SFB 649 Discussion Paper 2015-031

Simultaneous likelihood-based bootstrap confidence sets for a large number of models

Mayya Zhilova*

* Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Simultaneous likelihood-based bootstrap confidence sets for a large number of models

Mayya Zhilova*,†

Weierstrass-Institute,
Mohrenstr. 39,
10117 Berlin, Germany,
zhilova@wias-berlin.de

June 18, 2015

Abstract

The paper studies a problem of constructing simultaneous likelihood-based confidence sets. We consider a simultaneous multiplier bootstrap procedure for estimating the quantiles of the joint distribution of the likelihood ratio statistics, and for adjusting the confidence level for multiplicity. Theoretical results state the bootstrap validity in the following setting: the sample size n is fixed, the maximal parameter dimension $p_{\max}$ and the number of considered parametric models K are s.t. $(\log K)^{12} p_{\max}^3/n$ is small. We also consider the situation when the parametric models are misspecified. If the models' misspecification is significant, then the bootstrap critical values exceed the true ones and the simultaneous bootstrap confidence set becomes conservative. Numerical experiments for local constant and local quadratic regressions illustrate the theoretical results.

JEL classification codes: C13, C15

Keywords: simultaneous inference, correction for multiplicity, family-wise error, misspecified model, multiplier/weighted bootstrap

*I am very grateful to Vladimir Spokoiny for many helpful discussions and comments.

†Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 649 "Economic Risk" is gratefully acknowledged.

Contents

1 Introduction **3**

2 The multiplier bootstrap procedure **9**

3 Theoretical justification of the bootstrap procedure **11**
3.1 Overview of the theoretical approach . . . 11
3.2 Main results . . . 14

4 Numerical experiments **16**
4.1 Local constant regression . . . 16
4.2 Local quadratic regression . . . 16
4.3 Simulated data . . . 17
4.4 Effect of the modeling bias on a width of a bootstrap confidence band . . . 17
4.5 Effective coverage probability (local constant estimate) . . . 18
4.6 Correction for multiplicity . . . 21

5 Conditions **22**
5.1 Basic conditions . . . 22
5.2 Conditions required for the bootstrap validity . . . 23
5.3 Dependence of the involved terms on the sample size and cardinality of the parameters' set . . . 24

A Approximation of the joint distributions of ℓ_2-norms **25**
A.1 Joint Gaussian approximation of ℓ_2-norm of sums of independent vectors by Lindeberg's method . . . 27
A.2 Gaussian comparison . . . 35
A.3 Simultaneous anti-concentration for ℓ_2-norms of Gaussian vectors . . . 37
A.4 Proof of Proposition A.1 . . . 39

B Square-root Wilks approximations **42**
B.1 Finite sample theory . . . 42
B.2 Finite sample theory for the bootstrap world . . . 44
B.3 Simultaneous square-root Wilks approximations . . . 46

C Proofs of the main results **47**
C.1 Bernstein matrix inequality . . . 47
C.2 Bootstrap validity for the case of one parametric model . . . 48
C.3 Proof of Theorem 3.1 . . . 49
C.4 Proof of Theorem 3.2 . . . 51
C.5 Proof of Theorem 3.3 . . . 55

References **56**

1 Introduction

The problem of simultaneous confidence estimation appears in numerous practical applications when a confidence statement has to be made simultaneously for a collection of objects, e.g. in safety analysis in clinical trials, gene expression analysis, population biology, functional magnetic resonance imaging and many others. See e.g. Miller (1981); Westfall (1993); Manly (2006); Benjamini (2010); Dickhaus (2014), and references therein. This problem is also closely related to construction of simultaneous confidence bands in curve estimation, which goes back to Working and Hotelling (1929). For an extensive literature review about constructing the simultaneous confidence bands we refer to Hall and Horowitz (2013), Liu (2010), and Wasserman (2006).

A simultaneous confidence set requires a probability bound to be constructed jointly for several possibly dependent statistics. Therefore, the critical values of the corresponding statistics should be chosen in such a way that the joint probability distribution achieves a required family-wise confidence level. This choice can be made by multiplicity correction of the marginal confidence levels. The Bonferroni correction method (Bonferroni (1936)) uses a probability union bound, the corrected marginal significance levels are taken equal to the total level divided by the number of models. This procedure can be very conservative if the considered statistics are positively correlated and if their number is large. The Šidák correction method (Šidák (1967)) is more powerful than Bonferroni correction, however, it also becomes conservative in the case of large number of dependent statistics.

Most of the existing results about simultaneous bootstrap confidence sets and resampling-based multiple testing are asymptotic (with sample size tending to infinity), see e.g. Beran (1988, 1990); Hall and Pittelkow (1990); Härdle and Marron (1991); Shao and Tu (1995); Hall and Horowitz (2013), and Westfall (1993); Dickhaus (2014). The results based on asymptotic distribution of maximum of an approximating Gaussian process (see Bickel and Rosenblatt (1973); Johnston (1982); Härdle (1989)) require a huge sample size n, since they yield a coverage probability error of order $(\log(n))^{-1}$ (see Hall (1991)). Some papers considered an alternative approach in context of confidence band estimation based on the approximation of the underlying empirical processes by its bootstrap counterpart. In particular, Hall (1993) showed that such an approach leads to a significant improvement of the error rate (see also Neumann and Polzehl (1998); Claeskens and Van Keilegom (2003)). Chernozhukov et al. (2014a) constructed honest confidence bands for nonparametric density estimators without requiring the existence of limit distribution of the supremum of the studentized empirical process: instead, they used an approximation between sup-norms of an empirical and Gaussian processes, and anti-concentration

property of suprema of Gaussian processes.

