2)$. Let $\mathcal{X}$ be a Borel subset of $\mathbb{R}^n$ and X_i be fixed elements of $\mathcal{X}$. Denote by $f : \mathcal{X} \to \mathbb{R}$ the unknown regression function, then with $\boldsymbol{f} = (f(X_1), \ldots, f(X_n))^\top$ model (1.1) can be written as

$$Y_i = f(X_i) + \sigma_{0,i}\,\varepsilon_i, \quad i = 1, \ldots, n. \tag{1.2}$$

Given a point $x \in \mathcal{X}$, the target of estimation is the value of $f(x)$. The idea is to replace model (1.2) by a local parametric model

$$y_i = f_{\boldsymbol{\theta}}(X_i) + \sigma_i\,\epsilon_i, \quad i : X_i \in U_h(x), \tag{1.3}$$

where $\sigma_i > 0$ are known, $U_h(x) \stackrel{\text{def}}{=} \{t : \|t - x\| \le h/2\}$ and $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ is an unknown parameter to be estimated. Denote by $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_n)$ a $p \times n$ design matrix. In the considered set-up the covariance matrix Σ_0 is not known exactly and $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ stands for the available covariance matrix. Then the approximate model used instead of the true one reads as follows:

$$\boldsymbol{y} = \boldsymbol{\Psi}^\top\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\varepsilon}. \tag{1.4}$$

Employing inside of $U_h(x)$ one of the well-developed parametric methods we can estimate $\boldsymbol{\theta}$ by $\widetilde{\boldsymbol{\theta}}(y_1, \ldots, y_d; x)$ and then use the estimator $f_{\widetilde{\boldsymbol{\theta}}(Y_1,\ldots,Y_d)}(x)$ based on the observations from the "true" model (1.2) for estimation of $f(x)$. Therefore we have to choose the local model (correspondingly, the collection of estimators of $f_{\boldsymbol{\theta}}(\cdot), \boldsymbol{\theta} \in \Theta$) and the appropriate degree of locality h. This method of local approximation originated from [37, 8, 19, 38, 20, 21, 40, 22].

In what follows we consider approximation by local linear models of the type:

$$y_i = \Psi_i\boldsymbol{\theta} + \sigma_i\,\epsilon_i, \quad i : X_i \in U_h(x), \tag{1.5}$$

where $\Psi_i = \Psi(X_i) = (\psi_1(X_i - x), \dots, \psi_p(X_i - x))^\top$ is a vector of basis functions $\{\psi_j(\cdot)\}$ which already are fixed. Thus the model is misspecified in two places: in the form of the regression function and in the error distribution. The main issue then is to choose the appropriate bandwidth h such that the estimator

$$f_{\widetilde{\boldsymbol{\theta}}_h}(x) \stackrel{\text{def}}{=} \sum_{j=1}^{p} \widetilde{\theta}_h^{(j)}(x)\psi_j(0) \tag{1.6}$$

built on the base of localized data would provide a relevant estimator for $f(x)$. For this purposes the bandwidths selection should be done in a data-driven way, and this problem can be formulated as adaptive selection from the finite family $\{f_{\widetilde{\boldsymbol{\theta}}_h}(x)\}_{h>0}$. Notice also that the coefficients $\theta^{(j)}(x)$ as well as their estimators depend on x and should be calculated for every particular point of interest x. On the other side the localization reduces influence of the choice of the functions $\{\psi_j(\cdot)\}$ allowing to use simple collections.

The proposed approach includes the important class of polynomial regressions, see [11, 22, 31, 41]. For example in the univariate case $x \in \mathbb{R}$, due to the Taylor theorem, the approximation of the unknown function $f(t)$ for t close to x can be written in the following form: $f_{\boldsymbol{\theta}}(t) = \theta^{(0)} + \theta^{(1)}(t-x) + \cdots + \theta^{(p-1)}(t-x)^{p-1}/(p-1)!$ with the parameter $\boldsymbol{\theta} = (\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(p-1)})^\top$ corresponding to the values of $f(\cdot)$ and its derivatives at the point x, if they exist. The design matrix $\boldsymbol{\Psi}$ then consists of the columns

$$\Psi_i = \bigl(1,\, X_i - x, \dots, (X_i - x)^{p-1}/(p-1)!\bigr)^\top \ , \quad i = 1, \dots, n,$$

and corresponds to the well known polynomial smoothing. If the regression function is sufficiently smooth, then, up to a reminder term, for any t close to x, $f(t) \approx f_{\boldsymbol{\theta}}(t)$ and the estimator of $f(x)$ at x is given by $\widetilde{f}(x) = f_{\widetilde{\boldsymbol{\theta}}(x)}(x) = \widetilde{\theta}^{(0)}$. The local constant fit at a given point $x \in \mathbb{R}$ is covered as well with $p = 1$. In this case the "design" matrix is a row $\boldsymbol{\Psi} = (1, \dots, 1)$ and $f_{\boldsymbol{\theta}}(X_i) = \boldsymbol{\Psi}_i^\top \boldsymbol{\theta} = \theta^{(0)} = f_{\boldsymbol{\theta}}(x)$, $i = 1, \dots, n$. This type of approximation in our set-up with known constant noise is treated in [23] and [36].

Nonparametric estimation in heteroscedastic regression under the L_2 losses was studied in [17, 18] and series of papers [13, 14]. One should mention very interesting paper [9] on aggregation estimation under empirical losses in heteroscedastic Gaussian regression. For estimation of the mean with L_2-risk in Gaussian homoscedastic model with unknown variance the penalties allowing to deal with the complexity of such a collection of models were proposed in [4]. However the problem of "local model selection" addressed in the present paper is quite different to the model selection in the sense of [5] and [32] related to estimation with global risk. In this set-up an amazing progress is achieved for the model selection in heteroscedastic not necessary Gaussian regression model in [2, 3, 34]. The minimax pointwise estimation in heteroscedastic regression is in focus of [7].

2. Estimation procedure

2.1. Local parametric estimation

Using the conceptual framework given in the introduction we choose the maximum likelihood estimation as a parametric method used inside of a smoothing window. Let us briefly recall the idea of the local likelihood method dating back to [6] and [39].

If the response variables Y_i are independent and have a density $v(y, s(X_i))$, then the joint log-density of the sample is given by $\mathrm{L}(s) = \sum_{i=1}^{n} \log v(Y_i, s(X_i))$ leading to the "global" maximum likelihood estimation. Let as before $f_{\boldsymbol{\theta}}(\cdot)$ be a function entirely described by a vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$. The local likelihood model does not assume that $s(X_i) = f_{\boldsymbol{\theta}}(X_i)$, but one fits the "parametric" model locally within the smoothing window described by weights $\mathcal{W}(x) = \{w_i(x)\}_{i=1}^{n}$. The *local log-likelihood* is defined as

$$\mathrm{L}(\mathcal{W}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log v(Y_i, f_{\boldsymbol{\theta}}(X_i)) w_i(x). \tag{2.1}$$

The local likelihood estimator $\widetilde{\boldsymbol{\theta}}(x)$ is a maximizer of this weighted sum, $\widetilde{\boldsymbol{\theta}}(x) = \operatorname{argmax}_{\boldsymbol{\theta}} \mathrm{L}(\mathcal{W}, \boldsymbol{\theta})$. It is worth pointing out that in spite of the term "local likelihood" seems to be standard, see [31] for example, if the weights w_i are allowed to take values different from zero and one, the quantity defined by (2.1) is not a log-likelihood in the probabilistic sense even if the data indeed locally follows the parametric model with $v(Y_i, f_{\boldsymbol{\theta}}(X_i))$ for all $i : w_i(x) > 0$. However, the local, or more correctly, weighted log-likelihood inherits most of useful properties from its "global" counterpart, c.f. Proposition 4.3. And – what is of particular importance – the true value of the parameter $\boldsymbol{\theta}$ maximizes the expectation of (2.1), see [31] p.72. This property in more general set-up leads to the minimum contrast $-\mathrm{L}(\mathcal{W}, \boldsymbol{\theta})$ estimation.

Leaving the computational aspects aside, the key issue of this method is a proper choice of the largest smoothing window where the parametric fit $f_{\boldsymbol{\theta}}$ is still adequate. Putting differently, if we consider a finite collection of smoothing windows and corresponding (quasi) MLE's, the target is a data-driven selection from this family. In what follows we explore this approach.

Fix a point $x \in \mathbb{R}^d$ as a center of localization and basis $\{\psi_j\}$. Denote by

$$\Psi_i = \Psi(X_i - x) = (\psi_1(X_i - x), \ldots, \psi_p(X_i - x))^\top, \quad i = 1, \ldots, n,$$

the vectors of basis functions centered at x. For the next nonparametric "selection" step we need a sequence of nested windows. Let for every x a finite sequence of scales $\mathcal{W}_k(x)$, $k = 1, \ldots, K$, be given by matrices

$$\mathcal{W}_k(x) = \operatorname{diag}(w_{k,1}(x), \ldots, w_{k,n}(x)),$$

where the weights $w_{k,i}(x) \in [0, 1]$ can be understood, for instance, as smoothing kernels $w_{k,i}(x) = W((X_i - x)h_k^{-1})$. A particular localizing function $w_{(\cdot,\cdot)}(x)$ is

assumed to be fixed; the aim is to choose on the base of available data an index k of an "optimal" scale. To simplify the notation we sometimes suppress the dependence on the reference point x. Denote by

$$\mathbf{W}_k \stackrel{\mathrm{def}}{=} \Sigma^{-1/2}\mathcal{W}_k\Sigma^{-1/2} = \mathrm{diag}\left(\frac{w_{k,1}}{\sigma_1^2},\dots,\frac{w_{k,n}}{\sigma_n^2}\right), \quad k=1,\dots,K. \tag{2.2}$$

Let Θ be a compact subset of $\mathbb{R}^p$. Inside of any window given by $\mathcal{W}_k$, $k=1,\dots K$, according to (2.1) for each k we calculate the (quasi) MLE $\widetilde{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_k(x) = (\widetilde{\theta}_k^{(0)}(x),\dots,\widetilde{\theta}_k^{(p-1)}(x))^\top$ of $\boldsymbol{\theta}$:

$$\widetilde{\boldsymbol{\theta}}_k \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathrm{L}(\mathbf{W}_k,\boldsymbol{\theta}), \tag{2.3}$$

where $\mathrm{L}(\mathbf{W}_k,\boldsymbol{\theta})$ is the weighted log-likelihood corresponding to the joint distribution of independent sample with $Y_i \sim \mathcal{N}\left(\Psi_i^\top\boldsymbol{\theta},\sigma_i^2\right)$:

$$\begin{aligned}\mathrm{L}(\mathbf{W}_k,\boldsymbol{\theta}) &= -\frac{1}{2}\sum_{i=1}^n |Y_i-\Psi_i^\top\boldsymbol{\theta}|^2\frac{w_{k,i}}{\sigma_i^2}+R \\ &= -\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{\Psi}^\top\boldsymbol{\theta})^\top\mathbf{W}_k(\boldsymbol{Y}-\boldsymbol{\Psi}^\top\boldsymbol{\theta})+R.\end{aligned} \tag{2.4}$$

Here R stands for the terms independent of $\boldsymbol{\theta}$. If the $p\times p$ matrix $\mathbf{B}_k = \mathbf{B}_k(x)$ given by

$$\mathbf{B}_k \stackrel{\mathrm{def}}{=} \boldsymbol{\Psi}\mathbf{W}_k\boldsymbol{\Psi}^\top = \sum_{i=1}^n \Psi_i\Psi_i^\top\frac{w_{k,i}}{\sigma_i^2} \tag{2.5}$$

is positive definite at the point x, $\mathbf{B}_k(x)\succ 0$, then $\widetilde{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_k(x)$ given by

$$\widetilde{\boldsymbol{\theta}}_k = \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\boldsymbol{Y} = \mathbf{B}_k^{-1}\sum_{i=1}^n \Psi_i Y_i\frac{w_{k,i}}{\sigma_i^2} \tag{2.6}$$

is a linear estimator. Recall that in the case of polynomial basis $\{t^q,\, q=0,\dots,p-1\}$ for every fixed k the first coordinate of $\widetilde{\boldsymbol{\theta}}_k(x)$ is the local polynomial estimator for the value of $f(x)$.

In what follows we assume that $n>p$ and $\det\mathbf{B}_k(x)>0$ for any $k=1,\dots,K$. One needs to keep in mind that for example, if $w_{1,\cdot} = W((\cdot-x)h_1^{-1})$ is a finitely supported kernel function, one can always find a bandwidth h_1 so small that the matrix $\mathbf{B}_1(x)$ is degenerated. This implies that the smallest value of h_1 should be chosen in order to guarantee $\mathbf{B}_1(x)\succ 0$. More precisely we assume the following:

(*A*1) *The $p\times n$ matrix $\boldsymbol{\Psi}\mathcal{W}_1(x)$ is of full row rank, that is its rows are linearly independent as the Euclidean vectors.*

Remark 2.1. In view of Assumption $(A2)$ below in Section 2.2, it is sufficient to formulate this assumption only for $k=1$, the positive definiteness of other $\mathbf{B}_k$'s follows automatically.

Remark 2.2. The empirical semi-norm of a function $g(\cdot)$ given by $\|g\|_n^2 = n^{-1}\sum_{i=1}^n g^2(X_i)$, $X_i \in \mathcal{X}$, is generated by the "empirical" scalar product associating the scalar product in $\mathbb{R}^n$: $\langle g, f\rangle_n = n^{-1}\sum_{i=1}^n g(X_i)f(X_i)$ with the functions $g(\cdot)$ and $f(\cdot)$. Given a weight function $s(\cdot) > 0$ one can define in a similar way a weighted empirical scalar product

$$\langle g, f\rangle_{n,s} = n^{-1}\sum_{i=1}^n g(X_i)f(X_i)s(X_i)$$

and the corresponding weighted empirical semi-norm. Thus we see that given $\sigma(\cdot) > 0$ and a collection of functions $\{w_{k,\cdot}(x),\, k = 1,\dots,K\}$, the matrices $n^{-1}\mathbf{B}_k(x)$ are the Gram matrices of the localized basis functions $\psi_1,\dots,\psi_p$ centered at x. That is for any $k = 1,\dots,K$ we have

$$n^{-1}\mathbf{B}_k(x) = (\ \langle \psi_\nu(\cdot - x)\sqrt{w_{k,\cdot}}\,,\ \psi_\eta(\cdot - x)\sqrt{w_{k,\cdot}}\rangle_{n,\sigma}\ )_{1\le\nu\le\eta\le p},$$

where

$$\langle\ g(\cdot - x)\sqrt{w_{k,\cdot}}\,,\ f(\cdot - x)\sqrt{w_{k,\cdot}}\ \rangle_{n,\sigma} = n^{-1}\sum_{i=1}^n g(X_i - x)f(X_i - x)w_{k,i}(x)\sigma_i^{-2}$$

with $\sigma_i = \sigma(X_i) > 0$ and $w_{k,i}(x) = w_{k,X_i}(x)$. It is well known that any Gram matrix is non-negative definite. Correspondingly, $\mathbf{B}_k \succ 0$ if and only if the rows of $\mathbf{\Psi}\mathcal{W}_1(x)^{1/2}$ are linearly independent. In view of $(A2)$ it is sufficient to formulate this assumption only for $k = 1$. We require slightly more: the rows of $\mathbf{\Psi}\mathcal{W}_1(x)$ to be independent. This guarantees that all the variances $\mathrm{Var}[\widetilde{\boldsymbol{\theta}}_k]$ are non-degenerated. Indeed, from (3.1) below we have $\mathrm{Var}[\widetilde{\boldsymbol{\theta}}_k] = \mathbf{B}_k^{-1}\widetilde{\mathbf{B}_k}\mathbf{B}_k^{-1}$, where $\widetilde{\mathbf{B}_k}(x) = \mathbf{\Psi}\mathbf{W}_k\Sigma_0\mathbf{W}_k\mathbf{\Psi}^\top$ is a Gram matrix of the same as $\mathbf{B}_k(x)$ type, but generated by the scalar products $\langle g(\cdot - x)w_{k,\cdot}\,, f(\cdot - x)w_{k,\cdot}\rangle_{n,\sigma}$

The formulas in (2.6) give a sequence of estimators $\{\widetilde{\boldsymbol{\theta}}_k(x),\, k = 1,\dots,K\}$. It was noticed in [1] that in the case when the true data distribution is unknown the QMLE is a natural estimator for the parameter maximizing the expected log-likelihood. That is for every $k = 1,\dots,K$, the estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ can be considered as an estimator of

$$\begin{aligned}\boldsymbol{\theta}_k^*(x) &\stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta}\mathbb{E}\,\mathrm{L}\,(\mathbf{W}_k,\boldsymbol{\theta}) &(2.7)\\ &= \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta}(\boldsymbol{f} - \mathbf{\Psi}^\top\boldsymbol{\theta})^\top\mathbf{W}_k(\boldsymbol{f} - \mathbf{\Psi}^\top\boldsymbol{\theta})\\ &= \mathbf{B}_k^{-1}\mathbf{\Psi}\mathbf{W}_k\boldsymbol{f} = \mathbf{B}_k^{-1}\sum_{i=1}^n \Psi_i f(X_i)\frac{w_{k,i}}{\sigma_i^2}. &(2.8)\end{aligned}$$

Recall that we do not assume $\boldsymbol{f} = \mathbf{\Psi}^\top\boldsymbol{\theta}$ even locally. It is known from [42] that in the presence of a model misspecification for every k the QMLE $\widetilde{\boldsymbol{\theta}}_k$ is a strongly consistent estimator for $\boldsymbol{\theta}_k^*(x)$, which also is the minimizer of the

weighted Kullback-Leibler [26] information criterion:

$$\begin{aligned}\boldsymbol{\theta}_k^*(x) &= \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^n \mathbb{KL}\left(\mathcal{N}\left(f(X_i),\sigma_i\right),\mathcal{N}\left(\Psi_i^\top\boldsymbol{\theta},\sigma_i\right)\right) w_{k,i}(x)\\ &= \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^n |f(X_i)-\Psi_i^\top\boldsymbol{\theta}|^2\frac{w_{k,i}(x)}{\sigma_i^2}\end{aligned}$$

with $\mathbb{KL}(P,P_{\boldsymbol{\theta}}) \stackrel{\text{def}}{=} \mathbb{E}_P\big[\log\left(\mathrm{d}P/\mathrm{d}P_{\boldsymbol{\theta}}\right)\big]$. For properties of the Kullback-Leibler divergence see, for example, [41].

It follows from the above definition of $\boldsymbol{\theta}_k^*(x)$ and from (2.3) that the QMLE $\widetilde{\boldsymbol{\theta}}_k$ admits a decomposition into deterministic and stochastic parts:

$$\widetilde{\boldsymbol{\theta}}_k = \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k(\boldsymbol{f}+\Sigma_0^{1/2}\boldsymbol{\varepsilon}) = \boldsymbol{\theta}_k^* + \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0^{1/2}\boldsymbol{\varepsilon} \tag{2.9}$$

$$\mathbb{E}\widetilde{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k^*, \tag{2.10}$$

where $\boldsymbol{\varepsilon}\sim\mathcal{N}(0,I_n)$. Notice that if $\boldsymbol{f}\equiv\boldsymbol{\Psi}^\top\boldsymbol{\theta}$, then $\boldsymbol{\theta}_k^*\equiv\boldsymbol{\theta}$ for any k, and the classical parametric set-up takes place.

2.2. Adaptive bandwidth selection

Let a point $x\in\mathcal{X}\subset\mathbb{R}^n$, basis $\{\psi_j\}$ and method of localization $w_{(\cdot,\cdot)}(x)$ be fixed. The crucial assumption for the procedure under consideration to work is that the localizing schemes (scales) $\mathcal{W}_k(x)=\operatorname{diag}(w_{k,1}(x),\ldots,w_{k,n}(x))$ are nested, see Remark 2.6. We say that the localizing schemes are nested if for the corresponding matrices the following *ordering condition* is fulfilled:

(A2) *For any fixed x and the method of localization with $w_{(\cdot,\cdot)}(x)\geq 0$ the following relation holds:*

$$\mathcal{W}_1(x)\leq\ldots\leq\mathcal{W}_k(x)\leq\ldots\leq\mathcal{W}_K(x).$$

The inequalities are understood componentwise: for $1\leq l\leq k\leq K$ $\mathcal{W}_l(x)\leq\mathcal{W}_k(x)\Leftrightarrow w_{k,i}(x)-w_{l,i}(x)\geq 0$ for all $i=1,\ldots,n$. For the kernel smoothing this condition means the following. Given a sequence of bandwidths $0<h_1<\ldots<h_k<\ldots<h_K\leq 1$ let $w_{k,i}(x)=W((X_i-x)h_k^{-1})\in[0,1]$ be such that $W(u/h_l)\leq W(u/h_k)$ for any $0<h_l<h_k<1$, and $W(u)\to 0$ as $\|u\|\to\infty$, or even is compactly supported. Also it is intrinsically assumed that, starting from the smallest window, at every step of the procedure every new window contains at least p new design points.