In many modern applications the sample size cannot be large, and/or can be smaller than a parameter dimension, for example, in genomics, brain imaging, spatial epidemiology and microarray data analysis, see Leek and Storey (2008); Kim and van de Wiel (2008); Arlot et al. (2010); Cao and Kosorok (2011), and references therein.

For the recent results on resampling-based simultaneous confidence sets in high-dimensional finite sample set-up we refer to the papers by Arlot et al. (2010) and Chernozhukov et al. (2013a, 2014a,b). Arlot et al. (2010) considered i.i.d. observations of a Gaussian vector with a dimension possibly much larger than the sample size, and with unknown covariance matrix. They examined multiple testing problems for the mean values of its coordinates and provided non-asymptotic control for the family-wise error rate using resampling-type procedures. Chernozhukov et al. (2013a) presented a number of non-asymptotic results on Gaussian approximation and multiplier bootstrap for maxima of sums of high-dimensional vectors (with a dimension possibly much larger than a sample size) in a very general set-up. As an application the authors considered the problem of multiple hypothesis testing in the framework of approximate means. They derived non-asymptotic results for the general stepdown procedure by Romano and Wolf (2005) with improved error rates and in high-dimensional setting. Chernozhukov et al. (2014a) showed how this technique applies to the problem of constructing an honest confidence set in nonparametric density estimation. Chernozhukov et al. (2014b) extended the results from maxima to the class of sparsely convex sets.

The present paper studies simultaneous likelihood-based bootstrap confidence sets in the following setting:

1. the sample size n is fixed;

2. the parametric models can be misspecified;

3. the number K of the parametric models can be exponentially large w.r.t. n;

4. the maximal dimension $p_{\max}$ of the considered parametric models can be dependent on the sample size n.

This set-up, in contrast with the paper by Chernozhukov et al. (2014b), does not require the sparsity condition , in particular the dimension $p_1, \ldots, p_K$ of each parametric family may grow with the sample size. Moreover, the simultaneous likelihood-based confidence sets are not necessarily convex, and the parametric assumption can be violated.

The considered simultaneous multiplier bootstrap procedure involves two main steps: estimation of the quantile functions of the likelihood ratio statistics, and multiplicity correction of the marginal confidence level. Theoretical results of the paper state the

bootstrap validity in the setting 1-4 taking in account the multiplicity correction. The resulting approximation bound requires the quantity $(\log K)^{12} p_{\max}^3 / n$ to be small. The log-factor here is suboptimal and can probably be improved. The paper particularly focuses on the impact of the model misspecification. We distinguish between slight and strong misspecifications. Under the so called small modeling bias condition **(SmB)** given in Section 5.2 the bootstrap approximation is accurate. This condition roughly means that all the parametric models are close to the true distribution. If the **(SmB)** condition is not fulfilled, then the simultaneous bootstrap confidence set is still applicable, however, it becomes conservative. This property is nicely confirmed by the numerical experiments in Section 4.

Let the random data

$$\boldsymbol{Y} \stackrel{\text{def}}{=} (Y_1, \ldots, Y_n)^\top \tag{1.1}$$

consist of *independent* observations Y_i, and belong to the probability space $(\Omega, \mathcal{F}, I\!P)$. The sample size n is *fixed*. $I\!P$ is an *unknown probability distribution* of the sample $\boldsymbol{Y}$. Consider K regular parametric families of probability distributions:

$$\{I\!P_k(\boldsymbol{\theta})\} \stackrel{\text{def}}{=} \{I\!P_k(\boldsymbol{\theta}) \ll \mu_0, \boldsymbol{\theta} \in \Theta_k \subset I\!R^{p_k}\}, \quad k = 1, \ldots, K.$$

Each parametric family induces the quasi log-likelihood function for $\boldsymbol{\theta} \in \Theta_k \subset I\!R^{p_k}$

$$\begin{aligned} L_k(\boldsymbol{Y}, \boldsymbol{\theta}) &\stackrel{\text{def}}{=} \log\left(\frac{dI\!P_k(\boldsymbol{\theta})}{d\mu_0}(\boldsymbol{Y})\right) \\ &= \sum\nolimits_{i=1}^n \log\left(\frac{dI\!P_k(\boldsymbol{\theta})}{d\mu_0}(Y_i)\right). \end{aligned} \tag{1.2}$$

It is important that we *do not require* that $I\!P$ belongs to any of the known parametric families $\{I\!P_k(\boldsymbol{\theta})\}$, that is why the term *quasi* log-likelihood is used here. Below in this section we consider two popular examples of simultaneous confidence sets in terms of the quasi log-likelihood functions (1.2). Namely, the simultaneous confidence band for local constant regression, and multiple quantiles regression.

The target of estimation for the misspecified log-likelihood $L_k(\boldsymbol{\theta})$ is such a parameter $\boldsymbol{\theta}_k^*$, that minimises the Kullback-Leibler distance between the unknown true measure $I\!P$ and the parametric family $\{I\!P_k(\boldsymbol{\theta})\}$:

$$\boldsymbol{\theta}_k^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta_k} I\!E L_k(\boldsymbol{\theta}). \tag{1.3}$$

The maximum likelihood estimator is defined as:

$$\widetilde{\boldsymbol{\theta}}_k \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta_k} L_k(\boldsymbol{\theta}).$$

The parametric sets Θ_k have dimensions p_k, therefore, $\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^* \in I\!R^{p_k}$. For $1 \le k, j \le K$ and $k \ne j$ the numbers p_k and p_j can be unequal.

The likelihood-based confidence set for the target parameter $\boldsymbol{\theta}_k^*$ is

$$\mathcal{E}_k(\mathfrak{z}) \stackrel{\text{def}}{=} \left\{ \boldsymbol{\theta} \in \Theta_k : L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}) \le \mathfrak{z}^2/2 \right\} \subset I\!R^{p_k}. \tag{1.4}$$

Let $\mathfrak{z}_k(\alpha)$ denote the $(1-\alpha)$-quantile of the corresponding square-root likelihood ratio statistic:

$$\mathfrak{z}_k(\alpha) \stackrel{\text{def}}{=} \inf \left\{ \mathfrak{z} \ge 0 : I\!P \left(L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*) > \mathfrak{z}^2/2 \right) \le \alpha \right\}. \tag{1.5}$$

Together with (1.4) this implies for each $k = 1, \ldots, K$:

$$I\!P\Big(\boldsymbol{\theta}_k^* \in \mathcal{E}_k\left(\mathfrak{z}_k(\alpha)\right) \Big) \ge 1 - \alpha. \tag{1.6}$$

Thus $\mathcal{E}_k(\mathfrak{z})$ and the quantile function $\mathfrak{z}_k(\alpha)$ fully determine the marginal $(1-\alpha)$-confidence set. The simultaneous confidence set requires a correction for multiplicity. Let $\mathfrak{c}(\alpha)$ denote a maximal number $c \in (0, \alpha]$ s.t.

$$I\!P \left(\bigcup\nolimits_{k=1}^{K} \left\{ \sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > \mathfrak{z}_k(c) \right\} \right) \le \alpha. \tag{1.7}$$

This is equivalent to

$$\mathfrak{c}(\alpha) \stackrel{\text{def}}{=} \sup \left\{ c \in (0, \alpha] : I\!P \left(\max_{1 \le k \le K} \left\{ \sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} - \mathfrak{z}_k(c) \right\} > 0 \right) \le \alpha \right\}. \tag{1.8}$$

Therefore, taking the marginal confidence sets with the same confidence levels $1 - \mathfrak{c}(\alpha)$ yields the simultaneous confidence bound of the total level $1-\alpha$. The value $\mathfrak{c}(\alpha) \in (0, \alpha]$ is the correction for multiplicity. In order to construct the simultaneous confidence set using this correction, one has to estimate the values $\mathfrak{z}_k(\mathfrak{c}(\alpha))$ for all $k = 1, \ldots, K$. By its definition this problem splits into two subproblems:

1. **Marginal step.** Estimation of the marginal quantile functions $\mathfrak{z}_1(\alpha), \ldots, \mathfrak{z}_K(\alpha)$ given in (1.5).

2. **Correction for multiplicity.** Estimation of the correction for multiplicity $\mathfrak{c}(\alpha)$ given in (1.8).

If the 1-st problem is solved for any $\alpha \in (0,1)$, the 2-nd problem can be treated by calibrating the value α s.t. (1.8) holds. It is important to take into account the correlation between the likelihood ratio statistics $L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*)$, $k = 1, \ldots, K$, otherwise the estimate of the correction $\mathfrak{c}(\alpha)$ can be too conservative. For instance, the Bonferroni

correction would lead to the marginal confidence level $1-\alpha/K$, which may be very conservative if K is large and the statistics $L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}^*_k)$ are highly correlated.

In Section 2 we suggest a multiplier bootstrap procedure, which performs the steps 1 and 2 described above. Theoretical justification of the procedure is given in Section 3. The proofs are based on several approximation bounds: non-asymptotic square-root Wilks theorem, simultaneous Gaussian approximation for ℓ_2-norms, Gaussian comparison, and simultaneous Gaussian anti-concentration inequality.

Spokoiny and Zhilova (2014) considered the 1-st subproblem for the case of a single parametric model ($K=1$): a multiplier bootstrap procedure was applied for construction of a likelihood-based confidence set, and justified theoretically for a fixed sample size and for possibly misspecified parametric model. In the present paper we extend that approach for the case of simultaneously many parametric models.

Below we illustrate the definitions (1.2)-(1.8) of the simultaneous likelihood-based confidence sets with two popular examples.