Given the point $x\in\mathcal{X}$, basis $\{\psi_j\}$ and method of localization $w_{(\cdot,\cdot)}(x)$, we look for the estimator $f_{\widehat{\theta}}(x)$ of $f(x)$ having form (1.6), where the coefficients $\widehat{\theta}^{(j)}(x)$ are the components of the estimator

$$\widehat{\boldsymbol{\theta}}(x)\stackrel{\text{def}}{=}\widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)=(\widetilde{\theta}_{\widehat{k}}^{(1)}(x),\ldots,\widetilde{\theta}_{\widehat{k}}^{(p)}(x))^\top, \tag{2.11}$$

corresponding to the adaptive choice of the index $\widehat{k} \in \{1, \ldots, K\}$, i.e. to the choice of the scale. One should keep in mind that $\widehat{k}$ is a random variable taking values in $\{1, \ldots, K\}$.

The selection of $\widehat{\boldsymbol{\theta}}(x)$ from $\{\widetilde{\boldsymbol{\theta}}_k(x)\}$, $k = 1, \ldots, K$, can be done by application of the Lepski [27] method to comparing of the maximized log-likelihoods $\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k)$. This is the idea of the *fitted local likelihood* (FLL) technique suggested in [23]. More precisely, to describe the test statistic, define for any $\boldsymbol{\theta}$, $\boldsymbol{\theta}' \in \Theta$ the corresponding log-likelihood ratio:

$$\mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}, \boldsymbol{\theta}') \stackrel{\mathrm{def}}{=} \mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}) - \mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}'), \tag{2.12}$$

with $\mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta})$ defined by (2.4).

For every $l = 1, \ldots, K$, the "fitted" log-likelihood ratio is defined as follows:

$$\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \boldsymbol{\theta}') \stackrel{\mathrm{def}}{=} \max_{\boldsymbol{\theta} \in \Theta} \mathrm{L}(\mathbf{W}_l, \boldsymbol{\theta}, \boldsymbol{\theta}').$$

By Lemma 4.2, for any scale index l and parameter vector $\boldsymbol{\theta}$ this quantity is quadratic in $\boldsymbol{\theta}$:

$$2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \boldsymbol{\theta}) = (\widetilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta})^\top \mathbf{B}_l (\widetilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}).$$

This prompts, see Remark 2.6, to use the *FLL-statistics*:

$$\begin{aligned} T_{lk} &\stackrel{\mathrm{def}}{=} 2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k) \\ &= (\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_k)^\top \mathbf{B}_l (\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_k)\,, \quad l < k. \end{aligned} \tag{2.13}$$

In the algorithm (2.14) the scale corresponding to $k = 1$ is assumed to provide $\mathbf{B}_1 \succ 0$ and to be sufficiently small assuring nonsignificant deviation of the parametric fit from the true model and $k = 1$ is always accepted. Then the adaptive index $\widehat{k}$ is selected by Lepski's selection rule with the statistics $\{T_{lm}\}$:

$$\widehat{k} = \max\{k \le K : T_{lm} \le \mathfrak{z}_l,\ 1 \le l < m \le k\}\,. \tag{2.14}$$

Finally put $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{k}}$.

The procedure (2.14) involves parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$. As in the classical Lepski procedure, c.f. [27] and [29], the inequalities in (2.14) control the risk of estimators for the case of dominating bias. The opposite case of the negligible w.r.t. the noise bias can be easily controlled in view of the Wilks-type result of Proposition 4.3, c.f. Corollary 4.4 and Remark 2.6:

$$\mathbb{E}|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)|^r \le C(p, r) \tag{2.15}$$

with the constant $C(p, r)$ explicitly given by (4.8) in Appendix.

Let $\widehat{\boldsymbol{\theta}}_k$ denote the last accepted estimate after the first k steps of the procedure:

$$\widehat{\boldsymbol{\theta}}_k \stackrel{\mathrm{def}}{=} \widetilde{\boldsymbol{\theta}}_{\min\{k, \widehat{k}\}}. \tag{2.16}$$

Suppose at this step that the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ have being fixed satisfying the following set of $K - 1$ conditions:

Definition 2.1 (Propagation conditions (PC))**.** Let for a given $\alpha \in (0,1]$ and $r > 0$ the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ satisfy

$$\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r \le \alpha C(p,r) \quad \text{for all } \ k = 2, \ldots, K, \tag{2.17}$$

where $C(p,r)$ is defined by (4.8) and $\mathbb{E}_{0,\Sigma}$ stands for the expectation w.r.t. the measure $\mathcal{N}(0,\Sigma)$.

Remark 2.3. *"True" value of* $\boldsymbol{\theta}$. Lemma 4.1 from Section 4 shows that in the "no bias" situation the Gaussian distribution provides a nice pivotality property: the actual value of the parameter $\boldsymbol{\theta}$ is not important for the risk of adaptive estimate, so one can put $\boldsymbol{\theta} = 0$ in (2.17).

Remark 2.4. *Calculation of the thresholds.* Clearly at any step $k \le K$ of the algorithm the "current value" of the adaptive estimator $\widetilde{\boldsymbol{\theta}}_k$ depends on the thresholds $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$. The theoretical aspects related to the heteroscedasticity of model and incorrectly known variance is the focus of the present paper. Thus we do not detail the practical aspects of the thresholds calibration only mentioning that in practice this can be done by Monte Carlo simulations under the known "parametric" measure $\mathcal{N}(0,\Sigma)$. Moreover one needs to calculate them only once. For detailed consideration of the practical aspects of the calibration as well as for the computational results see [36] or [23] focused on the image denoising by local constant fitting, where the similar idea was proposed. Demo-versions of the software are available on the web page http://www.cs.tut.fi/˜lasip/.

Remark 2.5. *Loss power* r *and "confidence" level* α. The choice of the parameters α and r is free and depends only on desired accuracy results and procedure performance. The basic oracle result of Theorem 3.3 is formulated in terms of polynomial loss function with index $r/2$. Therefore the choice of r in the PC's determines the final risk bounds. The constant α appears in the second order term of the bound.

A detailed explanation of the heuristics behind the PC's and the role of the parameters r and α from the hypothesis testing point of view is given in [36], pp. 2789-2790. Below in Remark 2.6 we present other heuristics for the procedure and PC's, also explaining why $\alpha \le 1$. Here we just mention that the result of Proposition 3.1 shows that up to the constants the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ are of the form $\mathfrak{z}_k = C_1 r(K-k) + C_2 \log(K/\alpha) + C_3$. Therefore the high value of r along with small α enlarge $\mathfrak{z}_k$'s and make the procedure less sensitive to deviations of the parametric fit from the true model resulting in acceptance of a larger smoothing window. Small r and α close to one may result in a less stable performance of the procedure and undersmoothing. The free choice of these parameters allows a practical adjustment of the procedure to a particular data set.

Remark 2.6. *Some heuristics behind the procedure.* Let us give an explanation in the spirit of the example with two Hölder classes (naturally nested w.r.t. the smoothness parameters!) from [27], p.2. Let we have only two scales $\mathcal{W}_1(x) \le \mathcal{W}_2(x)$ and, correspondingly, two MLE estimators $\widetilde{\boldsymbol{\theta}}_1(x)$ and $\widetilde{\boldsymbol{\theta}}_2(x)$. The aim is

to select automatically from $\{\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2\}$. Assume that the noise is known and that either the parametric model (1.4) is true "globally", i.e. on $\mathcal{W}_2$ and consequently (due to (A2)) on $\mathcal{W}_1$, either (1.4) is satisfied only on $\mathcal{W}_1$. Two wrong choices are possible:

(I) $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_1$ in the global parametric situation when the correct estimator is $\widetilde{\boldsymbol{\theta}}_2$;
(II) $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_2$ when the parametric model is true only on $\mathcal{W}_1$ and the correct estimator is $\widetilde{\boldsymbol{\theta}}_1$.

These two situations are highly asymmetric.

Consider (I). Here $\boldsymbol{\theta}_1^* = \boldsymbol{\theta}_2^* = \boldsymbol{\theta}$ and $\mathbb{E} = \mathbb{E}_{\boldsymbol{\theta}}$, that is $\mathbb{E}_{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1^* = \boldsymbol{\theta}_2^* = \mathbb{E}_{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}_2$. We have accepted the worst estimator corresponding to the smaller amount of data $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_1$ with larger variance. Since $\mathcal{W}_1(x) \le \mathcal{W}_2(x)$, by (3.2) we have $\operatorname{Var}\widetilde{\boldsymbol{\theta}}_1 = \mathbf{B}_1^{-1} \succeq \mathbf{B}_2^{-1} = \operatorname{Var}\widetilde{\boldsymbol{\theta}}_2$ for the binary weights; for the non-binary weights in $[0,1]$ $\operatorname{Var}\widetilde{\boldsymbol{\theta}}_l \preceq \mathbf{B}_l^{-1}$, $l = 1, 2$, and the matrices $\mathbf{B}_l^{-1}$ serve as monotonized bounds for the variances. Adding and subtracting $\mathrm{L}(\mathbf{W}_2, \widetilde{\boldsymbol{\theta}}_2)$ we get

$$\mathrm{L}(\mathbf{W}_2, \widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})\mathbb{I}\{\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_1\} = \mathrm{L}(\mathbf{W}_2, \widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2) + \mathrm{L}(\mathbf{W}_2, \widetilde{\boldsymbol{\theta}}_2, \boldsymbol{\theta}).$$

Let $r = 1$. The risk of this log-likelihood ratio is

$$\mathbb{E}_{\boldsymbol{\theta}}|2\,\mathrm{L}(\mathbf{W}_2, \widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})|\mathbb{I}\{\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_1\} \le \mathbb{E}_{\boldsymbol{\theta}}|2\,\mathrm{L}(\mathbf{W}_2, \widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2)| + \mathbb{E}_{\boldsymbol{\theta}}|2\,\mathrm{L}(\mathbf{W}_2, \widetilde{\boldsymbol{\theta}}_2, \boldsymbol{\theta})|.$$

The second term of the RHS is bounded with $C(p,1)$ by Corollary 4.4. On the contrary, the first term related to the "pure noise" (the value $\boldsymbol{\theta}$ cancels in $\widetilde{\boldsymbol{\theta}}_2 - \widetilde{\boldsymbol{\theta}}_1$) by the second statement of Lemma 4.7 can be much larger than $C(p,1)$. However, because the distribution of this quantity does not depend on the unknown parameter $\boldsymbol{\theta}$, its risk can be easily controlled by the choice of the threshold $\mathfrak{z}_1$. Thus we have arrived at the PC: $\mathfrak{z}_1$ should provide $\mathbb{E}_{\boldsymbol{\theta}}|2\,\mathrm{L}(\mathbf{W}_2, \widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2)| \le \alpha C(p,1)$ with some $\alpha \in (0,1]$.

In a general case at this place one needs exponential inequalities to bound the large deviations of the stochastic term in the "no noise" situation. For analysis of large deviations of a contrast function related to the considered here approach see [16].

Turn now to (II). Here $\boldsymbol{\theta}_1^* \ne \boldsymbol{\theta}_2^*$. Similarly to the previous case we have

$$\mathbb{E}|2\,\mathrm{L}(\mathbf{W}_1, \widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_1^*)|\mathbb{I}\{\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_2\} \le \mathbb{E}|2\,\mathrm{L}(\mathbf{W}_1, \widetilde{\boldsymbol{\theta}}_2, \widetilde{\boldsymbol{\theta}}_1)| + \mathbb{E}|2\,\mathrm{L}(\mathbf{W}_1, \widetilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1^*)|$$

and as in (I) the second term of the RHS is bounded with $C(p,1)$. But one can say nothing about the first term and the only way to control it is the procedure: we say that the choice $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_2$ is acceptable in this situation if $2\,\mathrm{L}(\mathbf{W}_1, \widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2) \le \mathfrak{z}_1$, where $\mathfrak{z}_1$ is the threshold fixed by the PC. To choose from more than two estimators the selection rule at every step k accepts the estimator $\widetilde{\boldsymbol{\theta}}_k$ if and only if $2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k) \le \mathfrak{z}_l$ for all $l < k$ with the proviso that $\widetilde{\boldsymbol{\theta}}_{k-1}$ had been accepted at the previous step of the procedure.

Note also that exactly this part of the procedure can cause the well-known oversmoothing effect of the Lepski-type procedures, because one admits oversmoothing in the range of threshold. The threshold corresponding to the oracle

scale presents also in the leading term of the risk, c.f. Theorem 3.3. That is why it is so important to select the smallest possible sequence of thresholds and it is shown in [36] p. 2791 that the PC's provide such a sequence. However, to fix the thresholds by simulations as in [36] the exact knowledge of the noise is required. This explains the interest of the author to the noise misspecification and generalization of the propagation approach to this set-up.

3. Theoretical study

In order to infer on the admissible level of misspecification for "model" covariance matrix from (1.4) we need to introduce a parameter δ reflecting the relative variability in errors:

(A3) *There exists* $\delta \in [0,1)$ *such that*

$$1-\delta \le \sigma_{0,i}^2/\sigma_i^2 \le 1+\delta \quad \textit{for all} \quad i=1,\ldots,n.$$

Remark 3.1. Clearly, the value of δ is not available. This parameter is used to trace the influence of the erroneously known noise. The procedure given by (2.13), (2.14) and (2.17) does not require knowledge of δ or of the true covariance matrix Σ_0.

3.1. Upper bound for the critical values

For any real symmetric matrices A and B we write $A \preceq B$ if $\vartheta^\top A\,\vartheta \le \vartheta^\top B\,\vartheta$ for all vectors ϑ, or, equivalently, if and only if the matrix $B-A$ is nonnegative definite. Assuming $(A3)$, the true covariance matrix $\Sigma_0 \preceq \Sigma(1+\delta)$, and the variance of the estimate $\widetilde{\boldsymbol{\theta}}_k$ is bounded with $\mathbf{B}_k^{-1}$:

$$\begin{aligned} V_k \stackrel{\mathrm{def}}{=} \operatorname{Var}\widetilde{\boldsymbol{\theta}}_k &= \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0\mathbf{W}_k\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1} \qquad (3.1)\\ &\preceq (1+\delta)\mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma\mathbf{W}_k\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}\\ &= (1+\delta)\mathbf{B}_k^{-1}\boldsymbol{\Psi}\Sigma^{-1/2}\mathcal{W}_k^2\Sigma^{-1/2}\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}\\ &\preceq (1+\delta)\mathbf{B}_k^{-1}\boldsymbol{\Psi}\Sigma^{-1/2}\mathcal{W}_k\Sigma^{-1/2}\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}\\ &= (1+\delta)\mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}\\ &= (1+\delta)\mathbf{B}_k^{-1}. \qquad (3.2)\end{aligned}$$

The last inequality follows from the observation that all entries of the diagonal "weight" matrix $\mathcal{W}_k$ do not exceed one, implying $\mathcal{W}_k^2 \preceq \mathcal{W}_k$. The strict equality takes place if $\{w_{k,i}\} \in \{0,1\}$ and the noise is known, i.e. if $\delta = 0$. To justify the procedure it is necessary to show that the critical values fixed by (PC) are finite. This will be obtained under the following assumption:

(A4) *Let for some constants* u_0 *and* u *such that* $1 < u_0 \le u$ *for any* $2 \le k \le K$ *the matrices* $\mathbf{B}_k$ *satisfy*

$$u_0 I_p \preceq \mathbf{B}_{k-1}^{-1/2}\,\mathbf{B}_k\,\mathbf{B}_{k-1}^{-1/2} \preceq u I_p$$

Remark 3.2. In the "one dimensional case" $p = 1$, that is for the local constant fitting, the "matrix" $\mathbf{B}_k = \sum_{i=1}^n w_{k,i}\sigma_i^{-2} \geq \mathbf{B}_{k-1}$ is just a weighted "local sample size". Assume for simplicity that $\sigma_i^2 \equiv \sigma^2$, the weights are rectangular kernels $w_{k,i}(x) = \mathbb{I}\{|X_i - x| \leq h_k/2\}$ and the design is equidistant. Then for n sufficiently large

$$\frac{1}{n}\mathbf{B}_k = \frac{1}{n\sigma^2}\sum_{i=1}^n \mathbb{I}\left\{\left|\frac{i}{n} - x\right| \leq \frac{h_k}{2}\right\} \approx \frac{h_k}{\sigma^2},$$

and Assumption $(A4)$ with $u_0 = u$ means that the bandwidths grow geometrically: $h_k = uh_{k-1}$.

Now we are able to demonstrate the finiteness of the critical values.

Proposition 3.1 (Theoretical choice of the critical values)**.** *Assume* $(A1)-(A2)$ *and* $(A4)$. *The adaptive procedure defined by* (2.13), (2.14) *and* (2.17) *is well defined in the sense that the choice of the critical values of the form*

$$\mathfrak{z}_k = \frac{4}{\mu}\left\{r(K-k)\log u + \log(K/\alpha) - \frac{p}{4}\log(1-4\mu) - \log(1-u^{-r}) + \overline{C}(p,r)\right\} \tag{3.3}$$

provides the conditions (2.17) *for all* $k \leq K$. *Particularly,*

$$\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_K - \widehat{\boldsymbol{\theta}})^\top \mathbf{B}_K(\widetilde{\boldsymbol{\theta}}_K - \widehat{\boldsymbol{\theta}})|^r \leq \alpha C(p,r). \tag{3.4}$$

In (3.3) $\mu \in (0, 1/4)$ *is an arbitrary constant,* $u > 1$ *is given by Assumption* $(A4)$, $r > 0$ *and* $\alpha \in (0, 1]$ *are from the PC's, and*

$$\overline{C}(p,r) = \log\left\{\frac{2^{2r}[\Gamma(2r + p/2)\Gamma(p/2)]^{1/2}}{\Gamma(r + p/2)}\right\}.$$

The proof is given in Section 4.2.

Remark 3.3. *Dependence of the thresholds on the parameters* r *and* α in connection with the performance of the procedure is discussed in Remark 2.5.

Dependance on the number of scales. For kernel estimators, c.f. Remark 3.2, Assumption $(A4)$ essentially means a geometrical grid of bandwidths $h_{k-1} = u^{-1}h_k$ implying $h_1 = u^{-(K-1)}h_K$. Thus $(K-1)\log u = \log(h_K/h_1)$, where $\log u$ is a fixed constant, say equal to $\log 2$. Since $h_K \leq 1$ and $h_1 \geq 1/n$, the number of scales is at most of order $\log n$, that is $K \asymp \log(h_K/h_1) \leq \log n$ and is related to the "adaptive factor" to pay for the pointwise adaptation, c.f. (2.11) in [30] p. 2518 and the discussion therein. The leading term in (3.3) is $const.(K-k)$ and it shows that the thresholds $\mathfrak{z}_k$ linearly decrease in k providing stability of the procedure at the first steps and sensitivity to deviations of the parametric fit from the true model at the further steps of the algorithm. The thresholds are at most of order $\log n$ and this "log" disappears at the "last point" $k = K$. That is if the parametric assumption is true, there is no "log-payment", c.f. Remark 3.6.

3.2. Quality of estimation in the nearly parametric case

The critical values of the procedure $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ were selected by the propagation conditions (2.17) under the measure $\mathcal{N}(\boldsymbol{\theta}, \Sigma)$ that is probably not confirmed by the data. Let now the maximizers $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*$ of the expected local log-likelihoods are only approximately equal, say to $\boldsymbol{\theta}$, up to some $k \le K$ and the covariance matrix is Σ_0. The meaning of "approximately equal" will be explained below.

The aim is to justify the use of the critical values in this situation. For this purposes we study the discrepancy between the joint distributions of linear estimators $\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_k$ for $k = 1, \ldots, K$ under the "no bias" assumption corresponding to the distributions with mean $\boldsymbol{\theta}_1^* = \cdots = \boldsymbol{\theta}_k^* = \boldsymbol{\theta}$ and possibly incorrectly specified covariance matrix Σ, and in the general situation with $\boldsymbol{\theta}_1^* \ne \cdots \ne \boldsymbol{\theta}_k^*$ and covariance Σ_0. Denote the expectations w.r.t. these measures by $\mathbb{E}_{\boldsymbol{\theta},\Sigma} := \mathbb{E}_{k,\boldsymbol{\theta},\Sigma}$ and $\mathbb{E}_{\boldsymbol{f},\Sigma_0} := \mathbb{E}_{k,\boldsymbol{f},\Sigma_0}$ respectively and the $p \times k$ matrix of the first k estimators and the expectations correspondingly by

$$\begin{aligned}
\widetilde{\boldsymbol{\Theta}}_k &\stackrel{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_k), \\
\boldsymbol{\Theta}_k^* &\stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{f},\Sigma_0} \widetilde{\boldsymbol{\Theta}}_k = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*), \\
\boldsymbol{\Theta}_k &\stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{\theta},\Sigma} \widetilde{\boldsymbol{\Theta}}_k = (\boldsymbol{\theta}, \ldots, \boldsymbol{\theta}).
\end{aligned}$$

Let $A \otimes B$ stand for the Kronecker product of matrices $A = (a_{i,j})_{1 \le i \le m, 1 \le j \le n}$ and B defined as

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}.$$

Denote the $pk \times pk$ covariance matrices of $\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_k^\top = (\widetilde{\boldsymbol{\theta}}_1^\top, \ldots, \widetilde{\boldsymbol{\theta}}_k^\top) \in \mathbb{R}^{pk}$ by

$$\boldsymbol{\Sigma}_k \stackrel{\text{def}}{=} \operatorname{Var}_{\boldsymbol{\theta},\Sigma}[\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_k] = \mathbf{D}_k (J_k \otimes \Sigma) \mathbf{D}_k^\top, \tag{3.5}$$

$$\boldsymbol{\Sigma}_{k,0} \stackrel{\text{def}}{=} \operatorname{Var}_{\boldsymbol{f},\Sigma_0}[\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_k] = \mathbf{D}_k (J_k \otimes \Sigma_0) \mathbf{D}_k^\top, \tag{3.6}$$

where the matrix J_k is a $k \times k$ matrix with all its elements equal to 1, and the $pk \times nk$ block diagonal matrix $\mathbf{D}_k$ is defined as follows:

$$\begin{aligned}
\mathbf{D}_k &\stackrel{\text{def}}{=} D_1 \oplus \cdots \oplus D_k = \operatorname{diag}(D_1, \ldots, D_k), \\
D_l &\stackrel{\text{def}}{=} \mathbf{B}_l^{-1} \boldsymbol{\Psi} \mathbf{W}_l, \quad l = 1, \ldots, k.
\end{aligned} \tag{3.7}$$

By Lemma 4.8 from Section 4 under Assumption $(A3)$ with the same δ the similar relation holds for the covariance matrices $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Sigma}_{k,0}$ of the sets of linear estimators:

$$(1 - \delta) \boldsymbol{\Sigma}_k \preceq \boldsymbol{\Sigma}_{k,0} \preceq (1 + \delta) \boldsymbol{\Sigma}_k, \quad k \le K. \tag{3.8}$$

In spite of by Lemma 4.10 the moment generating function of $\operatorname{vec}\widetilde{\boldsymbol{\Theta}}_K$ has the form corresponding to the multivariate normal distribution this representation makes sense only if $\boldsymbol{\Sigma}_K$ is nonsingular. Notice that $\operatorname{rank}(J_K \otimes \Sigma) = n$. From $J_K \otimes \Sigma \succeq 0$ it follows only that $\boldsymbol{\Sigma}_K \succeq 0$, similarly, $\boldsymbol{\Sigma}_{K,0} \succeq 0$. However, without any additional assumptions it is easy to show, see Lemma 4.9, that for rectangular kernels $\boldsymbol{\Sigma}_K \succ 0$. On the other hand, due to (3.8), it is enough to require nonsingularity only for the matrix $\boldsymbol{\Sigma}_K$ corresponding to the approximate model (1.4), and its choice belongs to a statistician. In what follows we assume that $\boldsymbol{\Sigma}_K \succ 0$.