Example 1 (Simultaneous confidence band for local constant regression): Let $Y_1, \ldots, Y_n$ be independent random scalar observations and $X_1, \ldots, X_n$ some deterministic design points. Consider the following quadratic likelihood function reweighted with the kernel functions $K(\cdot)$:

$$L(\boldsymbol{\theta}, x, h) \stackrel{\text{def}}{=} -\frac{1}{2}\sum\nolimits_{i=1}^{n}(Y_i - \boldsymbol{\theta})^2 w_i(x,h),$$

$$w_i(x,h) \stackrel{\text{def}}{=} K(\{x - X_i\}/h),$$

$$K(x) \in [0,1], \ \int_{I\!R} K(x)dx = 1, \ K(x) = K(-x).$$

Here $h > 0$ denotes bandwidth, the local smoothing parameter. The target point and the local MLE read as:

$$\boldsymbol{\theta}^*(x,h) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n w_i(x,h) I\!E Y_i}{\sum_{i=1}^n w_i(x,h)}, \qquad \widetilde{\boldsymbol{\theta}}(x,h) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n w_i(x,h) Y_i}{\sum_{i=1}^n w_i(x,h)}.$$

$\widetilde{\boldsymbol{\theta}}(x,h)$ is also known as Nadaraya-Watson estimate. Fix a bandwidth h and consider the range of points $x_1, \ldots, x_K$. They yield K local constant models with the target parameters $\boldsymbol{\theta}^*_k \stackrel{\text{def}}{=} \boldsymbol{\theta}^*(x_k, h)$ and the likelihood functions $L_k(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}, x_k, h)$ for $k = 1, \ldots, K$. The confidence intervals for each model are defined as

$$\mathcal{E}_k(\mathfrak{z}, h) \stackrel{\text{def}}{=} \left\{ \boldsymbol{\theta} \in \Theta : L(\widetilde{\boldsymbol{\theta}}(x_k,h), x_k, h) - L(\boldsymbol{\theta}, x_k, h) \le \mathfrak{z}^2/2 \right\},$$

for the quintile functions $\mathfrak{z}_k(\alpha)$ and for the multiplicity correction $\mathfrak{c}(\alpha)$ from (1.5) and (1.8) they form the following simultaneous confidence band:

$$I\!P\left(\bigcap\nolimits_{k=1}^{K} \left\{ \boldsymbol{\theta}^*_k \in \mathcal{E}_k\big(\mathfrak{z}_k\left(\mathfrak{c}(\alpha)\right)\big) \right\} \right) \ge 1 - \alpha.$$

In Section 4 we provide results of numerical experiments for this model.

Example 2 (Multiple quantiles regression): Quantile regression is an important method of statistical analysis, widely used in various applications. It aims at estimating conditional quantile functions of a response variable, see Koenker (2005). Multiple quantiles regression model considers simultaneously several quantile regression functions based on a range of quantile indices, see e.g. Liu and Wu (2011); Qu (2008); He (1997). Let $Y_1, \ldots, Y_n$ be independent random scalar observations and $X_1, \ldots, X_n \in \mathbb{R}^d$ some deterministic design points, as in Example 1. Consider the following quantile regression models for $k = 1, \ldots, K$:

$$Y_i = g_k(X_i) + \varepsilon_{k,i}, \quad i = 1, \ldots, n,$$

where $g_k(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ are unknown functions, the random values $\varepsilon_{k,1}, \ldots, \varepsilon_{k,n}$ are independent for each fixed k, and

$$\mathbb{P}(\varepsilon_{k,i} < 0) = \tau_k \quad \text{for all } i = 1, \ldots, n.$$

The range of quantile indices $\tau_1, \ldots, \tau_K \in (0,1)$ is known and fixed. We are interested in simultaneous parametric confidence sets for the functions $g_1(\cdot), \ldots, g_K(\cdot)$. Let $f_k(\boldsymbol{x}, \boldsymbol{\theta}) : \mathbb{R}^d \times \mathbb{R}^{p_k} \mapsto \mathbb{R}$ be known regression functions. Using the quantile regression approach by Koenker and Bassett Jr (1978), this problem can be treated with the quasi maximum likelihood method and the following log-likelihood functions:

$$L_k(\boldsymbol{\theta}) = -\sum_{i=1}^n \rho_{\tau_k}\left(Y_i - f_k(X_i, \boldsymbol{\theta})\right),$$
$$\rho_{\tau_k}(x) \stackrel{\text{def}}{=} x\left(\tau_k - \mathbb{1}\{x < 0\}\right).$$

for $k = 1, \ldots, K$. This quasi log-likelihood function corresponds to the Asymmetric Laplace distribution with the density $\tau_k(1-\tau_k)\mathrm{e}^{-\rho_{\tau_k}(x-a)}$. If $\tau = 1/2$, then $\rho_{1/2}(x) = |x|/2$ and $L(\boldsymbol{\theta}) = -\sum_{i=1}^n |Y_i - f_k(X_i, \boldsymbol{\theta})|/2$, which corresponds to the median regression.

The paper is organised as follows: Section 2 describes the multiplier bootstrap procedure, Section 3 explains the ideas of the theoretical approach and provides main results in Sections 3.1 and 3.2 correspondingly. All the necessary conditions are given in Section 5. In Section 5.3 and in statements of the main theoretical results we provide information about dependence of the involved terms on the sample size and parametric dimensions in the case of i.i.d. observations. Proofs of the main results are given in Section C. Statements from Sections A and B are used for the proofs in Section C. Numerical experiments are described in Section 4: we construct simultaneous confidence corridors for local constant and local quadratic regressions using both bootstrap and Monte Carlo

procedures. The quality of the bootstrap procedure is checked by computing the effective simultaneous coverage probabilities of the bootstrap confidence sets. We also compare the widths of the confidence bands and the values of multiplicity correction obtained with bootstrap and with Monte Carlo procedures. The experiments confirm that the multiplier bootstrap and the bootstrap multiplicity correction become conservative if the local parametric model is considerably misspecified.

The results given here are valid on a random set of probability $1 - Ce^{-\mathtt{x}}$ for some explicit constant $C > 0$. The number $\mathtt{x} > 0$ determines this dominating probability level. For the case of the i.i.d. observations (see Secion 5.3) we take $\mathtt{x} = \mathtt{C}\log n$. Throughout the text $\|\cdot\|$ denotes the Euclidean norm for a vector and spectral norm for a matrix. $\|\cdot\|_{\max}$ is the maximal absolute value of elements of a vector (or a matrix), $p_{\text{sum}} \stackrel{\text{def}}{=} p_1 + \cdots + p_K$, $p_{\max} \stackrel{\text{def}}{=} \max\limits_{1 \le k \le K} p_k$.

2 The multiplier bootstrap procedure

Let $\ell_{i,k}(\boldsymbol{\theta})$ denote the log-density from the k-th parametric distribution family evaluated at the i-th observation:

$$\ell_{i,k}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log\left(\frac{d\mathbb{P}_k(\boldsymbol{\theta})}{d\mu_0}(Y_i)\right), \tag{2.1}$$

then due to independence of $Y_1, \ldots, Y_n$

$$L_k(\boldsymbol{\theta}) = \sum\nolimits_{i=1}^{n} \ell_{i,k}(\boldsymbol{\theta}) \quad \forall\, k = 1, \ldots, K.$$

Consider i.i.d. scalar random variables u_i independent of the data $\boldsymbol{Y}$, s.t. $\mathbb{E}u_i = 1$, $\operatorname{Var} u_i = 1$, $\mathbb{E}\exp(u_i) < \infty$ (e.g. $u_i \sim \mathcal{N}(1,1)$ or $u_i \sim exp(1)$ or $u_i \sim 2Bernoulli(0.5)$). Multiply the summands of the likelihood function $L_k(\boldsymbol{\theta})$ with the new random variables:

$$L_k^{\circ}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^{n} \ell_{i,k}(\boldsymbol{\theta}) u_i, \tag{2.2}$$

then it holds $\mathbb{E}^{\circ} L_k^{\circ}(\boldsymbol{\theta}) = L_k(\boldsymbol{\theta})$, where $\mathbb{E}^{\circ}$ stands for the conditional expectation given $\boldsymbol{Y}$.

Therefore, the quasi MLE for the $\boldsymbol{Y}$-world is a target parameter for the bootstrap world for each $k = 1, \ldots, K$:

$$\operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_k} \mathbb{E}^{\circ} L_k^{\circ}(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_k} L_k(\boldsymbol{\theta}) = \widetilde{\boldsymbol{\theta}}_k.$$

The corresponding bootstrap MLE is:

$$\widetilde{\boldsymbol{\theta}}_k^{\circ} \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_k} L_k^{\circ}(\boldsymbol{\theta}).$$

The k-th likelihood ratio statistic in the bootstrap world equals to $L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)$, where all the elements: the function $L_k^\circ(\boldsymbol{\theta})$ and the arguments $\widetilde{\boldsymbol{\theta}}_k^\circ$, $\widetilde{\boldsymbol{\theta}}_k$ are known and available for computation. This means, that given the data $\boldsymbol{Y}$, one can estimate the distribution or quantiles of the statistic $L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)$ by generating many independent samples of the bootstrap weights $u_1, \ldots, u_n$ and computing with them the bootstrap likelihood ratio.

Let us introduce similarly to (1.5) the $(1-\alpha)$-quantile for the bootstrap square-root likelihood ratio statistic:

$$\mathfrak{z}_k^\circ(\alpha) \stackrel{\text{def}}{=} \inf\left\{\mathfrak{z} \geq 0 : I\!P^\circ\left(L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k) > \mathfrak{z}^2/2\right) \leq \alpha\right\}, \tag{2.3}$$

here $I\!P^\circ$ denotes probability measure conditional on the data $\boldsymbol{Y}$, therefore, $\mathfrak{z}_k^\circ(\alpha)$ is a random value dependent on $\boldsymbol{Y}$.

Spokoiny and Zhilova (2014) considered the case of a single parametric model ($K = 1$), and showed that the bootstrap quantile $\mathfrak{z}_k^\circ(\alpha)$ is close to the true one $\mathfrak{z}_k(\alpha)$ under a so called "Small Modeling Bias" (SmB) condition, which is fulfilled when the true distribution is close to the parametric family or when the observations are i.i.d. When the SmB condition does not hold, the bootstrap quantile is still valid, however, it becomes conservative. Therefore, for each fixed $k = 1, \ldots, K$ the bootstrap quantiles $\mathfrak{z}_k^\circ(\alpha)$ are rather good estimates for the true unknown ones $\mathfrak{z}_k(\alpha)$, however, they are still "pointwise" in k, i.e. the confidence bounds (1.6) hold for each k separately. Our goal here is to estimate $\mathfrak{z}_1(\alpha), \ldots, \mathfrak{z}_K(\alpha)$ and $\mathfrak{c}(\alpha)$ according to (1.7) and (1.8). Let us introduce the bootstrap correction for multiplicity:

$$\mathfrak{c}^\circ(\alpha) \stackrel{\text{def}}{=} \sup\left\{c \in (0,\alpha] : I\!P^\circ\left(\bigcup\nolimits_{k=1}^K \left\{\sqrt{2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - 2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)} > \mathfrak{z}_k^\circ(c)\right\}\right) \leq \alpha\right\}. \tag{2.4}$$

By its definition $\mathfrak{c}^\circ(\alpha)$ depends on the random sample $\boldsymbol{Y}$.