Denote by $\mathbb{P}^k_{\boldsymbol{\theta},\Sigma} = \mathcal{N}(\operatorname{vec}\boldsymbol{\Theta}_k, \boldsymbol{\Sigma}_k)$ and by $\mathbb{P}^k_{\boldsymbol{f},\Sigma_0} = \mathcal{N}(\operatorname{vec}\boldsymbol{\Theta}^*_k, \boldsymbol{\Sigma}_{k,0})$, $k = 1,\dots,K$, the distributions of $\operatorname{vec}\widetilde{\boldsymbol{\Theta}}_k$ under the assumption that the parametric model (1.4) is true up to the scale k and under the assumption that nonparametric model (1.1) takes place. Denote also the Radon-Nikodym derivative by

$$Z_k \stackrel{\text{def}}{=} \frac{\mathrm{d}\mathbb{P}^k_{\boldsymbol{f},\Sigma_0}}{\mathrm{d}\mathbb{P}^k_{\boldsymbol{\theta},\Sigma}}. \tag{3.9}$$

Then Lemma 4.11 gives the Kullback-Leibler divergence between these measures:

$$\begin{aligned} 2\mathbb{KL}(\mathbb{P}^k_{\boldsymbol{f},\Sigma_0}, \mathbb{P}^k_{\boldsymbol{\theta},\Sigma}) &\stackrel{\text{def}}{=} 2\mathbb{E}_{\boldsymbol{f},\Sigma_0}\log(Z_k) \\ &= \Delta(k) + \log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) + \operatorname{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{k,0}) - pk, \end{aligned} \tag{3.10}$$

where

$$\Delta(k) \stackrel{\text{def}}{=} b(k)^\top \boldsymbol{\Sigma}_k^{-1} b(k) \tag{3.11}$$

$$b(k) \stackrel{\text{def}}{=} \operatorname{vec}\boldsymbol{\Theta}^*_k - \operatorname{vec}\boldsymbol{\Theta}_k. \tag{3.12}$$

If there would be no any "noise misspecification", i.e. if $\delta \equiv 0$ implying $\Sigma = \Sigma_0$, then $\Delta(k) = b(k)^\top \boldsymbol{\Sigma}_k^{-1} b(k) = 2\mathbb{KL}(\mathbb{P}^k_{\boldsymbol{f},\Sigma}, \mathbb{P}^k_{\boldsymbol{\theta},\Sigma})$. Under Assumption $(A2)$, the quantity $\Delta(k)$ grows with k, so following the terminology suggested in [36], we introduce the *small modeling bias condition*:

(*SMB*) *Let for some $k \le K$ and $\boldsymbol{\theta}$ exist a finite constant $\Delta \ge 0$ such that $\Delta(k) \le \Delta$.*

Monotonicity of $\Delta(k)$ and (SMB) immediately imply that

$$\sup_{1\le l\le k} \Delta(l) \le \Delta.$$

Relation (3.8) yields $-pk\delta \le \operatorname{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{k,0}) - pk \le pk\delta$. Thus the statement of Lemma 4.11 gives a bound for the Kullback-Leibler divergence in terms of δ:

$$\begin{aligned} -\frac{pk}{2}\log(1+\delta) + \frac{\Delta(k)}{2} - \frac{pk\delta}{2} &\le \mathbb{KL}(\mathbb{P}^k_{\boldsymbol{f},\Sigma_0}, \mathbb{P}^k_{\boldsymbol{\theta},\Sigma}) \\ &\le -\frac{pk}{2}\log(1-\delta) + \frac{\Delta(k)}{2} + \frac{pk\delta}{2}. \end{aligned} \tag{3.13}$$

Moreover, if $\delta = \delta(n)$ and $\delta(n) \to 0+$ as $n \to \infty$

$$\Delta(k) - 2pk\delta + o(\delta) \le 2\mathbb{KL}(\mathbb{P}^k_{\boldsymbol{f},\Sigma_0}, \mathbb{P}^k_{\boldsymbol{\theta},\Sigma}) \le \Delta(k) + 2pk\delta + o(\delta). \tag{3.14}$$

This means that, if for some k Assumption (SMB) is fulfilled and $\delta = O(1/K)$, then the Kullback-Leibler divergence between the measures$\mathbb{P}^k_{\boldsymbol{\theta},\Sigma}$ and $\mathbb{P}^k_{\boldsymbol{f},\Sigma_0}$ is bounded by a small constant.

Now one can state the crucial property for obtaining the final oracle result.

Theorem 3.2 (Propagation property)**.** *Assume* $(A1) - (A4)$ *and* (PC)*. Then for any* $k \le K$ *the following upper bounds hold:*

$$\begin{aligned}
&\mathbb{E}|(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})|^{r/2} \\
\le\ & C(p,r)^{1/2}(1+\delta)^{pk/4}(1-\delta)^{-3pk/4} \exp\left\{\varphi(\delta)\frac{\Delta(k)}{2(1-\delta)}\right\}, \\
&\mathbb{E}|(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^{r/2} \\
\le\ & (\alpha C(p,r))^{1/2}(1+\delta)^{pk/4}(1-\delta)^{-3pk/4} \exp\left\{\varphi(\delta)\frac{\Delta(k)}{2(1-\delta)}\right\},
\end{aligned}$$

where $\varphi(\delta) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{for homogeneous errors,} \\ \frac{2(1+\delta)}{(1-\delta)^2} - 1 & \text{otherwise.} \end{cases}$

Here $\widetilde{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_k(x)$ *is the QMLE defined by* (2.3)*,* $\boldsymbol{\theta}$ *is the parameter from* (1.4)*,* $\widehat{\boldsymbol{\theta}}_k(x) = \widetilde{\widetilde{\boldsymbol{\theta}}}_{\min k,\widehat{k}}(x)$ *is the adaptive estimate at the* k*th step of the procedure,* $C(p,r)$ *is the constant from the PC's defined in* (4.8) *and* p *is the number of basis functions used for the linear fitting.*

The proof is given in Subsection 4.4.

Remark 3.4. Bounds (4.32) and (4.31) obtained in the proof of the theorem (Section 4.4) give a condition on the relative error in the noise misspecification. Let $\delta = \delta(n) \to 0+$ as $n \to \infty$. Then for every $k \le K$

$$\varphi(\delta)\frac{\Delta(k)}{1+\delta} - 2pk\delta + o(\delta) \le \log \mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2] \le \varphi(\delta)\frac{\Delta(k)}{1-\delta} + 2pk\delta + o(\delta)$$

with Z_k defined by (3.9). This bound implies, up to the additive constant $0.5\log\big(\alpha C(p,r)\big)$, the same asymptotic behavior for the logarithm of the risk of adaptive estimate $\log \mathbb{E}\|\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)\|^r$ at each step of the procedure. Because by (SMB) the quantity $\Delta(k)$ is supposed to be bounded by a small constant, and K is of order $\log n$, see Remark 3.3, the expectation $\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]$ is small if $\delta = O(1/\log n)$ and, consequently, the risk $\mathbb{E}\|\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)\|^r$ is bounded, c.f. (2.17). This means that for a plug-in estimator of the variance only the logarithmic in sample size quality is needed. This observation is of particular importance, since it is known from [35] that over classes of functions with bounded second derivative the rate $n^{-1/2}$ of variance estimation is achievable only for the dimension $d \le 8$.

Remark 3.5. The propagation property guarantees that the adaptive procedure does not stop with large probability while $\Delta(k)$ is small, i.e. under (SMB), and if the relative error δ in the noise is sufficiently small.

3.3. Quality of estimation in the nonparametric case: The oracle result

Define the *oracle index* as the largest index $k \le K$ such that (SMB) holds:

$$k^* \stackrel{\text{def}}{=} \max\{k \le K : \Delta(k) \le \Delta\}. \tag{3.15}$$

Theorem 3.3. *Let $\Delta(1) \le \Delta$, i.e. the first estimate is always accepted in the testing procedure. Let $\mathbf{\Sigma}_K \succ 0$ and k^* be the oracle index. Let $\widetilde{\boldsymbol{\theta}}_{k^*}$ be the nonadaptive estimator defined by (2.6) corresponding to k^* and $\widehat{\boldsymbol{\theta}}$ be an output of the procedure (2.13) – (2.14). Then under (PC) and assumptions $(A1) - (A4), (SMB)$ and the risk between the adaptive and oracle estimators is bounded with the following expression:*

$$\begin{aligned} &\mathbb{E}|(\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})|^{r/2} \\ \le\ & \mathfrak{z}_{k^*}^{r/2} + (\alpha C(p,r))^{1/2} (1+\delta)^{pk^*/4} (1-\delta)^{-3pk^*/4} \exp\left\{\varphi(\delta) \frac{\Delta}{2(1-\delta)}\right\}, \end{aligned} \tag{3.16}$$

where $\varphi(\delta)$ is as in Theorem 3.2 and $C(p,r)$ is the constant from the PC's defined in (4.8).

Remark 3.6. The second term in the RHS of (3.16) is bounded with a constant with the proviso that $\delta = O(1/\log n)$, see Remark 3.4, and the leading term is $\mathfrak{z}_{k^*}^{r/2}$ that by Proposition 3.1 has the form $\mathfrak{z}_{k^*} = C_1 r(K - k^*) + C_2 \log(K/\alpha) + C_3$. The leading term K is at most of order $\log n$, see Remark 3.3, and is the unavoidable payment for the pointwise adaptation, see Theorem 2 on the lower bound in [27]. This term cancels if $k^* = K$, that is when the deviation of the parametric fit from the true model is not significant for all observations. This means that the parametric set-up takes place globally and there is no adaptation involved. The canceling of the *log* term at the last point of the range of adaptation in the rate is a common feature of this type procedures, sf. [27, 29, 30].

The LHS of the inequality (3.16) is the mathematical expectation of the oracle log-likelihood ratio $|2\,\mathrm{L}(\mathbf{W}_{k^*}, \widetilde{\boldsymbol{\theta}}_{k^*}, \widehat{\boldsymbol{\theta}})|^{r/2}$, or the risk of the difference between the adaptive estimator $\widehat{\boldsymbol{\theta}}$ and its nonadaptive counterpart $\widetilde{\boldsymbol{\theta}}_{k^*}$ normalized by the bound for the variance of the oracle estimator. Recall that by (3.2) in the case of binary weights the matrix $\mathbf{B}_{k^*}^{-1} = \operatorname{Var} \widetilde{\boldsymbol{\theta}}_{k^*}$; generally we have only $\operatorname{Var} \widetilde{\boldsymbol{\theta}}_{k^*} \preceq \mathbf{B}_{k^*}^{-1}$. Loosely speaking, the result says that the risk of adaptive estimator is of order of the oracle variance multiplied by the logarithmic in sample size factor $\mathfrak{z}_{k^*}$.

Proof. By the definition of the adaptive estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{k}}$. Because the events $\{\widehat{k} \le k^*\}$ and $\{\widehat{k} > k^*\}$ are disjunct, one can write

$$\begin{aligned} & \mathbb{E}|(\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})|^{r/2} \\ = \; & \mathbb{E}|(\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}})|^{r/2} \mathbb{I}\{\widehat{k} \le k^*\} \\ + \; & \mathbb{E}|(\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}})|^{r/2} \mathbb{I}\{\widehat{k} > k^*\}. \end{aligned}$$

If $\widehat{k} \le k^*$ then $\widehat{\boldsymbol{\theta}}_{k^*} \stackrel{\text{def}}{=} \widetilde{\boldsymbol{\theta}}_{\min\{k^*, \widehat{k}\}} = \widetilde{\boldsymbol{\theta}}_{\widehat{k}}$. Thus, to bound the first summand, it is enough to apply Theorem 3.2 with $k = k^*$.

To bound the second expectation, i.e. to bound the fluctuations of adaptive estimate $\widehat{\boldsymbol{\theta}}$ at the steps of the procedure for which the (SMB) condition is not fulfilled anymore, just notice that for $\widehat{k} > k^*$ the quadratic form coincides with the test statistic $T_{k^*,\widehat{k}}$

$$\begin{aligned} & (\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}}) \\ = \; & (\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}}) = T_{k^*,\widehat{k}}. \end{aligned}$$

But the index $\widehat{k}$ was accepted by the procedure, this means that $T_{l,\widehat{k}} \le \mathfrak{z}_l$ for all $l < \widehat{k}$ and therefore for $l = k^*$. Thus

$$\mathbb{E}|(\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})|^{r/2} \mathbb{I}\{\widehat{k} > k^*\} \le \mathfrak{z}_{k^*}^{r/2}.$$

□

3.4. Componentwise oracle risk bounds

Theorem 3.3 provides the oracle risk bound for the adaptive estimator $\widehat{\boldsymbol{\theta}}(x) = \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)$ of the parameter vector $\boldsymbol{\theta}(x) \in \mathbb{R}^p$ corresponding to the estimator $\widehat{f}_{\widehat{\boldsymbol{\theta}}}(x)$ of type (1.6). It is interesting to have a look at the oracle quality of estimation of the components $\theta^{(1)}, \dots, \theta^{(p)}$ of the vector $\boldsymbol{\theta}$ having in mind that the choice of polynomial basis leads to the direct estimation of the value of regression function and the derivatives by the coordinates of $\widehat{\boldsymbol{\theta}}$.

Denote by $LP_k(p-1)$ a local polynomial estimator of order $p-1$ corresponding to the kth degree of localization and by $LP^{ad}(p-1)$ its adaptive counterpart, i.e. $LP^{ad}(p-1) \stackrel{\text{def}}{=} LP_{\widehat{k}}(p-1)$. If the basis is polynomial and the regression function $f(\cdot)$ is sufficiently smooth in a neighborhood of x, then $\widehat{\boldsymbol{\theta}}(x)$ is the $LP^{ad}(p-1)$ of the vector $(f(x), f'(x), \dots, f^{(p-1)}(x))^\top$ of the values of the function f and its derivatives at the reference point $x \in \mathbb{R}^d$.

Therefore we are going to obtain a similar to the previous section oracle result for the components of the vector $\widehat{\boldsymbol{\theta}}(x)$, namely for $\boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}(x)$, $j = 1, \dots, p$, where $\boldsymbol{e}_j$ is the jth canonical basis vector in $\mathbb{R}^p$. As a corollary of this general result in the case of polynomial basis we get an oracle risk bound for $LP^{ad}(p-1)$ estimator of the function f and its derivatives at the point x.

$LP_k(p-1)$ estimator of $f^{(j-1)}(x)$ is given by

$$\begin{aligned} \widetilde{f}_k^{(j-1)}(x) &= e_j^\top \widetilde{\boldsymbol{\theta}}_k(x),\ j=1,\dots,p, \\ \widetilde{f}_k(x) &= \widetilde{f}_k^{(0)}(x) = e_1^\top \widetilde{\boldsymbol{\theta}}_k(x). \end{aligned} \tag{3.17}$$

Then the adaptive local polynomial estimators are defined as follows:

$$\begin{aligned} \widehat{f}^{(j-1)}(x) &= e_j^\top \widehat{\boldsymbol{\theta}}(x),\ j=1,\dots,p, \\ \widehat{f}(x) &= e_1^\top \widehat{\boldsymbol{\theta}}(x). \end{aligned} \tag{3.18}$$

Similarly, the adaptive estimators of the function f and its derivatives corresponding to the kth step of the procedure are given by

$$\widehat{f}_k^{(j-1)}(x) \stackrel{\text{def}}{=} e_j^\top \widehat{\boldsymbol{\theta}}_k(x),\ j=1,\dots,p. \tag{3.19}$$

Thus, if the basis is polynomial, the estimator $\widehat{f}(x) \stackrel{\text{def}}{=} \widehat{f}^{(0)}(x)$ is the $LP^{ad}(p-1)$ estimator of the value $f(x)$, and $\widehat{f}^{(j-1)}(x)$ with $j=2,\dots,p$ are, correspondingly, the $LP^{ad}(p-1)$ estimators of the values of its derivatives. However the results of Theorems 3.3 and 3.9 hold for any basis satisfying the conditions of the theorems. For the study below we need the following assumptions:

(A5) *There exists a positive finite number $\sigma_{max}(k)$ such that for $i : X_i \in U_{h_k}(x)$, with the neighborhood of the estimation point $U_{h_k}(x)$ given by $\mathbf{W}_k$ the variances of errors from the parametric (known) model (1.4) are locally uniformly bounded:*

$$\sigma_i^2 \le \sigma_{max}^2(k).$$

(A6) *Let assumption (A5) be satisfied. There exists a number $\Lambda_0 > 0$ such that for any $k=1,\dots,K$ the smallest eigenvalue $\lambda_p(\mathbf{B}_k) \ge nh_k^d\Lambda_0\sigma_{max}^{-2}(k)$ for n sufficiently large.*

Remark 3.7. The first assumption is not restrictive at all, since it is about the known variance from the model we use for the construction of estimators. The last assumption is stronger than the requirement $\mathbf{B}_k(x) \succ 0$. Lemmas 1.5, 1.4 in [41] shows that this assumption holds for nonnegative kernels, which are bounded from below on a set of positive Lebesgue measure. The constant Λ_0 is related to the smallest eigenvalue of the matrix $\boldsymbol{B}$ from Lemma 3.13.

Thus for any $k=1,\dots K$ and for any $\gamma \in \mathbb{R}^p$ we have

$$\gamma^\top \mathbf{B}_k^{-1}\gamma \le \frac{\sigma_{max}^2(k)}{nh_k^d\Lambda_0}\|\gamma\|^2 \le \frac{\overline{\sigma}_{max}^2(k)}{nh_k^d\Lambda_0}\|\gamma\|^2, \tag{3.20}$$

where

$$\overline{\sigma}_{max}^2(k) \stackrel{\text{def}}{=} \max_{1\le l\le k}\sigma_{max}^2(l)\overline{\sigma}_{max}^2(k) \stackrel{\text{def}}{=} \max_{1\le l\le k}\sigma_{max}^2(l). \tag{3.21}$$

Thus we have the following bound:

Lemma 3.4. *Let* $(A5)$ *and* $(A6)$ *be satisfied. Then for any* $j = 1, \dots, p$ *and* $k, k' = 1, \dots K$ *the following bound holds:*

$$\left(\frac{nh_k^d\Lambda_0}{\overline{\sigma}^2_{max}(k)}\right)^{1/2} |e_j^\top\widetilde{\boldsymbol{\theta}}_k - e_j^\top\widetilde{\boldsymbol{\theta}}_{k'}| \le \|\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_{k'})\|.$$

Proof. By (3.20) taking $\gamma = \mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_{k'})$ we have

$$\begin{aligned} |e_j^\top\widetilde{\boldsymbol{\theta}}_k - e_j^\top\widetilde{\boldsymbol{\theta}}_{k'}|^2 &\le \|\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_{k'}\|^2 \\ &= \|\mathbf{B}_k^{-1/2}\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_{k'})\|^2 \\ &\le \frac{\overline{\sigma}^2_{max}(k)}{nh_k^d\Lambda_0}\|\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_{k'})\|^2. \end{aligned}$$

□

To obtain the "componentwise" oracle risk bounds we need to recheck the "propagation property". Firstly, notice that the "propagation conditions" (2.17) on the choice of the critical values $\mathfrak{z}_1, \dots, \mathfrak{z}_{K-1}$ imply the similar bounds for the components $e_j^\top\widehat{\boldsymbol{\theta}}_k(x)$. Recall that $\widehat{\boldsymbol{\theta}}_k \stackrel{\text{def}}{=} \widetilde{\boldsymbol{\theta}}_{\min\{k,\widehat{k}\}}$. By (2.17), Lemma 3.4 and the pivotality property from Lemma 4.1 we have the following simple observation that serves as a componentwise counterpart of PC:

Lemma 3.5. *Under the propagation conditions* (PC) *for any* $\boldsymbol{\theta} \in \mathbb{R}^p$ *and all* $k = 2, \dots, K$ *we have:*

$$\begin{aligned} \left(\frac{nh_k^d\Lambda_0}{\overline{\sigma}^2_{max}(k)}\right)^r \mathbb{E}_{\boldsymbol{\theta},\Sigma}|e_j^\top\widetilde{\boldsymbol{\theta}}_k(x) - e_j^\top\widehat{\boldsymbol{\theta}}_k(x)|^{2r} &\le \mathbb{E}_{0,\Sigma}\|\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)\|^{2r} \\ &\le \alpha C(p,r). \end{aligned}$$

Here $\mathbb{E}_{0,\Sigma}$ *stands for the expectation w.r.t.* $\mathcal{N}(0,\Sigma)$ *and* $C(p,r)$ *is given by* (4.8).