The multiplier bootstrap procedure below explains how to estimate the bootstrap quantile functions $\mathfrak{z}_k^\circ(\mathfrak{c}^\circ(\alpha))$ corrected for multiplicity.

The simultaneous bootstrap procedure:

Input: The data $\boldsymbol{Y}$ (as in (1.1)) and a fixed confidence level $(1-\alpha) \in (0,1)$.

Step 1: Generate B independent samples of i.i.d. bootstrap weights $\{u_1^{(b)}, \ldots, u_n^{(b)}\}$, $b = 1, \ldots, B$. For the bootstrap likelihood processes

$$L_k^{\circ(b)}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \ell_{i,k}(\boldsymbol{\theta}) u_i^{(b)}. \tag{2.5}$$

compute the bootstrap likelihood ratios $L_k^{\circ(b)}(\boldsymbol{\theta}_k^{\circ(b)}) - L_k^{\circ(b)}(\widetilde{\boldsymbol{\theta}}_k)$. For each fixed b the bootstrap likelihoods $L_1^{\circ(b)}(\boldsymbol{\theta}), \ldots, L_K^{\circ(b)}(\boldsymbol{\theta})$ are computed using

the same bootstrap sample $\{u_i^{(b)}\}$, s.t. the i-th summand $\ell_{i,k}(\boldsymbol{\theta})$ is always multiplied with the i-th weight $u_i^{(b)}$ as in (2.5).

Step 2: Estimate the marginal quantile functions $\mathfrak{z}_k^\circ(\alpha)$ defined in (2.3) separately for each $k = 1, \ldots, K$, using B bootstrap realisations of $L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)$ from **Step 1**.

Step 3: Find by an iterative procedure the maximum value $c \in (0, \alpha]$ s.t.

$$\mathbb{P}^\circ\left(\bigcup\nolimits_{k=1}^K \left\{\sqrt{2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - 2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)} \geq \mathfrak{z}_k^\circ(c)\right\}\right) \leq \alpha.$$

Otput: The resulting critical values are $\mathfrak{z}_k^\circ(c)$, $k = 1, \ldots, K$.

Remark 2.1. The requirement in Step 1 to use the same bootstrap sample $\{u_i^{(b)}\}$ for generation of the bootstrap likelihood ratios $L_k^{\circ(b)}(\boldsymbol{\theta}_k^{\circ(b)}) - L_k^{\circ(b)}(\widetilde{\boldsymbol{\theta}}_k)$, $k = 1, \ldots, K$ allows to preserve the correlation structure between the ratios and, therefore, to make a sharper simultaneous adjustment in Step 3.

This procedure is justified theoretically in the next section.

3 Theoretical justification of the bootstrap procedure

Before stating the main results in Section 3.2 we introduce in Section 3.1 the basic ingredients of the proofs. The general scheme of the theoretical approach here is taken from Spokoiny and Zhilova (2014). In the present work we extend that approach for the case of simultaneously many parametric models.

3.1 Overview of the theoretical approach

For justification of the described multiplier bootstrap procedure for simultaneous inference it has to be checked that the joint distributions of the sets of likelihood ratio statistics $\left\{L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*) : k = 1, \ldots, K\right\}$ and $\left\{L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k) : k = 1, \ldots, K\right\}$ are close to each other. These joint distributions are approximated using several non-asymptotic

steps given in the following scheme:

$$\begin{array}{lccccc} & & \text{uniform sq-Wilks theorem} & & \text{joint Gauss. approx. \& anti-concentr.}^{*} & \\ \boldsymbol{Y}\text{-world:} & \sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}^*_k)} & \underset{\frac{p_k+\log K}{\sqrt{n}}}{\approx} & \|\boldsymbol{\xi}_k\| & \approx & \|\overline{\boldsymbol{\xi}}_k\| \\ & & & \bigcap_{1\le k\le K} & & \wr\, w \ \text{simultaneous Gauss. compar.}^{**} \\ \text{Bootstrap world:} & \sqrt{2L^{\circ}_k(\widetilde{\boldsymbol{\theta}}^{\circ}_k) - 2L^{\circ}_k(\widetilde{\boldsymbol{\theta}}_k)} & \underset{\frac{p_k+\log K}{\sqrt{n}}}{\approx} & \|\boldsymbol{\xi}^{\circ}_k\| & \approx & \|\overline{\boldsymbol{\xi}}^{\circ}_k\|, \end{array} \tag{3.1}$$

* the accuracy of these approximating steps is $\mathtt{C}\left\{\frac{p^3_{\max}}{n}\log^9(K)\log^3(np_{\text{sum}})\right\}^{1/8}$;

** Gaussian comparison step yields an approximation error proportional to $\widehat{\delta}^2_{\text{smb}}\left(\frac{p^3_{\max}}{n}\right)^{1/4} p_{\max}\log^2(K)\log^{3/4}(np_{\text{sum}})$, where $\widehat{\delta}^2_{\text{smb}}$ comes from condition **(SmB)**, see also (3.4) below.

Here $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}^{\circ}_k$ denote normalized score vectors for the $\boldsymbol{Y}$ and bootstrap likelihood processes:

$$\boldsymbol{\xi}_k \stackrel{\text{def}}{=} D_k^{-1}\nabla_{\boldsymbol{\theta}} L_k(\boldsymbol{\theta}^*_k), \qquad \boldsymbol{\xi}^{\circ}_k \stackrel{\text{def}}{=} \boldsymbol{\xi}^{\circ}_k(\boldsymbol{\theta}^*_k) \stackrel{\text{def}}{=} D_k^{-1}\nabla_{\boldsymbol{\theta}} L_k(\boldsymbol{\theta}^*_k), \tag{3.2}$$

D^2_k is the full Fisher information matrix for the corresponding k-th likelihood:

$$D^2_k \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!\!E L_k(\boldsymbol{\theta}^*_k).$$

$\overline{\boldsymbol{\xi}}_k \sim \mathcal{N}(0, \operatorname{Var}\boldsymbol{\xi}_k)$ and $\overline{\boldsymbol{\xi}}^{\circ}_k \sim \mathcal{N}(0, \operatorname{Var}^{\circ}\boldsymbol{\xi}^{\circ}_k)$ denote approximating Gaussian vectors, which have the same covariance matrices as $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^{\circ}$. Moreover the vectors $\left(\overline{\boldsymbol{\xi}}_1^{\top}, \ldots, \overline{\boldsymbol{\xi}}_K^{\top}\right)^{\top}$ and $\left(\overline{\boldsymbol{\xi}}^{\circ\top}_1, \ldots, \overline{\boldsymbol{\xi}}^{\circ\top}_K\right)^{\top}$ are normally distributed and have the same covariance matrices as the vectors $\left(\boldsymbol{\xi}_1^{\top}, \ldots, \boldsymbol{\xi}_K^{\top}\right)^{\top}$ and $\left(\boldsymbol{\xi}^{\circ\top}_1, \ldots, \boldsymbol{\xi}^{\circ\top}_K\right)^{\top}$ correspondingly. $\operatorname{Var}^{\circ}$ and $\operatorname{Cov}^{\circ}$ denote variance and covariance operators w.r.t. the probability measure $I\!\!P^{\circ}$ conditional on $\boldsymbol{Y}$.

The first two approximating steps: square root Wilks and Gaussian approximations are performed in parallel for both $\boldsymbol{Y}$ and bootstrap worlds, which is shown in the corresponding lines of the scheme (3.1). The two worlds are connected in the last step: Gaussian comparison for ℓ_2-norms of Gaussian vectors. All the approximations are performed simultaneously for K parametric models.

Let us consider each step in more details. Non-asymptotic square-root Wilks approximation result had been obtained recently by Spokoiny (2012a, 2013). It says that for

a fixed sample size and misspecified parametric assumption: $I\!P \notin \{I\!P_k\}$, it holds with exponentially high probablity:

$$\left| \sqrt{2\{L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*)\}} - \|\boldsymbol{\xi}_k\| \right| \leq \Delta_{k,\mathrm{W}} \simeq \frac{p_k}{\sqrt{n}},$$

here the index k is fixed, i.e. this statement is for one parametric model. The precise statement of this result is given in Section B.1, and its simultaneous version – in Section B.3. The approximating value $\|\boldsymbol{\xi}_k\|$ is ℓ_2-norm of the score vector $\boldsymbol{\xi}_k$ given in (3.2). The next approximating step is between the joint distributions of $\|\boldsymbol{\xi}_1\|, \ldots, \|\boldsymbol{\xi}_K\|$ and $\|\overline{\boldsymbol{\xi}}_1\|, \ldots, \|\overline{\boldsymbol{\xi}}_K\|$. This is done in Section A.1 for general centered random vectors under bounded exponential moments assumptions. The main tools for the simultaneous Gaussian approximation are: Lindeberg's telescopic sum, smooth maximum function and three times differentiable approximation of the indicator function $\mathbb{1}\{x \in I\!R : x > 0\}$. The simultaneous anti-concentration inequality for the ℓ_2-norms of Gaussian vectors is obtained in Section A.3. The result is based on approximation of the ℓ_2-norm with a maximum over a finite grid on a hypersphere, and on the anti-concentration inequality for maxima of a Gaussian random vector by Chernozhukov et al. (2014c). The same approximating steps are performed for the bootstrap world, the square-root bootstrap Wilks approximation is given in Sections B.2, B.3. The last step in the scheme (3.1) is comparison of the joint distributions of the sets of ℓ_2-norms of Gaussian vectors: $\|\overline{\boldsymbol{\xi}}_1\|, \ldots, \|\overline{\boldsymbol{\xi}}_K\|$ and $\|\overline{\boldsymbol{\xi}}_1^{\circ}\|, \ldots, \|\overline{\boldsymbol{\xi}}_K^{\circ}\|$ by Slepian interpolation (see Section A.2 for the result in a general setting). The error of approximation is proportional to

$$\max_{1 \leq k_1, k_2 \leq K} \left\| \mathrm{Cov}(\boldsymbol{\xi}_{k_1}, \boldsymbol{\xi}_{k_2}) - \mathrm{Cov}^{\circ}(\boldsymbol{\xi}_{k_1}^{\circ}, \boldsymbol{\xi}_{k_2}^{\circ}) \right\|_{\max}. \tag{3.3}$$