As before we suppress the dependence on x. To get the propagation property we study for $k = 1, \dots, K$ the joint distributions of $e_j^\top\widetilde{\boldsymbol{\theta}}_1, \dots, e_j^\top\widetilde{\boldsymbol{\theta}}_k$, that is the distribution of $e_j^\top\widetilde{\boldsymbol{\Theta}}_k$, the jth row of the matrix $\widetilde{\boldsymbol{\Theta}}_k$. Obviously,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{f},\Sigma_0}[e_j^\top\widetilde{\boldsymbol{\Theta}}_k] &= e_j^\top\boldsymbol{\Theta}_k^* = (e_j^\top\boldsymbol{\theta}_1^*, \dots, e_j^\top\boldsymbol{\theta}_k^*), \\ \mathbb{E}_{\boldsymbol{\theta},\Sigma}[e_j^\top\widetilde{\boldsymbol{\Theta}}_k] &= e_j^\top\boldsymbol{\Theta}_k = (e_j^\top\boldsymbol{\theta}, \dots, e_j^\top\boldsymbol{\theta}). \end{aligned}$$

Recall that the matrices $\boldsymbol{\Sigma}_{k,0}$ and $\boldsymbol{\Sigma}_k$ have a block structure. Now, for instance, to study the estimator of the first coordinate of the vector $\boldsymbol{\theta} = \boldsymbol{\theta}(x)$, or of $f(x)$ in the case of the polynomial basis, we take the first elements of each block and so on.

Denote the covariance matrices of the jth elements of the vectors $\widetilde{\boldsymbol{\theta}}_1, \dots, \widetilde{\boldsymbol{\theta}}_k$

by

$$\begin{aligned}\boldsymbol{\Sigma}_{k,j} &\stackrel{\text{def}}{=} \left\{\operatorname{cov}_{\boldsymbol{\theta},\Sigma}\left[\widetilde{\theta}_l^{(j)},\widetilde{\theta}_m^{(j)}\right]\right\}_{1\le l\le m\le k}\\ &= \mathbf{D}_{k,j}(J_k\otimes\Sigma)\mathbf{D}_{k,j}^{\top}, \qquad (3.22)\\ \boldsymbol{\Sigma}_{k,0,j} &\stackrel{\text{def}}{=} \left\{\operatorname{cov}_{\boldsymbol{f},\Sigma_0}\left[\widetilde{\theta}_l^{(j)},\widetilde{\theta}_m^{(j)}\right]\right\}_{1\le l\le m\le k}\\ &= \mathbf{D}_{k,j}(J_k\otimes\Sigma_0)\mathbf{D}_{k,j}^{\top}, \qquad (3.23)\end{aligned}$$

where J_k is a $k\times k$ matrix with all its elements equal to 1, and the $k\times nk$ block diagonal matrices $\mathbf{D}_{k,j}$ is defined by

$$\begin{aligned}\mathbf{D}_{k,j} &\stackrel{\text{def}}{=} \boldsymbol{e}_j^{\top}D_1\oplus\cdots\oplus\boldsymbol{e}_j^{\top}D_k,= \left(I_k\otimes\boldsymbol{e}_j^{\top}\right)\mathbf{D}_k\\ D_l &\stackrel{\text{def}}{=} \mathbf{B}_l^{-1}\boldsymbol{\Psi}\mathbf{W}_l,\quad l=1,\dots,k. \qquad (3.24)\end{aligned}$$

Moreover, the following representation holds:

$$\begin{aligned}\boldsymbol{\Sigma}_{k,j} &= \left(I_k\otimes\boldsymbol{e}_j^{\top}\right)\mathbf{D}_k\left(J_k\otimes\Sigma\right)\mathbf{D}_k^{\top}\left(I_k\otimes\boldsymbol{e}_j^{\top}\right)^{\top}\\ &= \left(I_k\otimes\boldsymbol{e}_j\right)^{\top}\boldsymbol{\Sigma}_k\left(I_k\otimes\boldsymbol{e}_j\right), \qquad (3.25)\end{aligned}$$

where $\boldsymbol{\Sigma}_k$ is defined by (3.5). Similarly,

$$\boldsymbol{\Sigma}_{k,0,j}=\left(I_k\otimes\boldsymbol{e}_j\right)^{\top}\boldsymbol{\Sigma}_{k,0}\left(I_k\otimes\boldsymbol{e}_j\right). \qquad (3.26)$$

Thus, the important relation (3.8) is preserved for $\boldsymbol{\Sigma}_{k,j}$ and $\boldsymbol{\Sigma}_{k,0,j}$ obtained by picking up the (j,j)th elements of each block of $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Sigma}_{k,0}$ respectively.

With usual notation $\gamma^{(j)}$ for the jth component of $\gamma\in\mathbb{R}^k$, denote by

$$\begin{aligned}b_j(k) &\stackrel{\text{def}}{=} \left(\boldsymbol{e}_j^{\top}(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}),\dots,\boldsymbol{e}_j^{\top}(\boldsymbol{\theta}_k^*-\boldsymbol{\theta})\right)^{\top}\\ &= \left((\boldsymbol{\theta}_1^*-\boldsymbol{\theta})^{(j)},\dots,(\boldsymbol{\theta}_k^*-\boldsymbol{\theta})^{(j)}\right)^{\top}\in\mathbb{R}^k \qquad (3.27)\\ \Delta_j(k) &\stackrel{\text{def}}{=} b_j(k)^{\top}\boldsymbol{\Sigma}_{k,j}^{-1}b_j(k). \qquad (3.28)\end{aligned}$$

Proposition 3.6 ("Componentwise" propagation property)**.** *Under the conditions* $(A1)-(A6)$ *and* (PC) *for any* $k\le K$ *the following upper bound holds:*

$$\begin{aligned}&\left(\frac{nh_k^d\Lambda_0}{\overline{\sigma}_{max}^2(k)}\right)^{r/2}\mathbb{E}|\boldsymbol{e}_j^{\top}\widetilde{\boldsymbol{\theta}}_k(x)-\boldsymbol{e}_j^{\top}\widehat{\boldsymbol{\theta}}_k(x)|^r\\ \le\ &(\alpha\mathbb{E}|\chi_p^2|^r)^{1/2}(1+\delta)^{pk/4}(1-\delta)^{-3pk/4}\exp\left\{\varphi(\delta)\frac{\Delta_j(k)}{2(1-\delta)}\right\} \qquad (3.29)\end{aligned}$$

with $\varphi(\delta)$ *as in Theorem 3.2,* $\overline{\sigma}_{max}^2(k)$ *defined in* (3.21)*,* Λ_0 *from* $(A6)$*,* $\alpha\in(0,1]$ *and* $r>0$ *from* (PC)*.*

Corollary 3.7. *Let the basis be polynomial. Then under the conditions of the preceding theorem* $\mathbb{E}|\widetilde{f}_k^{(j-1)}(x)-\widehat{f}_k^{(j-1)}(x)|^r$ *satisfies* (3.29)

Proof. The proof essentially follows the line of the proof of Theorem 3.2. If the distributions of $\operatorname{vec}\widetilde{\boldsymbol{\Theta}}_k$ were Gaussian, then any subvector is also Gaussian.

Denote by

$$\mathbb{P}^{k,j}_{\boldsymbol{\theta},\Sigma} = \mathcal{N}\left((\boldsymbol{e}_j^\top\boldsymbol{\theta},\dots,\boldsymbol{e}_j^\top\boldsymbol{\theta})^\top, \boldsymbol{\Sigma}_{k,j}\right)$$

and by

$$\mathbb{P}^{k,j}_{\boldsymbol{f},\Sigma_0} = \mathcal{N}\left((\boldsymbol{e}_j^\top\boldsymbol{\theta}^*_1,\dots,\boldsymbol{e}_j^\top\boldsymbol{\theta}^*_k)^\top, \boldsymbol{\Sigma}_{k,0,j}\right)$$

$k = 1,\dots,K$, the distributions of $e_j^\top\widetilde{\boldsymbol{\Theta}}_k$ under the parametric assumption and in the nonparametric case correspondingly.

By the Cauchy-Schwarz inequality and Lemma 3.5

$$\left(\frac{nh_k^d\Lambda_0}{\overline{\sigma}^2_{max}(k)}\right)^{r/2}\mathbb{E}|\boldsymbol{e}_j^\top\widetilde{\boldsymbol{\theta}}_k(x) - \boldsymbol{e}_j^\top\widehat{\boldsymbol{\theta}}(x)|^r \le (\alpha\mathbb{E}|\chi^2_p|^r)^{1/2}\left(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z^2_{k,j}]\right)^{1/2}$$

with the Radon-Nikodym derivative given by $Z_{k,j} = \mathrm{d}\mathbb{P}^{k,j}_{\boldsymbol{f},\Sigma_0}/\mathrm{d}\mathbb{P}^{k,j}_{\boldsymbol{\theta},\Sigma}$. By inequalities (3.25) and (3.26) the analog of $(A3)$ is preserved for $\boldsymbol{\Sigma}_{k,0,j}$ and $\boldsymbol{\Sigma}_{k,j}$, that is, there exists $\delta\in[0,1)$ such that

$$(1-\delta)\boldsymbol{\Sigma}_{k,j} \preceq \boldsymbol{\Sigma}_{k,0,j} \preceq (1+\delta)\boldsymbol{\Sigma}_{k,j} \tag{3.30}$$

for any $k\le K$ and $j = 1,\dots,p$. Then the assertion of the theorem follows by the Taylor expansion at the point $(\boldsymbol{e}_j^\top\boldsymbol{\theta},\dots,\boldsymbol{e}_j^\top\boldsymbol{\theta})^\top$ and (3.30) similarly to the proof of Theorem 3.2. □

At this point we introduce the "componentwise" small modeling bias conditions:

(*SMBj*) *Let for some $j = 1,\dots,p$, $k(j)\le K$ and $\theta^{(j)} = e_j^\top\boldsymbol{\theta}$ exist a finite constant $\Delta_j\ge 0$ such that*

$$\Delta_j(k(j)) \le \Delta_j, \tag{3.31}$$

where $\Delta_j(k)$ is defined by (3.28).

Definition 3.8. For each $j = 1,\dots,p$ the oracle index $k^*(j)$ is defined as the largest index of the scale for which the $(SMBj)$ condition holds, that is

$$k^*(j) = \max\{k\le K : \Delta_j(k)\le\Delta_j\}. \tag{3.32}$$

Proposition 3.9. *Assume $(A1)-(A6)$ and (PC). Let h_1, the smallest bandwidth, be such that the first estimator $\boldsymbol{e}_j^\top\widetilde{\boldsymbol{\theta}}_1(x)$ is always accepted by the adaptive procedure. Let $k^*(j)$ be the oracle index defined by (3.32). Then we have for the risk between the jth coordinates of the adaptive and oracle estimator:*

$$\left(\frac{nh^d_{k^*(j)}\Lambda_0}{\overline{\sigma}^2_{max}(k^*)}\right)^{r/2}\mathbb{E}|\boldsymbol{e}_j^\top\widetilde{\boldsymbol{\theta}}_{k^*(j)}(x) - \boldsymbol{e}_j^\top\widehat{\boldsymbol{\theta}}(x)|^r \tag{3.33}$$

$$\le\ \mathfrak{z}^{r/2}_{k^*(j)} + (\alpha\mathbb{E}|\chi^2_p|^r)^{1/2}(1+\delta)^{pk^*_j/4}(1-\delta)^{-3pk^*_j/4}\exp\left\{\varphi(\delta)\frac{\Delta_j}{2(1-\delta)}\right\},$$

where $\varphi(\delta)$ is as in Theorem 3.6.

Corollary 3.10. *For the polynomial basis under the conditions of the preceding theorem the risk between the adaptive and oracle estimators* $\mathbb{E}|\widetilde{f}^{(j-1)}_{k^*(j)}(x) - \widehat{f}^{(j-1)}(x)|^r$ *satisfies* (3.33).

Remark 3.8. The statements of this and the preceding proposition are of the same type that their vector counterparts. They are needed for asymptotical results of the last section.

Proof. To simplify the notation we suppress the dependence on j in the index k. Similarly to the proof of Theorem 3.3 we consider the disjunct events $\{\widehat{k} \le k^*\}$ and $\{\widehat{k} > k^*\}$. Therefore,

$$\begin{aligned} & \mathbb{E}|\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_{k^*}(x) - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}(x)|^r \\ = \; & \mathbb{E}|\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_{k^*}(x) - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}(x)|^r \, \mathbb{I}\{\widehat{k} \le k^*\} \\ + \; & \mathbb{E}|\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_{k^*}(x) - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}(x)|^r \, \mathbb{I}\{\widehat{k} > k^*\}. \end{aligned}$$

By Lemma 3.4 and the definition of the test statistic $T_{k^*,\widehat{k}}$ the second summand can be easily bounded:

$$\begin{aligned} & \left(\frac{nh^d_{k^*}\Lambda_0}{\overline{\sigma}^2_{max}(k^*)}\right)^{r/2} \mathbb{E}|\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_{k^*}(x) - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}(x)|^r \, \mathbb{I}\{\widehat{k} > k^*\} \\ \le \; & \mathbb{E}\|\mathbf{B}^{1/2}_{k^*}(\widetilde{\boldsymbol{\theta}}_{k^*}(x) - \widehat{\boldsymbol{\theta}}(x))\|^r \, \mathbb{I}\{\widehat{k} > k^*\} \\ \le \; & \mathfrak{z}^{r/2}_{k^*}. \end{aligned}$$

To bound the first summand we use the "componentwise" analog of Theorem 3.2, namely Theorem 3.6 that completes the proof. □

3.5. SMB and the bias-variance trade-off

It was shown in [36] that the small modeling bias ($SMB1$ here) condition given by (3.31) can be obtained from the "bias-variance trade-off" relations. Notice that our set-up includes the set-up from [36] as a particular ($p = 0$, $\delta = 0$, $\sigma(\cdot) \equiv \sigma$ is a known constant) case. To prove that the similar relation holds in the present case we need the following definition. Let the basis be polynomial. Given a point x and method of localization w, for any $j = 1, \dots, p$ the "ideal adaptive bandwidths", see [29] and [30], is defined as follows:

$$k^\star(j) = \max\{k \le K : \overline{b}_{k,f^{(j-1)}}(x) \le C_j(w)\sigma_{k,j}(x)\sqrt{d(n)}\}, \tag{3.34}$$

where $C_j(w)$ is a constant depending on the choice of the smoother w,

$$\begin{aligned} \overline{b}_{k,f^{(j-1)}}(x) &= \sup_{1\le l\le k} |\boldsymbol{e}_j^\top \boldsymbol{\theta}^*_l(x) - f^{(j-1)}(x)|, \\ \sigma^2_{k,j}(x) &= \mathrm{Var}_{\boldsymbol{f},\Sigma_0}[\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_k(x)], \\ d(n) &= \log(h_K/h_1), \end{aligned}$$

and $f^{(0)}$ stands for the function f itself. To bound the "modeling bias" $\Delta_j(k)$ we need the following assumption:

(A7) *There exists a constant $s_j > 0$ such that for all $k \le K$*

$$\boldsymbol{\Sigma}_{k,j}^{-1} \preceq s_j \boldsymbol{\Sigma}_{k,j,diag}^{-1} \tag{3.35}$$

where $\boldsymbol{\Sigma}_{k,j,diag} = \operatorname{diag}\big(\operatorname{Var}_{\boldsymbol{\theta},\Sigma}[\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_1(x)], \dots, \operatorname{Var}_{\boldsymbol{\theta},\Sigma}[\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_k(x)]\big)$ is a diagonal matrix composed of the diagonal elements of $\boldsymbol{\Sigma}_{k,j}$.

Remark 3.9. In order to understand the meaning and fulfillment of this assumption let us consider for simplicity the case of local constant fitting ($p = 0$). Then (3.35) can be rewritten as

$$\exists\, s > 0 : \boldsymbol{R}_k = \boldsymbol{\Sigma}_{k,diag}^{-1/2} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_{k,diag}^{-1/2} \succeq s^{-1} \boldsymbol{I}_k \quad \forall k \le K,$$

where $\boldsymbol{\Sigma}_k = (\operatorname{cov}[\widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_m])_{1 \le l \le m \le k}$ is a $k \times k$ positive definite matrix, $\boldsymbol{\Sigma}_{k,diag} = \operatorname{diag}(v_1, \dots, v_k)$ with $v_l = \operatorname{Var}_{\boldsymbol{\theta},\boldsymbol{\Sigma}}[\widetilde{\boldsymbol{\theta}}_l] > 0$, $l = 1, \dots, k$. We immediately see the following:

1. Since $\boldsymbol{R}_k \succ 0$, it is known that for any symmetric matrix $\boldsymbol{A}$ one can find a sufficiently small in absolute value real number τ s.t. $\boldsymbol{R}_k - \tau \boldsymbol{A} \succ 0$.
2. $\boldsymbol{R}_k = (\rho_{lm})_{1 \le l \le m \le k}$ is a correlation matrix with entries

$$\rho_{lm} = (v_l v_m)^{-1/2} \operatorname{cov}[\widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_m].$$

Moreover, $1 \ge \rho_{lm} > 0$ since for $w_{k,i} \in [0,1]$ the estimators are strictly positively correlated. Indeed,

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}_k &= \left(\sum_i \frac{w_{k,i}}{\sigma_i^2}\right)^{-1} \sum_i \frac{w_{k,i}}{\sigma_i^2} Y_i, \\ \operatorname{cov}[\widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_m] &= \left(\sum_i \frac{w_{l,i}}{\sigma_i^2}\right)^{-1} \left(\sum_i \frac{w_{m,i}}{\sigma_i^2}\right)^{-1} \sum_i \frac{w_{l,i} w_{m,i} \sigma_{0,i}^2}{\sigma_i^4} > 0. \end{aligned}$$

The strict inequality takes place because the estimators have a common support and therefore are dependent. Below we shall see that $(A7)$ essentially means that the estimators should not be correlated too strongly, which in its turn is provided by the assumption on the "geometrical growth of the scales", i.e. by $(A4)$. Indeed, since $\rho_{lm} > 0$, we have by direct calculations

$$(1 - \rho_{max}) \boldsymbol{I}_k \preceq \boldsymbol{R}_k \preceq (1 - \rho_{max}) \boldsymbol{I}_k + \rho_{max} \boldsymbol{J}_k,$$

where $\rho_{max} = \max_{1 \le l < m \le k} \{\rho_{lm}\}$ is the maximal correlation of the off-diagonal elements of $\boldsymbol{\Sigma}_k$ and $\boldsymbol{J}_k$ is a $k \times k$ matrix with all its elements equal to one. Thus we see that $s = (1 - \rho_{max})^{-1}$ explodes when the maximal correlation (except for the variations) is close to one.

3. Connection with $(A4)$. Assume that the weights $w_{l,i} = \mathbb{I}\{\|X_i - x\| \le h_l/2\}$. Then for $l < m$
$$\rho_{lm}^2 = \frac{\sum_i w_{l,i}\sigma_{0,i}^2\sigma_i^{-4}}{\sum_i w_{l,i}\sigma_{0,i}^2\sigma_i^{-4}} = \frac{\sum_i w_{l,i}}{\sum_i w_{m,i}}$$
for $\sigma_{0,i} = \sigma_i = /s$. Also we have $v_m/v_l = \mathbf{B}_l/\mathbf{B}_m = \rho_{lm}^2$. Since $u \ge u_0 > 1$, assumption $(A4)$ provides $0 < u^{-(m-l)/2} \le \rho_{lm} \le u_0^{-(m-l)/2} < 1$.

We have the following result:

Lemma 3.11. *Let the weights $\{w_{k,i}(x)\}$ satisfy (4.19) and the basis be polynomial $\{1, t-x, (t-x)^2/2!, \dots, (t-x)^{p-1}/(p-1)!\}$. Granted assumptions $(A1)$ – $(A4)$ and $(A7)$ for any (possibly fixed) n, any given point x, smoothing function w and $j = 1, \dots, p$ the choice of $k(j) = k^\star(j)$ defined by (3.34) with $d(n) = 1$ implies the $(SMBj)$ condition $\Delta_j(k(j)) \le \Delta_j$ with the constant $\Delta_j < 2s_jC_j^2(w)(1-u_0^{-1})^{-1} < \infty$.*

Proof. Consider the quantity $b_j(k)^\top \mathbf{\Sigma}_{k,j,diag}^{-1} b_j(k)$. For the polynomial basis $\boldsymbol{e}_j^\top \boldsymbol{\theta}(x) = f^{(j-1)}(x)$. In view of (4.19) the matrix $\mathbf{\Sigma}_{k,j,diag}$ is particularly simple:
$$\begin{aligned}\mathbf{\Sigma}_{k,j,diag} &= \operatorname{diag}(\boldsymbol{e}_j^\top \mathbf{B}_1^{-1}\boldsymbol{e}_j, \dots, \boldsymbol{e}_j^\top \mathbf{B}_k^{-1}\boldsymbol{e}_j) \\ &= \operatorname{diag}(\operatorname{Var}_{\boldsymbol{\theta},\Sigma}[\widetilde{\boldsymbol{\theta}}_1^{(j)}(x)], \dots, \operatorname{Var}_{\boldsymbol{\theta},\Sigma}[\widetilde{\boldsymbol{\theta}}_k^{(j)}(x)]),\end{aligned}$$
that is $\mathbf{\Sigma}_{k,j,diag}$ is a diagonal matrix of the variances of the jth coordinates of vectors $\widetilde{\boldsymbol{\theta}}_1, \dots, \widetilde{\boldsymbol{\theta}}_k$. Then by $(A4)$ and (3.2)
$$\begin{aligned} b_j(k)^\top \mathbf{\Sigma}_{k,j,diag}^{-1} b_j(k) &= \sum_{l=1}^k \frac{|\boldsymbol{e}_j^\top \boldsymbol{\theta}_l^* - f^{(j-1)}(x)|^2}{\boldsymbol{e}_j^\top \mathbf{B}_l^{-1}\boldsymbol{e}_j} \\ &\le \left(\overline{b}_{k,f^{(j-1)}}(x)\right)^2 \sum_{l=1}^k \frac{1}{\boldsymbol{e}_j^\top \mathbf{B}_l^{-1}\boldsymbol{e}_j} \\ &\le \frac{\left(\overline{b}_{k,f^{(j-1)}}(x)\right)^2}{\boldsymbol{e}_j^\top \mathbf{B}_k^{-1}\boldsymbol{e}_j} \sum_{l=1}^k u_0^{-(k-l)} \\ &\le \frac{\left(\overline{b}_{k,f^{(j-1)}}(x)\right)^2 (1+\delta)}{\sigma_{k,j}^2(x)(1-u_0^{-1})}. \end{aligned}$$
By (3.34) with $d(n) = 1$ the choice of $k = k^\star(j)$ implies $\left(\overline{b}_{k,f^{(j-1)}}(x)\right)^2 \le C_j^2(w)\sigma_{k,j}^2(x)$. Thus
$$b_j(k)^\top \mathbf{\Sigma}_{k,j,diag}^{-1} b_j(k) \le (1+\delta)C_j^2(w)(1-u_0^{-1})^{-1}$$
and
$$\begin{aligned}\Delta_j(k) &= b_j(k)^\top \mathbf{\Sigma}_{k,j}^{-1} b_j(k) \le s_jC_j^2(w)(1+\delta)(1-u_0^{-1})^{-1} \\ &< 2s_jC_j^2(w)(1-u_0^{-1})^{-1} < \infty,\end{aligned}$$
since u_0 from $(A4)$ is strictly larger than 1. □

Remark 3.10. The assumption that the weights $\{w_{k,i}(x)\}$ satisfy (4.19), that is that they are of the indicator-type, seems to bee too restrictive. This assumption allows to show the connection between the small modeling bias condition and the classical bias-variance trade-off without technical complications for any n, including the case of the fixed sample size. Relaxing of the consideration to the asymptotic case does not require such an assumption on the weights, see the lemmas below. Moreover, since this section essentially serves for checking the rate of convergence of the adaptive estimator at a point w.r.t. the Hölder classes of functions and since by $(A2)$ the windows are nested, to get the first impression it is enough to consider the design in $\mathbb{R}$, as in the case of the nested windows the generalization of the adaptive procedure to $\mathbb{R}^d$ is straightforward. On the contrary non-nested windows that are related to estimation on anisotropic classes require drastic modifications of the procedure, see [24] and [25].

Lemma 3.12. *Let the basis be polynomial and for each k the weight function $w_{k,\cdot}(x) = W((\cdot - x)h_k^{-1})$ be nonnegative, bounded with $\operatorname{supp} W(\cdot) \subset [0,1]$ and such that the Lebesgue measure of the set $\{u : W(u)^2 > 0\}$ is strictly positive. Let $X_i = i/n$, $i = 1, \dots, n$, and $h_k = h_k(n)$ be a sequence s.t. $h_k(n) \to 0$ and $nh_k(n) \to \infty$ as $n \to \infty$. Let the variance be either known ($\sigma_i \equiv \sigma_{0,i} = \sigma(\cdot)$) and continuous at the neighborhood of x, either the known "model" variance be locally bounded: i.e. $\exists\, 0 < \sigma_{min}(k) \le \sigma_{max}(k) < \infty$ s.t. $\sigma_{min}(k) \le \sigma_i \le \sigma_{max}(k)$ for $\forall\, i : w_{k,i}(x) > 0$. For a square matrix A by A_{diag} we denote a diagonal matrix with the same entries as the main diagonal of A. Then*

1. $$\boldsymbol{e}_j^\top \operatorname{Var}[\widetilde{\boldsymbol{\theta}}_k(x)]\boldsymbol{e}_l = O\Big(\frac{\sigma^2(x)}{nh_k^{j+l-1}}\Big) = O\Big(\boldsymbol{e}_j^\top \mathbf{B}_k^{-1}\boldsymbol{e}_l\Big),$$
   as $n \to \infty$;
2. *For n sufficiently large we have*
   $$\begin{aligned} & \sigma_{max}^{-2}(k)\operatorname{diag}(\mu_1(W), h_k^2\mu_2(W), \dots, h_k^{2(p-1)}\mu_{2(p-1)}(W)) \\ \lesssim\ & (nh_k)^{-1}(\mathbf{B}_k(x))_{diag} \\ \lesssim\ & \sigma_{min}^{-2}(k)\operatorname{diag}(\mu_1(W), h_k^2\mu_2(W), \dots, h_k^{2(p-1)}\mu_{2(p-1)}(W)) \end{aligned}$$
   with the moments of the kernel $W(\cdot)$ defined by
   $$\mu_\pi(W) = \int u^\pi W(u)\mathrm{d}u;$$
3. *By (3.1) $\operatorname{Var}\widetilde{\boldsymbol{\theta}}_k = \mathbf{B}_k^{-1}\widetilde{\mathbf{B}_k}\mathbf{B}_k^{-1}$, where $\widetilde{\mathbf{B}_k} = \boldsymbol{\Psi}\mathbf{W}_k\Sigma_0\mathbf{W}_k\boldsymbol{\Psi}^\top$ is a Gram matrix (c.f. Remark 2.2) and therefore the Hölder inequality is applicable to its off-diagonal elements. Since*
   $$(\widetilde{\mathbf{B}_k})_{diag} = \operatorname{diag}\Big(\sum_{i=1}^n \frac{w_{k,i}^2(x)}{\sigma_i^2}\frac{\sigma_{0,i}^2}{\sigma_i^2}, \dots, \sum_{i=1}^n \frac{(X_i - x)^{2(p-1)}}{((p-1)!)^2}\frac{w_{k,i}^2(x)}{\sigma_i^2}\frac{\sigma_{0,i}^2}{\sigma_i^2}\Big)$$

and assuming $(A3)$ *similarly to the statement 2 we have for* n *sufficiently large*

$$\begin{aligned}&(nh_k)^{-1}\boldsymbol{e}_j^\top\widetilde{\mathbf{B}_k}\boldsymbol{e}_j\\ \lesssim\ &\frac{1+\delta}{\sigma^2_{min}(k)}\operatorname{diag}(\mu_1(W^2),h_k^2\mu_2(W^2),\dots,h_k^{2(p-1)}\mu_{2(p-1)}(W^2))\end{aligned}$$

and the bounds for the variance of j*th coordinate of* $\widetilde{\boldsymbol{\theta}}_k$*:*

$$\frac{(1-\delta)\sigma^2_{min}(k)}{nh_k^{1+2(j-1)}}\lesssim\boldsymbol{e}_j^\top\operatorname{Var}[\widetilde{\boldsymbol{\theta}}_k]\boldsymbol{e}_j\lesssim\frac{(1+\delta)\sigma^2_{max}(k)}{nh_k^{1+2(j-1)}}.$$

That is $\boldsymbol{e}_j^\top\operatorname{Var}[\widetilde{\boldsymbol{\theta}}_k]\boldsymbol{e}_j=O(\boldsymbol{e}_j^\top\mathbf{B}_k^{-1}\boldsymbol{e}_j)$*. The constants depend on* $\sigma^2_{min}(k)$*,* $\sigma^2_{max}(k)$ *and the moments of* W *and* W^2*.*

Remark 3.11. When the constants are not the target in the study of convergence rate, the last display allows to substitute in the balance equation (3.34) the variance by the (j,j)th component of $\mathbf{B}_k^{-1}$, with the proviso that δ is "well behaved", c.f. Remark 3.4.

Proof. The statement of the lemma and its proof is essentially in the spirit of the Theorem 2.1 in [33] and Theorem 3.1 in [11], where the study was performed for the random design. □

Lemma 3.13. *Let for each* k *the weight function* $w_{k,\cdot}(x)=W((\cdot-x)h_k^{-1})$ *be nonnegative, bounded with* $\operatorname{supp}W(\cdot)\subset[0,1]$ *and such that the Lebesgue measure of the set* $\{u:W(u)>0\}$ *is strictly positive. Let* $X_i=i/n$*,* $i=1,\dots,n$*, and* $h_k=h_k(n)$ *be a sequence s.t.* $h_k(n)\to0$ *and* $nh_k(n)\to\infty$ *as* $n\to\infty$*. Let* $\Psi(u)=(1,u,\dots,u^{p-1}/(p-1)!)^\top$ *and* $\Psi_i\stackrel{\text{def}}{=}\Psi(i/n-x)$*.*

1. *Denote by* $\mathbf{B}_k^\sharp=\mathbf{B}_k^\sharp(x)=\boldsymbol{\Psi}\mathcal{W}_k\boldsymbol{\Psi}^\top=\sum_{i=1}^n\Psi_i\Psi_i^\top w_{k,i}$*. Then with* $H=\operatorname{diag}(1,h_k,\dots,h_k^{p-1})$ *we have*

$$(nh_k)^{-1}H^{-1}\mathbf{B}_k^\sharp H^{-1}\to\boldsymbol{B}=\int\Psi(u)\Psi^\top(u)W(u)\mathrm{d}u$$

as $n\to\infty$*, where the matrix* $\boldsymbol{B}$ *is positive definite and independent on* x *and* n*.*

2. *Moreover, assuming the known "model" variance be locally bounded: i.e.* $\exists\,0<\sigma_{min}(k)\le\sigma_{max}(k)<\infty$ *s.t.* $\sigma_{min}(k)\le\sigma_i\le\sigma_{max}(k)$ *for* $\forall i:$ $w_{k,i}(x)>0$ *we have for sufficiently large* n*:*

$$0\prec\sigma^{-2}_{max}(k)\boldsymbol{B}\preceq(nh_k)^{-1}H^{-1}\mathbf{B}_kH^{-1}\preceq\sigma^{-2}_{min}(k)\boldsymbol{B}.$$

Proof. The first statement of the lemma is based on the convergence of Riemann sums. The non-degenerateness of $\boldsymbol{B}$ is the Lemma 1.4 in [41] and follows from the fact that the polynomials of degree $\le p-1$ have at most $p-1$ different zeros.

To justify the second statement it is enough to remark that $\sigma_{max}^{-2}(k)\mathbf{B}_k^{\sharp} \preceq \mathbf{B}_k \preceq \sigma_{min}^{-2}(k)\mathbf{B}_k^{\sharp}$ and that the first statement implies $(nh_k)^{-1}\gamma^\top H^{-1}\mathbf{B}_k^{\sharp}H^{-1}\gamma \to \gamma^\top \boldsymbol{B}\gamma$ for any vector γ. □

Remark 3.12. Using the standard technique it is easy to derive from the above result that for estimation of functions over Hölder classes the methodology proposed in [23] and [36] and generalized in the present paper delivers the minimax rate of convergence up to a logarithmic factor, see the following subsection for details.

3.6. Rates of convergence

At this section $d = 1$ and the basis is polynomial with the columns of the design matrix $\boldsymbol{\Psi}$ given by

$$\Psi_i = \Psi(X_i - x) = \left(1,\, X_i - x, \ldots, (X_i - x)^{p-1}/(p-1)!\right)^\top.$$

. The polynomial weights $W_{l,\,i}^*$ are given by

$$W_{l,\,i}^*(x) = \boldsymbol{e}_1^\top \mathbf{B}_l^{-1}\Psi_i w_{l,i}(x)/\sigma_i^2 \tag{3.36}$$

with $\mathbf{B}_l \succ 0$ defined by (2.5) and the variance term given by $\sigma_l^2(x) \stackrel{\text{def}}{=} \mathbb{E}_f[|\boldsymbol{e}_1^\top \widetilde{\boldsymbol{\theta}}_l(x) - \boldsymbol{e}_1^\top \boldsymbol{\theta}_l^*(x)|^2]$. Here

$$\boldsymbol{e}_1^\top \boldsymbol{\theta}_l^*(x) = \mathbb{E}_f[\widetilde{f}_l(x)] = \sum_{i=1}^n W_{l,\,i}^*(x) f(X_i)$$

is a local linear smoother of the function f at the point x corresponding to lth scale. Define the "monotonized" bias by

$$\overline{b}_{k,f}(x) = \sup_{1 \le l \le k} |\boldsymbol{e}_1^\top \boldsymbol{\theta}_l^*(x) - f(x)|. \tag{3.37}$$

Before proceeding with analysis of the convergence rate we need to derive bounds for the bias and variance.

(*A*8) Let the known "model" variance be locally bounded: i.e. $\exists\, 0 < \sigma_{min}(k) \le \sigma_{max}(k) < \infty$ s.t. $\sigma_{min}(k) \le \sigma_i \le \sigma_{max}(k)$ for $\forall\, i : w_{k,i}(x) > 0$.

(*A*9) *There exists a real number $a_0 > 0$ such that for any interval $A \subseteq [0,1]$ and all $n \ge 1$*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i \in A\} \le a_0 \max\left\{\int_A \mathrm{d}t, \frac{1}{n}\right\}.$$

(*A*10) *The localizing functions (kernels) $w_{k,i}$ have compact supports belonging to $[0,1]$:*

$$w_{k,i}(x) = 0 \quad \text{if} \quad |X_i - x| > h_k.$$

This immediately implies the similar property for the local polynomial weights:

$$W^*_{k,i}(x) = 0 \quad \text{if} \quad |X_i - x| > h_k.$$

(A11) *There exists a finite number w_{max} such that*

$$\sup_{k,i} |w_{k,i}(x)| \le w_{max}.$$

Remark 3.13. Assumption $(A3)$ implies that the conditional number

$$\varkappa(\Sigma) \stackrel{\text{def}}{=} \frac{\sigma^2_{max}}{\sigma^2_{min}} \tag{3.38}$$

of covariance matrix from the misspecified model (1.4) is finite.

Lemma 3.14. *Assume $(A1)-(A3)$, $(A6)$ and $(A8)-(A11)$. Let h'_1 be the smallest bandwidth providing $(A6)$ and h''_1 be the smallest bandwidth s.t. the first estimator $\widetilde{\boldsymbol{\theta}}_1$ is accepted by the adaptive procedure. Denote by $h_1 \ge \max\{1/(2n), h'_1, h''_1\}$. Let the regression function $f(\cdot)$ belong to the Hölder class $\Sigma(\beta, L)$ on $[0,1]$, and let $\{\widetilde{f}_k(x)\}_{k=1}^K$ be the $LP_k(p-1)$ estimators of $f(x)$ with $p-1 = \lfloor\beta\rfloor$. Then for sufficiently large n and any h_k satisfying $h_K > \ldots > h_k > \ldots > h_1$, $k = 1, \ldots, K$, we have*

$$\begin{aligned} |\overline{b}_{k,f}(x)| &\le C_2 \varkappa(\Sigma) \frac{L h_k^\beta}{(p-1)!}, \\ \sigma^2_k(x) &\le (1+\delta) \frac{\sigma^2_{max}}{n h_k \Lambda_0} \end{aligned}$$

with $C_2 = 2 w_{max} a_0 \sqrt{e}/\Lambda_0$ and $\delta \in [0,1)$ from $(A3)$.

The proof is moved to Appendix.

Proposition 3.15. *Let the model (1.4) be satisfied. Assume $(A1)-(A4)$, $(A6)$ and $(A8)-(A11)$. Let h'_1 be the smallest bandwidth providing $(A6)$ and h''_1 be the smallest bandwidth s.t. the first estimator $\widetilde{\boldsymbol{\theta}}_1$ is accepted by the adaptive procedure. Denote by $h_1 \ge \max\{1/(2n), h'_1, h''_1\}$. Let the regression function $f(\cdot)$ belong to the Hölder class $\Sigma(\beta, L)$ on $[0,1]$ and let $\{\widetilde{f}_k(x)\}_{k=1}^K$ be the $LP_k(p-1)$ estimators of $f(x)$ with $p-1 = \lfloor\beta\rfloor$. Then for sufficiently large n for adaptive estimator obtained by the procedure we have*

$$\mathbb{E}|\widehat{f}(x) - f(x)|^r \asymp \left(\frac{\log n}{n}\right)^{\frac{\beta r}{2\beta+1}}.$$

Proof. If the model (1.4) is true, then $\Delta = 0$ and one can take k^* from (3.34) with $d(n) = \log n$ leading, in view of the preceding lemma, to the choice of the optimal bandwidth $h_{k^*}(x)$ of order $(\log n/n)^{\frac{1}{2\beta+1}}$. The oracle bound of Proposition 3.9 gives

$$\mathbb{E}|\widehat{f}(x) - \widetilde{f}_{k^*}(x)|^r \lesssim \left(\frac{\log n}{n h_{k^*(x)}}\right)^{r/2}.$$

Since k^* is an unknown but deterministic, $\widetilde{f}_{k^*}(x)$ is a standard local polynomial estimator. Therefore its quality of estimation is known:

$$\mathbb{E}|f(x) - \widetilde{f}_{k^*}(x)|^r \lesssim (1/n)^{\frac{\beta r}{2\beta+1}} \ll (\log n/n)^{\frac{\beta r}{2\beta+1}}$$

and the assertion follows by application of $(a+b)^r \le C(r)(a^r + b^r)$, $a, b \ge 0$, where the constant $C(r) = 2^{r-1}$ for $r \ge 1$ and is equal to one for $0 < r < 1$.

The rate $\left(\frac{\log n}{n}\right)^{\frac{\beta r}{2\beta+1}}$ is known to be optimal, c.f. [27]. □

Proposition 3.16. *Assume* $(A1)-(A4)$, $(A6)-(A11)$ *and* $\delta = O(1/\log n)$. *Let* h_1' *be the smallest bandwidth providing* $(A6)$ *and* h_1'' *be the smallest bandwidth s.t. the first estimator* $\widetilde{\boldsymbol{\theta}}_1$ *is accepted by the adaptive procedure. Denote by* $h_1 \ge \max\{1/(2n), h_1', h_1''\}$. *Let the regression function* $f(\cdot)$ *belong to the Hölder class* $\Sigma(\beta, L)$ *on* $[0,1]$, *and let* $\{\widetilde{f}_k(x)\}_{k=1}^K$ *be the* $LP_k(p-1)$ *estimators of* $f(x)$ *with* $p - 1 = \lfloor \beta \rfloor$. *Then for sufficiently large* n *for the adaptive estimator delivered by the procedure we have*

$$\mathbb{E}|\widehat{f}(x) - f(x)|^r \lesssim \left(\frac{\log^\gamma n}{n}\right)^{\frac{\beta r}{2\beta+1}}$$

with $\gamma = (2\beta+1)/(2\beta)$.

Proof. Because now we need to have (SMB) fulfilled, we must take k^* from (3.34) with $d(n) = 1$ leading to the suboptimal choice of $h_{k^*}(x)$ of order $(1/n)^{\frac{1}{2\beta+1}}$ and the assertion follows. □

4. Appendix

4.1. Pivotality and local parametric risk bounds

Lemma 4.1 (Pivotality property)**.** *Let* $(A2)$ *hold. Let* $\boldsymbol{\theta}_1^* = \cdots = \boldsymbol{\theta}_{\varkappa}^* = \boldsymbol{\theta}$ *for some* $\varkappa \le K$. *Then for any* $k \le \varkappa$ *the risk associated with the adaptive estimate at every step of the procedure does not depend on the parameter* $\boldsymbol{\theta}$*:*

$$\mathbb{E}_{\boldsymbol{\theta}}|(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r \quad = \quad \mathbb{E}_0|(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r,$$

where $\mathbb{E}_0$ *denotes the expectation w.r.t. the centered measures* $\mathcal{N}(0, \Sigma)$ *or* $\mathcal{N}(0, \Sigma_0)$.

Proof. At each step k of the procedure the adaptive estimator $\widehat{\boldsymbol{\theta}}_k$ coincides with one of the nonadaptive estimators $\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_k$. If $\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_k$, this means that the deviation from the parametric model is not significant and the procedure passes to the next step. On the contrary, $\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m$ for $m < k$ means that for some $l \le m$ the value of the test statistic $T_{l,\,m+1}$ is strictly larger than the threshold $\mathfrak{z}_l$ and

the procedure had terminated. Thus one can write the following decomposition:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}}|(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r \\
= &\sum_{m=1}^{k} \mathbb{E}_{\boldsymbol{\theta}} \|\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)\|^{2r} \mathbb{I}\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\} \\
= &\sum_{m=1}^{k-1} \mathbb{E}_{\boldsymbol{\theta}} \|\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)\|^{2r} \mathbb{I}\{\exists\, l \le m : \|\mathbf{B}_l^{1/2}(\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_{m+1})\|^2 > \mathfrak{z}_l\}.
\end{aligned}$$

In the last line the definition of $T_{l,\,m+1}$ given by (2.13) is used. Since for any $k \le \varkappa$ under the assumptions of lemma $\widetilde{\boldsymbol{\theta}}_k = \boldsymbol{\theta} + \mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k \Sigma_0^{1/2} \boldsymbol{\varepsilon}$, the value of $\boldsymbol{\theta}$ cancels in the differences $\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m$ and $\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_{m+1}$ for all $l \le m < k$, and therefore can be taken equal to zero. □

To justify the statistical properties of the considered procedure we need the following simple observation. Let for any $\boldsymbol{\theta}$, $\boldsymbol{\theta}' \in \Theta$ the corresponding log-likelihood ratio $\mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}, \boldsymbol{\theta}')$ be defined by (2.12). Then

$$2\,\mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}, \boldsymbol{\theta}') = \|\mathbf{W}_k^{1/2}(\boldsymbol{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}')\|^2 - \|\mathbf{W}_k^{1/2}(\boldsymbol{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta})\|^2.$$

Lemma 4.2 (Quadratic shape of the fitted log-likelihood)**.** *Let for every $k = 1, \ldots, K$ the fitted log likelihood (FLL) be defined as follows:*

$$\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}') \stackrel{\mathrm{def}}{=} \max_{\boldsymbol{\theta} \in \Theta} \mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}, \boldsymbol{\theta}').$$

Then

$$2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}) = (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}). \tag{4.1}$$

Proof. Notice that $\mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta})$ defined by (2.4) is quadratic in $\boldsymbol{\theta}$. The assertion follows from the second order Taylor series expansion around the point $\widetilde{\boldsymbol{\theta}}_k$, because it is the point of maximum, and the second derivative is the constant matrix $\mathbf{B}_k$. □

Let the matrix $\mathbf{S}$ be defined as follows:

$$\mathbf{S} \stackrel{\mathrm{def}}{=} \Sigma_0^{1/2} \mathbf{W}_k \boldsymbol{\Psi}^\top \mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k \Sigma_0^{1/2}. \tag{4.2}$$

Then for the distribution of $\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)$ one observes so-called "Wilks phenomenon", c.f. [12], described by the following theorem:

Proposition 4.3. *Let the regression model be given by (1.1) and the parameter $\boldsymbol{\theta}_k^* = \boldsymbol{\theta}_k^*(x)$ maximizing the expected local log-likelihood be defined by (2.7). Then for any $k = 1, \ldots, K$ the following equality in distribution takes place:*

$$2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) \stackrel{\mathrm{d}}{=} \lambda_1(\mathbf{S}) \overline{\varepsilon}_1^2 + \cdots + \lambda_p(\mathbf{S}) \overline{\varepsilon}_p^2 \tag{4.3}$$

with $p = \mathrm{rank}(\mathbf{B}_k) = \dim \Theta = p$. Here $\lambda_1(\mathbf{S}), \ldots, \lambda_p(\mathbf{S})$ are the non-zero eigenvalues of the matrix $\mathbf{S}$, and $\overline{\varepsilon}_i$ are independent standard normal random variables.

Moreover, under $(A3)$ *the maximal eigenvalue* $\lambda_{max}(\mathbf{S}) \le 1+\delta$, *and for any* $\mathfrak{z} > 0$

$$\mathbb{P}\left\{2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) \ge \mathfrak{z}\right\} \le \mathbb{P}\left\{\eta \ge \mathfrak{z}/(1+\delta)\right\}, \tag{4.4}$$

where η *is a random variable distributed according to the* χ^2 *law with* p *degrees of freedom.*

Remark 4.1. Generally, if $\mathbf{B}_k$ is degenerated, the number of terms in (4.3) is $p \le \dim\Theta$.

Proof. By Lemma 4.2 and the decomposition (2.9) it holds that:

$$\begin{aligned} 2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) &= (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) \\ &= (\mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0^{1/2}\boldsymbol{\varepsilon})^\top \mathbf{B}_k (\mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0^{1/2}\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\varepsilon}^\top \mathbf{S}\boldsymbol{\varepsilon}, \end{aligned}$$

where the symmetric matrix $\mathbf{S}$ is defined by (4.2). Then by the Schur theorem there exist an orthogonal matrix $\mathbf{M}$ and the diagonal matrix $\boldsymbol{\Lambda}$ composed of the eigenvalues of $\mathbf{S}$ such that $\mathbf{S} = \mathbf{M}^\top\boldsymbol{\Lambda}\mathbf{M}$. For $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$ and an orthogonal matrix $\mathbf{M}$ it holds that $\overline{\boldsymbol{\varepsilon}} \stackrel{\mathrm{def}}{=} \mathbf{M}\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$. Indeed, $\mathbb{E}\mathbf{M}\boldsymbol{\varepsilon} = \mathbb{E}\boldsymbol{\varepsilon} = 0$ and

$$\operatorname{Var}\mathbf{M}\boldsymbol{\varepsilon} = \mathbb{E}\mathbf{M}\boldsymbol{\varepsilon}(\mathbf{M}\boldsymbol{\varepsilon})^\top = \mathbf{M}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\mathbf{M}^\top = \mathbf{M}\mathbf{M}^\top = I_n.$$

Therefore,

$$2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) \stackrel{\mathrm{d}}{=} \overline{\boldsymbol{\varepsilon}}^\top\boldsymbol{\Lambda}\overline{\boldsymbol{\varepsilon}}\,, \quad \overline{\boldsymbol{\varepsilon}} \sim \mathcal{N}(0, I_n)\,.$$

On the other hand, the matrix $\mathbf{S}$ can be written as $\mathbf{S} = \Sigma_0^{1/2}\mathbf{W}_k^{1/2}\boldsymbol{\Pi}_k\mathbf{W}_k^{1/2}\Sigma_0^{1/2}$ with $\boldsymbol{\Pi}_k = \mathbf{W}_k^{1/2}\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k^{1/2}$. Since $\boldsymbol{\Pi}_k$ is symmetric and idempotent, i.e. $\boldsymbol{\Pi}_k^2 = \boldsymbol{\Pi}_k$, it is an orthogonal projector on the linear subspace of dimension $p = \operatorname{rank}(\mathbf{B}_k)$ spanned by the rows of $\boldsymbol{\Psi}$. Moreover, $\operatorname{rank}(\boldsymbol{\Pi}_k) = \operatorname{tr}(\boldsymbol{\Pi}_k) = \operatorname{tr}(\mathbf{W}_k^{1/2}\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k^{1/2}) = \operatorname{tr}(\mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\boldsymbol{\Psi}^\top) = \operatorname{tr}(\mathbf{B}_k^{-1}\mathbf{B}_k) = \operatorname{tr}(I_p) = p$. Therefore $\boldsymbol{\Pi}_k$ has only p unit eigenvalues and $n-p$ zero ones. Notice also that the $n \times n$ matrix $\mathbf{S}$ has $\operatorname{rank}(\mathbf{S}) = \operatorname{rank}(\boldsymbol{\Pi}_k\mathbf{W}_k^{1/2}\Sigma_0^{1/2}) = \operatorname{rank}(\boldsymbol{\Pi}_k) = p$ as well. Thus $2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) \stackrel{\mathrm{d}}{=} \lambda_1(\mathbf{S})\overline{\varepsilon}_1^2 + \cdots + \lambda_p(\mathbf{S})\overline{\varepsilon}_p^2$, where $\lambda_1(\mathbf{S}), \ldots, \lambda_p(\mathbf{S})$ are the non-zero eigenvalues of the matrix $\mathbf{S}$.

Recall the definition of the matrix norm induced by the L_2 vector norm:

$$\|A\|_{2,in} \stackrel{\mathrm{def}}{=} \sqrt{\lambda_{max}(A^\top A)}. \tag{4.5}$$

Assumption $(A3)$ allows to bound the induced L_2-norm of the matrix $\mathbf{S}$:

$$\begin{aligned} \|\mathbf{S}\|_{2,in} &= \|\Sigma_0^{1/2}\mathbf{W}_k^{1/2}\boldsymbol{\Pi}_k\mathbf{W}_k^{1/2}\Sigma_0^{1/2}\|_{2,in} \\ &\le \|\Sigma_0^{1/2}\mathbf{W}_k^{1/2}\|_{2,in}\|\boldsymbol{\Pi}_k\|_{2,in}\|\mathbf{W}_k^{1/2}\Sigma_0^{1/2}\|_{2,in} \\ &= \lambda_{max}(\mathbf{W}_k\Sigma_0)\lambda_{max}(\boldsymbol{\Pi}_k) \\ &= \max_i\{w_{k,i}\frac{\sigma_{0,i}^2}{\sigma_i^2}\} \\ &\le (1+\delta)\max_i\{w_{k,i}\} \le 1+\delta. \end{aligned}$$

Therefore, the largest eigenvalue of matrix $\mathbf{S}$ is bounded: $\lambda_{max}(\mathbf{S}) \leq 1+\delta$.

$$\mathbb{P}\left\{\lambda_1(\mathbf{S})\overline{\varepsilon}_1^2 + \cdots + \lambda_p(\mathbf{S})\overline{\varepsilon}_p^2 \geq \mathfrak{z}\right\} \leq \mathbb{P}\left\{\lambda_{max}(\mathbf{S})(\overline{\varepsilon}_1^2 + \cdots + \overline{\varepsilon}_p^2) \geq \mathfrak{z}\right\}$$

provides the last assertion. □

Corollary 4.4 (Quasi-parametric risk bounds)**.** *Let the model be given by* (1.1) *and* $\boldsymbol{\theta}_k^* = \boldsymbol{\theta}_k^*(x)$ *be defined by* (2.7)*. Assume* (*A*3)*. Then for any* $\mu < 1/(1+\delta)$ *we have*

$$\mathbb{E}\exp\{\mu\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)\} \leq [1-\mu(1+\delta)]^{-p/2}, \tag{4.6}$$

$$\mathbb{E}|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)|^r \leq (1+\delta)^r C(p,r)\,, \tag{4.7}$$

where

$$C(p,r) = \mathbb{E}|\chi_p^2|^r = 2^r\Gamma(r+p/2)/\Gamma(p/2). \tag{4.8}$$

Proof. By (4.3) and independence of $\overline{\varepsilon}_i$

$$\begin{aligned}
\mathbb{E}\exp\{\mu\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)\} &= \mathbb{E}\exp\left\{\frac{\mu}{2}\sum_{i=1}^p \lambda_i(\mathbf{S})\overline{\varepsilon}_i^2\right\} \\
&= \prod_{i=1}^p \mathbb{E}\exp\left\{\frac{\mu}{2}\lambda_i(\mathbf{S})\overline{\varepsilon}_i^2\right\} \\
&= \prod_{i=1}^p [1-\mu\,\lambda_i(\mathbf{S})]^{-1/2} \\
&\leq [1-\mu\,\lambda_{max}(\mathbf{S})]^{-p/2} \\
&\leq [1-\mu(1+\delta)]^{-p/2}.
\end{aligned}$$

Let $\eta \sim \chi_p^2$. Integration by parts yields the second inequality:

$$\begin{aligned}
\mathbb{E}|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)|^r &= \int_0^\infty \mathbb{P}\left\{2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) \geq \mathfrak{z}\right\} r\mathfrak{z}^{r-1}\mathrm{d}\mathfrak{z} \\
&\leq r\int_0^\infty \mathbb{P}\left\{\eta \geq \mathfrak{z}/(1+\delta)\right\}\mathfrak{z}^{r-1}\mathrm{d}\mathfrak{z} \\
&= (1+\delta)^r\,\mathbb{E}|\eta|^r.
\end{aligned}$$

□

4.2. Proof of the bounds for the critical values

Denote for any $l < k$ the variance of difference $\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_l$ by V_{lk}:

$$V_{lk} \stackrel{\mathrm{def}}{=} \mathrm{Var}(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_l) \succ 0. \tag{4.9}$$

Then there exists a unique matrix $V_{lk}^{1/2} \succ 0$ such that $(V_{lk}^{1/2})^2 = V_{lk}$.

Lemma 4.5. *Assume* $(A1)-(A4)$. *If* $\boldsymbol{\theta}_1^* = \cdots = \boldsymbol{\theta}_k^* = \boldsymbol{\theta}$ *for* $k \le K$, *then for any* $l < k$ *we have*

$$\begin{aligned}
\mathbb{P}\left\{2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k) \ge \mathfrak{z}\right\} &\le \mathbb{P}\left\{\eta \ge \mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\right\} \\
&\le \mathbb{P}\left\{\eta \ge \mathfrak{z}/t_0\right\}, \\
\mathbb{P}\left\{2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_l) \ge \mathfrak{z}\right\} &\le \mathbb{P}\left\{\eta \ge \mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})\right\} \\
&\le \mathbb{P}\left\{\eta \ge \mathfrak{z}/t_1\right\},
\end{aligned}$$

where $t_0 = 2(1+\delta)(1+u_0^{-(k-l)})$, $t_1 = 2(1+\delta)(1+u^{(k-l)})$, $1 < u_0 \le u$ *are the constants from the assumption* $(A4)$ *and* η *is a* χ^2_p*-distributed random variable.*

Proof. Decomposition (2.9) of $\widetilde{\boldsymbol{\theta}}_k$ into deterministic $\boldsymbol{\theta}_k^*$ and stochastic parts and the assumption of lemma imply

$$\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_k = \mathbf{B}_l^{-1}\boldsymbol{\Psi}\mathbf{W}_l\Sigma_0^{1/2}\boldsymbol{\varepsilon} - \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0^{1/2}\boldsymbol{\varepsilon} \stackrel{\mathrm{d}}{=} V_{lk}^{1/2}\xi,$$

where ξ is a standard normal vector in $\mathbb{R}^p$. Thus by Lemma 4.2 for any $l < k$

$$2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k) = \|\mathbf{B}_l^{1/2}(\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_k)\|^2 \stackrel{\mathrm{d}}{=} \xi^\top V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}\xi.$$

By the Schur theorem there exists an orthogonal matrix M such that

$$\xi^\top V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}\xi \stackrel{\mathrm{d}}{=} \overline{\varepsilon}^\top M^\top \Lambda_{lk} M\overline{\varepsilon},$$

where $\overline{\varepsilon}$ is a standard normal vector,

$$\Lambda_{lk} = \mathrm{diag}(\lambda_1(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})), \ldots, \lambda_p(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}))$$

and $p = \mathrm{rank}(\mathbf{B}_l)$. Therefore,

$$2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k) \stackrel{\mathrm{d}}{=} \lambda_1(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\overline{\varepsilon}_1^2 + \cdots + \lambda_p(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\overline{\varepsilon}_p^2,$$

where $\lambda_j(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})$, $j = 1, \ldots, p$, are the nonzero eigenvalues of $V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}$. Similarly,

$$2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_l) \stackrel{\mathrm{d}}{=} \lambda_1(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})\overline{\varepsilon}_1^2 + \cdots + \lambda_p(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})\overline{\varepsilon}_p^2.$$

Denoting by η a χ^2_p-distributed random variable we get

$$\begin{aligned}
\mathbb{P}\left\{2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k) \ge \mathfrak{z}\right\} &\le \mathbb{P}\left\{\eta \ge \mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\right\}, \\
\mathbb{P}\left\{2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_l) \ge \mathfrak{z}\right\} &\le \mathbb{P}\left\{\eta \ge \mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})\right\}.
\end{aligned}$$

For any square matrices A and B we have $(A-B)(A^\top - B^\top) \preceq 2(AA^\top + BB^\top)$. Applying this bound to the variance of the difference of estimators we obtain

$$\begin{aligned}
V_{lk} &= \left(\mathbf{B}_l^{-1}\boldsymbol{\Psi}\mathbf{W}_l\Sigma_0^{1/2} - \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0^{1/2}\right)\left(\mathbf{B}_l^{-1}\boldsymbol{\Psi}\mathbf{W}_l\Sigma_0^{1/2} - \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0^{1/2}\right)^\top \\
&\preceq 2(\mathbf{B}_l^{-1}\boldsymbol{\Psi}\mathbf{W}_l\Sigma_0\mathbf{W}_l\boldsymbol{\Psi}^\top\mathbf{B}_l^{-1} + \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0\mathbf{W}_k\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}) \\
&= 2V_l + 2V_k,
\end{aligned}$$

where $V_l = \operatorname{Var}\widetilde{\boldsymbol{\theta}}_l$, $l \le k$. By (3.2) and Assumption $(A4)$ we have

$$
\begin{aligned}
&V_l \preceq (1+\delta)\mathbf{B}_l^{-1},\\
&V_k \preceq (1+\delta)\mathbf{B}_k^{-1} \preceq (1+\delta)u_0^{-(k-l)}\mathbf{B}_l^{-1},\\
&V_{lk} \preceq 2(1+\delta)(1+u_0^{-(k-l)})\mathbf{B}_l^{-1}.
\end{aligned}
$$

Therefore,

$$
\mathbf{B}_l \preceq 2(1+\delta)(1+u_0^{-(k-l)})V_{lk}^{-1}. \tag{4.10}
$$

This provides the following bound:

$$
\begin{aligned}
\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}) &= \sup_{\|\gamma\|=1} \gamma^\top V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}\gamma\\
&\le 2(1+\delta)(1+u_0^{-(k-l)}).
\end{aligned} \tag{4.11}
$$

Similarly,

$$
\begin{aligned}
V_{lk} &\preceq 2(1+\delta)(1+u^{(k-l)})\mathbf{B}_k^{-1},\\
\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2}) &\le 2(1+\delta)(1+u^{(k-l)}).
\end{aligned} \tag{4.12}
$$

These bounds imply

$$
\begin{aligned}
\mathbb{P}\left\{2\,\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\widetilde{\boldsymbol{\theta}}_k)\ge \mathfrak{z}\right\} &\le \mathbb{P}\left\{\eta \ge \mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\right\}\\
&\le \mathbb{P}\left\{\eta \ge \mathfrak{z}\left[2(1+\delta)(1+u_0^{-(k-l)})\right]^{-1}\right\}\\
\mathbb{P}\left\{2\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)\ge \mathfrak{z}\right\} &\le \mathbb{P}\left\{\eta \ge \mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})\right\}\\
&\le \mathbb{P}\left\{\eta \ge \mathfrak{z}\left[2(1+\delta)(1+u^{(k-l)})\right]^{-1}\right\}
\end{aligned}
$$

□

Lemma 4.6. *Under the conditions of preceding lemma for any $l<k$, $\mu_0 < t_0^{-1}$, $\mu_1 < t_1^{-1}$ we have*

$$
\begin{aligned}
\mathbb{E}\exp\{\mu_0\,\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\widetilde{\boldsymbol{\theta}}_k)\} &\le [1-\mu_0 t_0]^{-p/2},\\
\mathbb{E}\exp\{\mu_1\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)\} &\le [1-\mu_1 t_1]^{-p/2},
\end{aligned}
$$

where $t_0 = 2(1+\delta)(1+u_0^{-(k-l)})$, $t_1 = 2(1+\delta)(1+u^{(k-l)})$ and the constants $1<u_0\le u$ are from Assumption $(A4)$.

Proof. The statement of the lemma is justified similarly to the proof of Corollary 4.4. The bounds (4.11) and (4.12) imply the bounds for the corresponding

moment generating functions:

$$\begin{aligned}
\mathbb{E}\exp\{\mu\,\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\widetilde{\boldsymbol{\theta}}_k)\} &= \prod_{j=1}^{p}\mathbb{E}\exp\{\frac{\mu}{2}\lambda_j(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\overline{\varepsilon}_j^2\}\\
&= \prod_{j=1}^{p}[1-\mu\lambda_j(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})]^{-1/2}\\
&\le [1-\mu\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})]^{-p/2}\\
&\le [1-2\mu(1+\delta)(1+u_0^{-(k-l)})]^{-p/2}\,,\\
\mathbb{E}\exp\{\mu\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)\} &\le [1-\mu\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})]^{-p/2}\\
&\le [1-2\mu(1+\delta)(1+u^{(k-l)})]^{-p/2}.
\end{aligned}$$

□

Lemma 4.7. *Under the conditions of Lemma 4.5 for any $l<k$ we have*

$$\begin{aligned}
\mathbb{E}|2\,\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\widetilde{\boldsymbol{\theta}}_k)|^r &\le 2^r C(p,r)(1+\delta)^r(1+u_0^{-(k-l)})^r,\\
\mathbb{E}|2\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)|^r &\le 2^r C(p,r)(1+\delta)^r(1+u^{(k-l)})^r,
\end{aligned}$$

where $C(p,r)$ is given by (4.8).

Remark 4.2. The RHS's of Lemmas 4.6 and 4.7 are highly asymmetric. Recall that here $\boldsymbol{\theta}_1^*=\cdots=\boldsymbol{\theta}_k^*=\boldsymbol{\theta}$, $l<k$ and $1<u_0\le u$. The bounds for the log-likelihood ratio corresponding to the l-th scale $\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\widetilde{\boldsymbol{\theta}}_k)$ are close to the bounds for their parametric counterpart $\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\boldsymbol{\theta})$ given by Corollary 4.4. It is not surprising because, if the parametric model is satisfied up to the scale k, for the MLE $\widetilde{\boldsymbol{\theta}}_k$ more data were used and the estimator $\widetilde{\boldsymbol{\theta}}_k$ w.r.t. $\widetilde{\boldsymbol{\theta}}_l$ acts approximately as the true parameter $\boldsymbol{\theta}$. On the contrary, the risk bounds for $\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)$ are quite large since for the larger k-th scale $\widetilde{\boldsymbol{\theta}}_l$ is a bad estimator with large variance.

Proof. Integration by parts and Lemma 4.5 yield for the second assertion

$$\begin{aligned}
\mathbb{E}|2\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)|^r &= r\int_0^\infty \mathbb{P}\left\{2\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)\ge\mathfrak{z}\right\}\mathfrak{z}^{r-1}\mathrm{d}\mathfrak{z}\\
&\le r\int_0^\infty \mathbb{P}\left\{\eta\ge\mathfrak{z}\left[2(1+\delta)(1+u^{(k-l)})\right]^{-1}\right\}\mathfrak{z}^{r-1}\mathrm{d}\mathfrak{z}\\
&= 2^r(1+\delta)^r(1+u^{(k-l)})^r\mathbb{E}|\eta|^r,
\end{aligned}$$

where $\eta\sim\chi_p^2$. The first assertion is proved similarly. □

Proof of Theorem 3.1 Theoretical choice of the critical values. The risk corresponding to the adaptive estimate can be represented as a sum of risks of the false alarms at each step of the procedure:

$$\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_k-\widehat{\boldsymbol{\theta}}_k)^\top\mathbf{B}_k(\widetilde{\boldsymbol{\theta}}_k-\widehat{\boldsymbol{\theta}}_k)|^r=\sum_{m=1}^{k-1}\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_k-\widetilde{\boldsymbol{\theta}}_m)^\top\mathbf{B}_k(\widetilde{\boldsymbol{\theta}}_k-\widetilde{\boldsymbol{\theta}}_m)|^r\mathbb{I}\{\widehat{\boldsymbol{\theta}}_k=\widetilde{\boldsymbol{\theta}}_m\}. \tag{4.13}$$

By the definition of the last accepted estimate $\widehat{\boldsymbol{\theta}}_k$, for any $m = 1, \ldots, k-1$, the event $\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\}$ happens if for some $l = 1, \ldots, m$ the statistic $T_{l,m+1} > \mathfrak{z}_l$. Thus

$$\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\} \subseteq \bigcup_{l=1}^{m} \{T_{l,m+1} > \mathfrak{z}_l\}.$$

It holds also that for any positive μ

$$\begin{aligned} \mathbb{I}\{T_{l,m+1} > \mathfrak{z}_l\} &= \mathbb{I}\{2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1}) - \mathfrak{z}_l > 0\} \\ &\le \exp\Big\{\frac{\mu}{2}\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1}) - \frac{\mu}{4}\mathfrak{z}_l\Big\}. \end{aligned}$$

This and the Cauchy-Schwarz inequality imply for $m = 1, \ldots, k-1$ the following bound:

$$\begin{aligned} &\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)|^r \mathbb{I}\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\} \qquad (4.14) \\ &= \mathbb{E}_{0,\Sigma}|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_m)|^r \mathbb{I}\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\} \\ &\le \sum_{l=1}^{m} e^{-\frac{\mu}{4}\mathfrak{z}_l} \mathbb{E}_{0,\Sigma}\Big[|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_m)|^r \exp\Big\{\frac{\mu}{2}\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1})\Big\}\Big] \\ &\le \sum_{l=1}^{m} e^{-\frac{\mu}{4}\mathfrak{z}_l} \Big\{\mathbb{E}_{0,\Sigma}\Big[|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_m)|^{2r}\Big]\Big\}^{\frac{1}{2}} \Big\{\mathbb{E}_{0,\Sigma}\Big[\exp\{\mu\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1})\}\Big]\Big\}^{\frac{1}{2}}. \end{aligned}$$

By the first statement of Lemma 4.6 with $\delta = 0$

$$\mathbb{E}_{0,\Sigma}\Big[\exp\{\mu\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1})\}\Big] \le [1 - 2\mu(1 + u_0^{-(m+1-l)})]^{-\frac{p}{2}}$$

for any $\mu < [2(1 + u_0^{-(m+1-l)}]^{-1}$. Since $u_0 > 1$ we have $[2(1 + u_0^{-(m+1-l)}]^{-1} > 1/4$ and the statement is valid for any $\mu \in (0, 1/4)$. Inequality $[1 - 2\mu(1 + u_0^{-(m+1-l)})]^{-p/2} < [1 - 4\mu]^{-p/2}$ provides for any $\mu \in (0, 1/4)$

$$\mathbb{E}_{0,\Sigma}\Big[\exp\{\mu\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1})\}\Big] < (1 - 4\mu)^{-p/2}. \qquad (4.15)$$

By the second statement of Lemma 4.7

$$\mathbb{E}_{0,\Sigma}|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_m)|^{2r} \le C(p, 2r) 2^{2r} (1 + u^{k-m})^{2r}. \qquad (4.16)$$

Putting together (4.13), (4.14), (4.15) and (4.16) we obtain

$$\begin{aligned} &\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r \\ &\le 2^r \sqrt{C(p, 2r)} (1 - 4\mu)^{-p/4} \sum_{m=1}^{k-1} \sum_{l=1}^{m} e^{-\frac{\mu}{4}\mathfrak{z}_l} (1 + u^{k-m})^r \\ &= 2^r \sqrt{C(p, 2r)} (1 - 4\mu)^{-p/4} \sum_{l=1}^{k-1} e^{-\frac{\mu}{4}\mathfrak{z}_l} \sum_{m=l}^{k-1} (1 + u^{k-m})^r \\ &\le 2^{2r} \sqrt{C(p, 2r)} (1 - 4\mu)^{-p/4} (1 - u^{-r})^{-1} \sum_{l=1}^{k-1} e^{-\frac{\mu}{4}\mathfrak{z}_l} u^{r(k-l)}, \end{aligned}$$

because $-(k-l) < -(m-l)$ and

$$\begin{aligned}\sum_{m=l}^{k-1}(1+u^{(k-m)})^r &= u^{r(k-l)}\sum_{m=l}^{k-1}(u^{-(k-l)}+u^{-(m-l)})^r \\ &< 2^r u^{r(k-l)}\sum_{m=l}^{k-1}u^{-r(m-l)} \\ &< 2^r u^{r(k-l)}(1-u^{-r})^{-1}.\end{aligned}$$

Since $u^{r(k-l)} \le u^{r(K-l)}$ for any $l < k \le K$ the choice of the threshold of the form

$$\mathfrak{z}_l = \frac{4}{\mu}\left\{r(K-l)\log u + \log(K/\alpha) - \frac{p}{4}\log(1-4\mu) - \log(1-u^{-r}) + \overline{C}(p,r)\right\}$$

with an arbitrary constant $\mu \in (0, 1/4)$, $u > 1$ from Assumption $(A4)$, $r > 0$ and $\alpha \in (0,1]$ from the PC's and with

$$\overline{C}(p,r) = \log\left\{\frac{2^{2r}[\Gamma(2r+p/2)\Gamma(p/2)]^{1/2}}{\Gamma(r+p/2)}\right\}$$

provides the required by PC bounds

$$\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_l - \widehat{\boldsymbol{\theta}}_l)^\top \mathbf{B}_l(\widetilde{\boldsymbol{\theta}}_l - \widehat{\boldsymbol{\theta}}_l)|^r \le \alpha C(p,r) \quad \text{for all } \ l = 2,\ldots,K.$$

□

4.3. Matrix results

Lemma 4.8. *The matrices $J_k \otimes \Sigma$ and $J_k \otimes \Sigma_0$ are positive semidefinite for any $k = 2,\ldots,K$.*

Moreover, under Assumption $(A3)$ with the same δ, the similar to $(A3)$ relation holds for the covariance matrices $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Sigma}_{k,0}$ of linear estimates:

$$(1-\delta)\boldsymbol{\Sigma}_k \preceq \boldsymbol{\Sigma}_{k,0} \preceq (1+\delta)\boldsymbol{\Sigma}_k\ , \quad k \le K.$$

Proof. Symmetry of J_k and Σ, (respectively, Σ_0) implies symmetry of $J_k \otimes \Sigma$, (respectively, $J_k \otimes \Sigma_0$). Notice that any vector $\gamma_{nk} \in \mathbb{R}^{nk}$ can be represented as a partitioned vector $\gamma_{nk}^\top = ((\gamma_{nk}^{(1)})^\top, (\gamma_{nk}^{(2)})^\top, \ldots, (\gamma_{nk}^{(k)})^\top)$, with $\gamma_{nk}^{(l)} \in \mathbb{R}^n$, $l = 1,\ldots,k$. Then

$$\gamma_{nk}^\top(J_k \otimes \Sigma)\gamma_{nk} = \big(\sum_{l=1}^k \gamma_{nk}^{(l)}\big)^\top \Sigma \big(\sum_{l=1}^k \gamma_{nk}^{(l)}\big) = \widetilde{\gamma}_n^\top\, \Sigma\, \widetilde{\gamma}_n, \tag{4.17}$$

where $\widetilde{\gamma}_n \stackrel{\text{def}}{=} \sum_{l=1}^k \gamma_{nk}^{(l)} \in \mathbb{R}^n$. Because $\Sigma \succ 0$ it implies $\widetilde{\gamma}_n^\top \Sigma \widetilde{\gamma}_n > 0$ for all $\widetilde{\gamma}_n \ne 0$. But even for $\gamma_{nk} \ne 0$, if its subvectors $\{\gamma_{nl}^{(l)}\}$ are linearly dependent, $\widetilde{\gamma}_n$

can be zero. Thus there exists a nonzero vector γ such that $\gamma^\top(J_k \otimes \Sigma)\gamma = 0$. This means positive semidefiniteness.

The second assertion follows from the observation that Assumption $(A3)$ due to the equality (4.17) also holds for the Kronecker product

$$(1-\delta)J_k \otimes \Sigma \preceq J_k \otimes \Sigma_0 \preceq (1+\delta)J_k \otimes \Sigma. \tag{4.18}$$

Therefore

$$(1-\delta)\mathbf{D}_k(J_k \otimes \Sigma)\mathbf{D}_k^\top \preceq \mathbf{D}_k(J_k \otimes \Sigma_0)\mathbf{D}_k^\top \preceq (1+\delta)\mathbf{D}_k(J_k \otimes \Sigma)\mathbf{D}_k^\top.$$

□

Lemma 4.9. *Fix $x \in \mathbb{R}^d$. Suppose that the weights $\{w_{l,\,i}(x)\}$ satisfy*

$$w_{l,\,i}(x)w_{m,\,i}(x) = w_{l,\,i}(x)\ , \quad l \le m. \tag{4.19}$$

Then under Assumptions $(A1)$, $(A2)$, $(A4)$ the covariance matrix $\mathbf{\Sigma}_k$ defined by (3.5) is nonsingular with

$$\det \mathbf{\Sigma}_k = \det \mathbf{B}_k^{-1} \prod_{l=2}^{k} \det(\mathbf{B}_{l-1}^{-1} - \mathbf{B}_l^{-1}) > 0\ , \quad k = 2, \dots, K. \tag{4.20}$$

Remark 4.3. The condition (4.19) holds for rectangular kernels with nested supports.

Proof. The condition (4.19) implies

$$\mathbf{W}_l \Sigma \mathbf{W}_m = \operatorname{diag}(w_{l,1}w_{m,1}/\sigma_1^2, \dots, w_{l,n}w_{m,n}/\sigma_n^2) = \mathbf{W}_l$$

for any $l \le m$. Thus the blocks of $\mathbf{\Sigma}_k$ simplify to

$$D_l \Sigma D_m^\top = \mathbf{B}_l^{-1}\mathbf{\Psi}\mathbf{W}_l\Sigma\mathbf{W}_m\mathbf{\Psi}^\top\mathbf{B}_m^{-1} = \mathbf{B}_l^{-1}\mathbf{\Psi}\mathbf{W}_l\mathbf{\Psi}^\top\mathbf{B}_m^{-1}$$

and $\mathbf{\Sigma}_k$ has a simple structure:

$$\mathbf{\Sigma}_k = \begin{pmatrix} \mathbf{B}_1^{-1} & \mathbf{B}_2^{-1} & \mathbf{B}_3^{-1} & \dots & \mathbf{B}_k^{-1} \\ \mathbf{B}_2^{-1} & \mathbf{B}_2^{-1} & \mathbf{B}_3^{-1} & \dots & \mathbf{B}_k^{-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{B}_k^{-1} & \mathbf{B}_k^{-1} & \mathbf{B}_k^{-1} & \dots & \mathbf{B}_k^{-1} \end{pmatrix}.$$

Then the determinant of $\mathbf{\Sigma}_k$ coincides with the determinant of the following irreducible block triangular matrix:

$$\det \mathbf{\Sigma}_k = \begin{vmatrix} \mathbf{B}_1^{-1} - \mathbf{B}_2^{-1} & \mathbf{B}_2^{-1} - \mathbf{B}_3^{-1} & \dots & \mathbf{B}_{k-1}^{-1} - \mathbf{B}_k^{-1} & \mathbf{B}_k^{-1} \\ \mathbf{0} & \mathbf{B}_2^{-1} - \mathbf{B}_3^{-1} & \dots & \mathbf{B}_{k-1}^{-1} - \mathbf{B}_k^{-1} & \mathbf{B}_k^{-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{B}_{k-1}^{-1} - \mathbf{B}_k^{-1} & \mathbf{B}_k^{-1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_k^{-1} \end{vmatrix}$$

implying

$$\det \boldsymbol{\Sigma}_k = \det(\mathbf{B}_1^{-1} - \mathbf{B}_2^{-1}) \det(\mathbf{B}_2^{-1} - \mathbf{B}_3^{-1}) \cdot \ldots \cdot \det(\mathbf{B}_{k-1}^{-1} - \mathbf{B}_k^{-1}) \det \mathbf{B}_k^{-1}.$$

Clearly the matrix $\boldsymbol{\Sigma}_k$ is nonsingular if all the matrices $\mathbf{B}_{l-1}^{-1} - \mathbf{B}_l^{-1}$ are nonsingular. By $(A1)$ and $(A2)$ $\mathbf{B}_l \succ 0$ for any l. By $(A4)$ there exists $u_0 > 1$ such that $\mathbf{B}_l \succeq u_0 \mathbf{B}_{l-1}$, therefore $\mathbf{B}_{l-1}^{-1} - \mathbf{B}_l^{-1} \succeq (1 - 1/u_0)\mathbf{B}_{l-1}^{-1} \succ \mathbf{B}_{l-1}^{-1} \succ 0$. □

Lemma 4.10. *In the "nonparametric situation" the moment generation function (mgf) of the joint distribution of* $\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_K$ *is*

$$\mathbb{E} \exp\bigl\{\gamma^\top (\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K - \operatorname{vec} \boldsymbol{\Theta}_K^*)\bigr\} = \exp\biggl\{\frac{1}{2}\gamma^\top \boldsymbol{\Sigma}_{K,0}\, \gamma\biggr\}. \tag{4.21}$$

Thus, provided that $\boldsymbol{\Sigma}_{K,0} \succ 0$, *it holds that* $\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K \sim \mathcal{N}(\operatorname{vec} \boldsymbol{\Theta}_K^*, \boldsymbol{\Sigma}_{K,0})$.

Similarly, in the "parametric situation", if $\boldsymbol{\Sigma}_K \succ 0$, *then the joint distribution of* $\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K$ *is* $\mathcal{N}(\operatorname{vec} \boldsymbol{\Theta}_K, \boldsymbol{\Sigma}_K)$ *with the mgf:*

$$\mathbb{E} \exp\bigl\{\gamma^\top (\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K - \operatorname{vec} \boldsymbol{\Theta}_K)\bigr\} = \exp\biggl\{\frac{1}{2}\gamma^\top \boldsymbol{\Sigma}_K\, \gamma\biggr\}. \tag{4.22}$$

Proof. Let $\gamma \in \mathbb{R}^{pK}$ be written in a partitioned form $\gamma^\top = (\gamma_1^\top, \ldots, \gamma_K^\top)$ with $\gamma_l \in \mathbb{R}^p$, $l = 1, \ldots, K$. Then the mgf for the centered random vector $\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K - \operatorname{vec} \boldsymbol{\Theta}_K^* \in \mathbb{R}^{pK}$, due to the decomposition (2.9) $\widetilde{\boldsymbol{\theta}}_l = \boldsymbol{\theta}_l^* + D_l \Sigma_0^{1/2} \boldsymbol{\varepsilon}$ with $D_l = \mathbf{B}_l^{-1} \boldsymbol{\Psi} \mathbf{W}_l$, can be represented as follows:

$$\begin{aligned} \mathbb{E} \exp\bigl\{\gamma^\top (\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K - \operatorname{vec} \boldsymbol{\Theta}_K^*)\bigr\} &= \mathbb{E} \exp\Bigl\{\sum_{l=1}^K \gamma_l^\top (\widetilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l^*)\Bigr\} \\ = \quad \mathbb{E} \exp\Bigl\{\sum_{l=1}^K \gamma_l^\top D_l \Sigma_0^{1/2} \boldsymbol{\varepsilon}\Bigr\} &= \mathbb{E} \exp\Bigl\{\Bigl(\sum_{l=1}^K D_l^\top \gamma_l\Bigr)^\top \Sigma_0^{1/2} \boldsymbol{\varepsilon}\Bigr\}. \end{aligned}$$

A trivial observation that $\sum_{l=1}^K D_l^\top \gamma_l$ is a vector in $\mathbb{R}^n$ and $\Sigma_0^{1/2} \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_0)$ by (1.1) implies by the definition of $\boldsymbol{\Sigma}_{K,0}$ the first assertion of the lemma, because

$$\begin{aligned} \mathbb{E} \exp\Bigl\{\Bigl(\sum_{l=1}^K D_l^\top \gamma_l\Bigr)^\top \Sigma_0^{1/2} \boldsymbol{\varepsilon}\Bigr\} &= \exp\biggl\{\frac{1}{2}\Bigl(\sum_{l=1}^K D_l^\top \gamma_l\Bigr)^\top \Sigma_0 \Bigl(\sum_{l=1}^K D_l^\top \gamma_l\Bigr)\biggr\} \\ = \quad \exp\biggl\{\frac{1}{2}\bigl(\mathbf{D}_K^\top \gamma\bigr)^\top (J_K \otimes \Sigma_0) \mathbf{D}_K^\top \gamma\biggr\} &= \exp\biggl\{\frac{1}{2}\gamma^\top \boldsymbol{\Sigma}_{K,0}\, \gamma\biggr\}, \end{aligned}$$

here $\mathbf{D}_K$ is defined by (3.24). □

4.4. Proof of the propagation property

Lemma 4.11. *The Kullback-Leibler divergence between the distributions of* $\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_k$ *under the true measure and under the "parametric" has the following*

form:

$$2\mathbb{KL}(\mathbb{P}^k_{\boldsymbol{f},\Sigma_0},\mathbb{P}^k_{\boldsymbol{\theta},\Sigma}) \stackrel{\text{def}}{=} 2\mathbb{E}_{\boldsymbol{f},\Sigma_0}\log\left(\frac{\mathrm{d}\mathbb{P}^k_{\boldsymbol{f},\Sigma_0}}{\mathrm{d}\mathbb{P}^k_{\boldsymbol{\theta},\Sigma}}\right)$$

$$= \Delta(k) + \log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) + \mathrm{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{k,0}) - pk, \tag{4.23}$$

where

$$b(k) \stackrel{\text{def}}{=} \mathrm{vec}\,\boldsymbol{\Theta}^*_k - \mathrm{vec}\,\boldsymbol{\Theta}_k \tag{4.24}$$

$$\Delta(k) \stackrel{\text{def}}{=} b(k)^\top\boldsymbol{\Sigma}_k^{-1}b(k). \tag{4.25}$$

Proof. Denote the Radon-Nikodym derivative by $Z_k \stackrel{\text{def}}{=} \mathrm{d}\mathbb{P}^k_{\boldsymbol{f},\Sigma_0}/\mathrm{d}\mathbb{P}^k_{\boldsymbol{\theta},\Sigma}$. Then

$$\log\big(Z_k(y)\big) = \frac{1}{2}\log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) - \frac{1}{2}\|\boldsymbol{\Sigma}_{k,0}^{-1/2}(y-\mathrm{vec}\,\boldsymbol{\Theta}^*_k)\|^2 + \frac{1}{2}\|\boldsymbol{\Sigma}_k^{-1/2}(y-\mathrm{vec}\,\boldsymbol{\Theta}_k)\|^2 \tag{4.26}$$

can be considered as a quadratic function of $\mathrm{vec}\,\boldsymbol{\Theta}_k$. By the Taylor expansion at the point $\mathrm{vec}\,\boldsymbol{\Theta}^*_k$ the last expression reads as follows

$$\log\big(Z_k(y)\big) = \frac{1}{2}\log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) - \frac{1}{2}\|\boldsymbol{\Sigma}_{k,0}^{-1/2}(y-\mathrm{vec}\,\boldsymbol{\Theta}^*_k)\|^2$$
$$+ \frac{1}{2}\|\boldsymbol{\Sigma}_k^{-1/2}(y-\mathrm{vec}\,\boldsymbol{\Theta}^*_k)\|^2 + b(k)^\top\boldsymbol{\Sigma}_k^{-1}(y-\mathrm{vec}\,\boldsymbol{\Theta}^*_k) + \frac{1}{2}\Delta(k).$$

Then the expression for the Kullback-Leibler divergence can be written in the following way:

$$\mathbb{KL}(\mathbb{P}^k_{\boldsymbol{f},\Sigma_0},\mathbb{P}^k_{\boldsymbol{\theta},\Sigma}) \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{f},\Sigma_0}\log\big(Z_k\big)$$
$$= \frac{1}{2}\log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) + \frac{1}{2}\Delta(k)$$
$$+ \frac{1}{2}\mathbb{E}\big\{\|\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{\Sigma}_{k,0}^{1/2}\xi\|^2 - \|\xi\|^2 + 2b(k)^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{k,0}^{1/2}\xi\big\},$$

where $\xi\sim\mathcal{N}(0,I_{pk})$. This implies

$$2\mathbb{KL}(\mathbb{P}^k_{\boldsymbol{f},\Sigma_0},\mathbb{P}^k_{\boldsymbol{\theta},\Sigma}) = \Delta(k) + \log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) + \mathrm{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{k,0}) - pk. \tag{4.27}$$

In the case of *homogeneous errors* with $\sigma_{0,i}=\sigma_0$ and $\sigma_i=\sigma, i=1,\dots,n$ the calculations simplify a lot. Now

$$\boldsymbol{\Sigma}_k = \sigma^2\mathbf{V}_k, \quad \boldsymbol{\Sigma}_{k,0} = \sigma_0^2\mathbf{V}_k$$

with a $pk\times pk$ matrix $\mathbf{V}_k$ defined as

$$\mathbf{V}_k = \big(\overline{D}_1\oplus\cdots\oplus\overline{D}_k\big)\big(J_k\otimes I_n\big)\big(\overline{D}_1\oplus\cdots\oplus\overline{D}_k\big)^\top,$$

where $\overline{D}_l = (\Psi \mathcal{W}_l \Psi^\top)^{-1} \Psi \mathcal{W}_l$, $l = 1, \ldots, k$ does not depend on σ. Then $\Delta(k) = \sigma^{-2}\Delta_1(k)$, with $\Delta_1(k) \stackrel{\mathrm{def}}{=} b(k)^\top \mathbf{V}_k^{-1} b(k)$, $\det \boldsymbol{\Sigma}_k / \det \boldsymbol{\Sigma}_{k,0} = (\sigma^2/\sigma_0^2)^{pk}$, and the expression for the Kullback-Leibler divergence reads as follows:

$$\begin{aligned} \mathbb{KL}(\mathbb{P}^k_{\boldsymbol{f},\Sigma_0}, \mathbb{P}^k_{\boldsymbol{\theta},\Sigma}) &= pk \log\Big(\frac{\sigma}{\sigma_0}\Big) + \frac{1}{2}\Delta(k) + \frac{pk}{2}\Big(\frac{\sigma_0^2}{\sigma^2} - 1\Big) \\ &= pk \log\Big(\frac{\sigma}{\sigma_0}\Big) + \frac{1}{2\sigma^2} b(k)^\top \mathbf{V}_k^{-1} b(k) + \frac{pk}{2}\Big(\frac{\sigma_0^2}{\sigma^2} - 1\Big), \end{aligned} \tag{4.28}$$

implying the same asymptotic behavior as in (3.13). □

Proof of Theorem 3.2 (Propagation property). Notice that for any nonnegative measurable function $g = g(\widetilde{\boldsymbol{\Theta}}_k)$ the Cauchy-Schwarz inequality implies

$$\mathbb{E}_{\boldsymbol{f},\Sigma_0}[g] = \mathbb{E}_{\boldsymbol{\theta},\Sigma}[g Z_k] \le \big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[g^2]\big)^{1/2} \big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]\big)^{1/2} \tag{4.29}$$

with the Radon-Nikodym derivative $Z_k = \mathrm{d}\mathbb{P}^k_{\boldsymbol{f},\Sigma_0} / \mathrm{d}\mathbb{P}^k_{\boldsymbol{\theta},\Sigma}$. One gets the first assertion taking $g = |(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})|^{r/2}$, and applying "the parametric risk bound" with $\delta = 0$ from (4.7):

$$\begin{aligned} \mathbb{E}[g] &\le \big(\mathbb{E}_{\boldsymbol{\theta},\Sigma} |(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})|^r\big)^{1/2} \big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]\big)^{1/2} \\ &= \big(\mathbb{E}_{\boldsymbol{\theta},\Sigma} |2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta})|^r\big)^{1/2} \big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]\big)^{1/2} \\ &\le \big(\mathbb{E}|\chi_p^2|^r\big)^{1/2} \big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]\big)^{1/2}. \end{aligned}$$

The second assertion of the theorem is treated similarly by application of the pivotality property from Lemma 4.1 and the propagation conditions (2.17).

To calculate $\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]$ let us consider $\log Z_k$ given by

$$\begin{aligned} \log\big(Z_k(y)\big) = \frac{1}{2}\log\Big(\frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}}\Big) &- \frac{1}{2}\|\boldsymbol{\Sigma}_{k,0}^{-1/2}(y - \operatorname{vec} \boldsymbol{\Theta}_k^*)\|^2 \\ &+ \frac{1}{2}\|\boldsymbol{\Sigma}_k^{-1/2}(y - \operatorname{vec} \boldsymbol{\Theta}_k)\|^2 \end{aligned}$$

as a function of $\operatorname{vec} \boldsymbol{\Theta}_k^*$. Application of the Taylor expansion at the point $\operatorname{vec} \boldsymbol{\Theta}_k$ yields

$$\begin{aligned} 2\log Z_k &= \log\frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}} - \|\boldsymbol{\Sigma}_{k,0}^{-1/2}(y - \operatorname{vec} \boldsymbol{\Theta}_k)\|^2 + \|\boldsymbol{\Sigma}_k^{-1/2}(y - \operatorname{vec} \boldsymbol{\Theta}_k)\|^2 \\ &+ 2b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1}(y - \operatorname{vec} \boldsymbol{\Theta}_k) - b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1} b(k). \end{aligned}$$

With $\xi \sim \mathcal{N}(0, I_{pk})$ the second moment of the Radon-Nikodym derivative reads as follows

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2] \\
&= \frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}} \exp\{-b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1} b(k)\} \\
&\times \mathbb{E} \exp\{-\|\boldsymbol{\Sigma}_{k,0}^{-1/2}\boldsymbol{\Sigma}_k^{1/2}\xi\|^2 + \|\xi\|^2 + 2b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2}\xi\} \\
&= \frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}} \big[\det\big(2\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2} - I_{pk}\big)\big]^{-1/2} \\
&\times \exp\{2b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2}\big(2\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2} - I_{pk}\big)^{-1}\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1} b(k) - b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1} b(k)\} \\
&= \frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}} \big[\prod_{j=1}^{pk}\{2\lambda_j(\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2}) - 1\}\big]^{-1/2} \\
&\times \exp\{b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1/2}\big[2\boldsymbol{\Sigma}_{k,0}^{-1/2}\boldsymbol{\Sigma}_k^{1/2}\big(2\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2} - I_{pk}\big)^{-1}\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1/2} - I_{pk}\big] \\
&\times \boldsymbol{\Sigma}_{k,0}^{-1/2} b(k)\}. \qquad (4.30)
\end{aligned}
$$

To estimate the obtained expression in terms of the level of noise misspecification δ notice that the condition (3.8) implies

$$
\left(\frac{1}{1+\delta}\right)^{pk} \le \frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}} \le \left(\frac{1}{1-\delta}\right)^{pk},
$$

$$
\left(\frac{1-\delta}{1+\delta}\right)^{\frac{pk}{2}} \le \big[\prod_{j=1}^{pk}\{2\lambda_j(\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2}) - 1\}\big]^{-1/2} \le \left(\frac{1+\delta}{1-\delta}\right)^{\frac{pk}{2}}.
$$

$$
\frac{1-\delta}{1+\delta} I_{pk} \preceq \left(2\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2} - I_{pk}\right)^{-1} \preceq \frac{1+\delta}{1-\delta} I_{pk}.
$$

Therefore the quantity in the exponent in (4.30) is bounded by:

$$
\begin{aligned}
&\left(2\frac{1-\delta}{(1+\delta)^2} - 1\right) b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1} b(k) \\
&\le b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1/2}\big[2\boldsymbol{\Sigma}_{k,0}^{-1/2}\boldsymbol{\Sigma}_k^{1/2}\big(2\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2} - I_{pk}\big)^{-1}\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1/2} - I_{pk}\big]\boldsymbol{\Sigma}_{k,0}^{-1/2} b(k) \\
&\le \left(2\frac{1+\delta}{(1-\delta)^2} - 1\right) b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1} b(k).
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
\frac{\Delta(k)}{1+\delta} &= \frac{1}{1+\delta} b(k)^\top \boldsymbol{\Sigma}_k^{-1} b(k) \\
&\le b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1} b(k) \\
&\le \frac{1}{1-\delta} b(k)^\top \boldsymbol{\Sigma}_k^{-1} b(k) = \frac{\Delta(k)}{1-\delta}.
\end{aligned}
$$

Finally,

$$\left(\frac{1-\delta}{(1+\delta)^3}\right)^{\frac{pk}{2}} \exp\left\{\left(\frac{2(1-\delta)}{(1+\delta)^2}-1\right)\frac{\Delta(k)}{1+\delta}\right\}$$
$$\leq \mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2] \leq \left(\frac{1+\delta}{(1-\delta)^3}\right)^{\frac{pk}{2}} \exp\left\{\left(\frac{2(1+\delta)}{(1-\delta)^2}-1\right)\frac{\Delta(k)}{1-\delta}\right\}. \quad (4.31)$$

In the *case of homogeneous errors* the expression for $\log Z_k$ reads as

$$\begin{aligned}\log Z_k &= pk\log\left(\frac{\sigma}{\sigma_0}\right) + \frac{1}{2}\left(\frac{1}{\sigma^2}-\frac{1}{\sigma_0^2}\right)\|\mathbf{V}_k^{-1/2}(y-\operatorname{vec}\boldsymbol{\Theta}_k)\|^2 \\ &+ \frac{1}{\sigma_0^2}b(k)^\top\mathbf{V}_k^{-1}(y-\operatorname{vec}\boldsymbol{\Theta}_k) - \frac{1}{2\sigma_0^2}b(k)^\top\mathbf{V}_k^{-1}b(k),\end{aligned}$$

implying

$$\mathbb{E}_{\boldsymbol{\theta},\sigma}[Z_k^2] = \left(\frac{\sigma^2}{\sigma_0^2}\right)^{pk}\left(\frac{\sigma_0^2}{2\sigma^2-\sigma_0^2}\right)^{\frac{pk}{2}}\exp\left\{\frac{b(k)^\top\mathbf{V}_k^{-1}b(k)}{2\sigma^2-\sigma_0^2}\right\}.$$

By Assumption $(A3)$

$$\begin{aligned}&\left(\frac{1-\delta}{(1+\delta)^3}\right)^{\frac{pk}{2}}\exp\left\{\frac{\Delta_1(k)}{\sigma^2(1+\delta)}\right\} \\ \leq \mathbb{E}_{\boldsymbol{\theta},\sigma}[Z_k^2] &\leq \left(\frac{1+\delta}{(1-\delta)^3}\right)^{\frac{pk}{2}}\exp\left\{\frac{\Delta_1(k)}{\sigma^2(1-\delta)}\right\}, \quad (4.32)\end{aligned}$$

where p is the dimension of the parameter set and k is the degree of the localization. □

4.5. Bounds for the bias and variance

Before proceeding with the proof we need to show that the weights $W_{l,i}^*(x)$ defined by (3.36) preserve the reproducing polynomials property:

Lemma 4.12. *Let $x \in \mathbb{R}$ be such that Assumptions $(A1)-(A2)$ hold. Then the weights defined by (3.36) satisfy*

$$\sum_{i=1}^n W_{l,i}^*(x) = 1, \quad (4.33)$$
$$\sum_{i=1}^n (X_i - x)^m W_{l,i}^*(x) = 0\,, \quad m = 1,\dots,p-1.$$

for all $l = 1,\dots,K$ and any design points $\{X_1,\dots,X_n\}$.

Proof. The assertion can be easily obtained similarly to the proof of Proposition 1.12 from [41]. □

Proof of Lemma 3.14. By Lemma 4.12 and the Taylor theorem with τ_i such that the points $\tau_i X_i$ are between X_i and x, and utilizing Assumption $(A10)$ we have with $b_{l,f}(x) = \boldsymbol{e}_1^\top \boldsymbol{\theta}_l^*(x) - f(x)$:

$$\begin{aligned} |b_{l,f}(x)| &\le \frac{1}{(p-1)!}\sum_{i=1}^n |f^{(p-1)}(\tau_i X_i) - f^{(p-1)}(x)||X_i - x|^{p-1}|W^*_{l,\,i}(x)| \\ &\le \frac{L}{(p-1)!}\sum_{i=1}^n |\tau_i X_i - x|^{\beta-(p-1)}|X_i - x|^{p-1}|W^*_{l,\,i}(x)| \\ &\le \frac{L h_l^\beta}{(p-1)!}\sum_{i=1}^n |W^*_{l,\,i}(x)|. \end{aligned}$$

Under the assumptions of the theorem the sum of the polynomial weights can be bounded as follows:

$$\begin{aligned} \sum_{i=1}^n |W^*_{l,\,i}(x)| &\le w_{max}\sum_{i=1}^n \sigma_i^{-2}\|\mathbf{B}_l^{-1}\Psi_i\| \\ &\le \varkappa(\Sigma)\frac{w_{max}}{\lambda_0 n h_l}\sum_{i=1}^n \|\Psi_i\|\,\mathbb{I}\{X_i \in [x-h_l, x+h_l]\} \\ &\le \varkappa(\Sigma)\frac{w_{max}\sqrt{e}}{\lambda_0} a_0 \max\{2, \frac{1}{n h_l}\} \\ &\le \varkappa(\Sigma)\frac{2a_0 w_{max}\sqrt{e}}{\lambda_0}, \end{aligned}$$

and the first assertion is justified in view of:

$$\overline{b}_{k,f}(x) \stackrel{\text{def}}{=} \sup_{1\le l\le k} |b_{l,f}(x)| \le \varkappa(\Sigma)\frac{2a_0 w_{max}\sqrt{e}a_0}{\lambda_0}\frac{L h_k^\beta}{(p-1)!}. \tag{4.34}$$

To bound the variance, just notice that by (3.20) for any $\gamma \in \mathbb{R}^p$

$$\gamma^\top \mathbf{B}_k^{-1}\gamma \le \frac{\sigma^2_{max}}{n h_k \Lambda_0}\|\gamma\|^2.$$

Then under Condition $(A3)$ by (3.2) for the variance term we have:

$$\begin{aligned} \sigma_k^2(x) &= \boldsymbol{e}_1^\top \operatorname{Var}\widetilde{\boldsymbol{\theta}}_k\, \boldsymbol{e}_1 \\ &\le (1+\delta)\boldsymbol{e}_1^\top \mathbf{B}_k^{-1}\boldsymbol{e}_1 \\ &\le (1+\delta)\frac{\sigma^2_{max}}{n h_k \Lambda_0}. \end{aligned}$$

□

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory (Tsahkadsor, 1971), Akadémiai Kiadó, Budapest* 267–281. MR0483125

[2] Arlot, S. (2009). Model selection by resampling penalization. *Electron. J. Stat.* **3** 557–624. MR2519533
[3] Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least squares regression. *J. Mach. Learn. Res.* **10**(Feb) 245–279.
[4] Baraud, Y., Giraud, C. and Huet, S. (2009). Gaussian model selection with an unknown variance. *Ann. Statist.* **37:2** 630–672. MR2502646
[5] Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society* **3:3** 203–268. MR1848946
[6] Brillinger, D. R. (1977). Discussion of Stone (1977). *Ann. Statist.* **5:4** 622–623. MR0448466
[7] Brua, J.-Y. (2009). Asymptotic efficient estimators for non-parametric heteroscedastic model. *Statistical Methodology* **6:1** 47–60. MR2655538
[8] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74:368** 829–836. MR0556476
[9] Dalalyan A. S. and Salmon J. (2011). Sharp Oracle Inequalities for Aggregation of Affine Estimators. Preprint arXiv:1104.3969v3.
[10] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrica* **81:3** 425–455 MR1311089
[11] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Monographs on Statistics and Applied Probability, 66. Chapman and Hall, London. MR1383587
[12] Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29:1** 153–193. MR1833962
[13] Galtchouk, L. and Pergamenshchikov, S. (2009). Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression *Journal of the Korean Statistical Society* **38:4** 305–322. MR2582478
[14] Galtchouk, L. and Pergamenshchikov, S. (2009). Sharp non-asymptotic oracle inequalities for nonparametric heteroscedastic regression models. *J. Nonparametr. Stat.* **21:1** 1–16. MR2483856
[15] Goldenshluger, A. and Nemirovski, A. (1994). On spatial adaptive estimation of nonparametric regression. *Research report, Technion-Israel Inst. Technology, Haifa, Israel.*
[16] Golubev, Y. and Spokoiny, V. (2009). Exponential bounds for minimum contrast estimators. *Electron. J. Stat.* **3** 712–746. MR2534199
[17] Efromovich, S. and Pinsker, M. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statistica Sinica* **6:4** 925–942. MR1422411
[18] Efromovich, S. (2007). Sequential design and estimation in heteroscedastic nonparametric regression. *Sequential Analysis* **26:1** 3–25. MR2293406
[19] Katkovnik, V. Ja. (1979). Linear and nonlinear methods of nonparametric regression analysis. (Russian) *Soviet Automat. Control* **5** 35–46, 93. MR0582402
[20] Katkovnik, V. Ja. (1983). Convergence of linear and nonlinear nonparametric estimates of "kernel" type. *Automat. Remote Control* **44:4** 495–506; translated from *Avtomat. i Telemekh.* 1983 **4** 108–120 (Russian). MR0728628

[21] Katkovnik, V. Ja. (1985). *Nonparametric Identification and Data Smoothing: The Method of Local Approximation*. Nauka, Moscow (Russian). MR0874985

[22] Katkovnik, V., Egiazarian, K. and Astola, J. (2006). *Local Approximation Techniques in Signal and Image Processing*. Bellingham, WA: SPIE Press.

[23] Katkovnik, V. and Spokoiny, V. (2008). Spatially adaptive estimation via fitted local likelihood techniques. *IEEE Trans. Signal Process.*, **56:3** 873–886. MR2518663

[24] Kerkyacharian, G., Lepski, O. and Picard, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields* **121:2** 137–170. MR1863916

[25] Kerkyacharian, G., Lepski, O. and Picard, D. (2007). Nonlinear estimation in anisotropic multiindex denoising. Sparse case. *Teor. Veroyatn. Primen.* **52:1** 150–171; translation in *Theory Probab. Appl.* (2008). **52:1** 58–77. MR2354574

[26] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics* **22** 79–86. MR0039968

[27] Lepskii, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. (Russian) *Teor. Veroyatnost. i Primenen.* **35:3** 459–470; translation in *Theory Probab. Appl.* **35:3** 454–466. MR1091202

[28] Lepskii, O. V. (1992). Asymptotic minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. (Russian) *Teor. Veroyatnost. i Primenen.* **37:3** 468–481; translation in *Theory Probab. Appl.* **37:3** 433–448. MR1214353

[29] Lepski, O. V., Mammen, E. and Spokoiny, V.G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Stat.* **25:3** 929–947. MR1447734

[30] Lepski, O. V. and Spokoiny, V.G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Stat.* **25:6** 2512–2546. MR1604408

[31] Loader, C. (1999). *Local Regression and Likelihood.* Statistics and Computing. Springer-Verlag, New York. MR1704236

[32] Massart, P. (2003). *Concentration Inequalities and Model Selection* (2007). Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. Lecture Notes in Mathematics, 1896. Springer Berlin. MR2319879

[33] Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Stat.* **22:**3 1346–1370. MR1311979

[34] Saumard, A. (2010). Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression, Preprint hal-00512304, v1.

[35] Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *J. Multivariate Anal.* **82:1** 111–133. MR1918617

[36] Spokoiny, V. and Vial, C. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.* **37:5B** 2783–2807. MR2541447

[37] Stone, C. J. (1977). Consistent nonparametric regression. With discussion and a reply by the author. *Ann. Statist.* **5:4** 595–645. MR0443204

[38] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8:6** 1348–1360. MR0594650

[39] Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82:398** 559–567. MR0898359

[40] Tsybakov, A. B. (1986). Robust reconstruction of functions by a local approximation method. (Russian) *Problemy Peredachi Informatsii* **22:2** 69–84 (*Problems of Information Transmission*, 1986 **22:2** 133–146). MR0855002

[41] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer-Verlag, New York. MR2724359

[42] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50:1** 1–25. MR0640163