It is shown, using Bernstein matrix inequality (Sections C.1 and C.3), that the value (3.3) is bounded from above (up to a constant) on a random set of dominating probability with

$$\max_{1 \leq k \leq K} \left\| H_k^{-1} B_k^2 H_k^{-1} \right\| \leq \widehat{\delta}_{\mathrm{smb}}^2 \tag{3.4}$$

for

$$\begin{aligned} B_k^2 &\stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n I\!E \left\{ \nabla_{\boldsymbol{\theta}} \ell_{i,k}(\boldsymbol{\theta}_k^*) \right\} I\!E \left\{ \nabla_{\boldsymbol{\theta}} \ell_{i,k}(\boldsymbol{\theta}_k^*) \right\}^{\top}, \\ H_k^2 &\stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n I\!E \left\{ \nabla_{\boldsymbol{\theta}} \ell_{i,k}(\boldsymbol{\theta}_k^*) \nabla_{\boldsymbol{\theta}} \ell_{i,k}(\boldsymbol{\theta}_k^*)^{\top} \right\}. \end{aligned} \tag{3.5}$$

The value $\left\| H_k^{-1} B_k^2 H_k^{-1} \right\|$ is responsible for the modelling bias of the k-th model. If the parametric family $\{I\!P_k(\boldsymbol{\theta})\}$ contains the true distribution $I\!P$ or if the observations Y_i are i.i.d., then B_k^2 equals to zero. Condition **($\widehat{\mathbf{SmB}}$)** assumes that all the values $\left\| H_k^{-1} B_k^2 H_k^{-1} \right\|$ are rather small.

3.2 Main results

The following theorem shows the closeness of the joint cumulative distribution functions (c.d.f-s.) of $\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}^*_k)}, k = 1,\ldots,K\right\}$ and $\left\{\sqrt{2L^\circ_k(\widetilde{\boldsymbol{\theta}}^\circ_k) - 2L^\circ_k(\widetilde{\boldsymbol{\theta}}_k)}, k = 1,\ldots,K\right\}$. The approximating error term Δ_{total} equals to a sum of the errors from all the steps in the scheme (3.1).

Theorem 3.1. *Under the conditions of Section 5 it holds with probability* $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$ *for* $z_k \geq C\sqrt{p_k}$, $1 \leq C < 2$

$$\left| I\!P\left(\bigcup\nolimits_{k=1}^K \left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}^*_k)} > z_k\right\}\right) - I\!P^\circ\left(\bigcup\nolimits_{k=1}^K \left\{\sqrt{2L^\circ_k(\widetilde{\boldsymbol{\theta}}^\circ_k) - 2L^\circ_k(\widetilde{\boldsymbol{\theta}}_k)} > z_k\right\}\right)\right| \leq \Delta_{\text{total}}.$$

The approximating total error $\Delta_{\text{total}} \geq 0$ *is deterministic and in the case of i.i.d. observations (see Section 5.3) it holds:*

$$\Delta_{\text{total}} \leq \mathtt{C}\left(\frac{p^3_{\max}}{n}\right)^{1/8} \log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\left\{\left(\widehat{\mathfrak{a}}^2 + \widehat{\mathfrak{a}}^2_B\right)\left(1 + \delta^2_{\widehat{\mathcal{V}}}(\mathtt{x})\right)\right\}^{3/8}, \quad (3.6)$$

where the deterministic terms $\widehat{\mathfrak{a}}^2, \widehat{\mathfrak{a}}^2_B$ *and* $\delta^2_{\widehat{\mathcal{V}}}(\mathtt{x})$ *come from the conditions* $\widehat{(\boldsymbol{\mathcal{I}})}$, $\widehat{(\boldsymbol{\mathcal{I}_B})}$ *and* $\widehat{(\boldsymbol{SD_1})}$. Δ_{total} *is defined in* (C.5).

Remark 3.1. The obtained approximation bound is mainly of theoretical interest, although it shows the impact of $p_{\max}$, K and n on the quality of the bootstrap procedure. For more details on the error term see Remark A.1.

The next theorem justifies the bootstrap procedure under the $\widehat{(\mathbf{SmB})}$ condition. The theorem says that the bootstrap quantile functions $\mathfrak{z}^\circ_k(\cdot)$ with the bootstrap-corrected for multiplicity confidence levels $1 - \mathfrak{c}^\circ(\alpha)$ can be used for construction of the simultaneous confidence set in the $\boldsymbol{Y}$-world.

Theorem 3.2 (Bootstrap validity for a small modeling bias)**.** *Assume the conditions of Theorem 3.1, and* $\mathfrak{c}(\alpha), 0.5\mathfrak{c}^\circ(\alpha) \geq \Delta_{\text{full},\max}$, *then for* $\alpha \leq 1 - 8\mathrm{e}^{-\mathtt{x}}$ *it holds with probability* $1 - 12\mathrm{e}^{-\mathtt{x}}$

$$I\!P\left(\bigcup\nolimits_{k=1}^K \left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}^*_k)} \geq \mathfrak{z}^\circ_k\left(\mathfrak{c}^\circ(\alpha) - 2\Delta_{\text{full},\max}\right)\right\}\right) - \alpha \leq \Delta_{\mathfrak{z},\text{total}},$$

$$I\!P\left(\bigcup\nolimits_{k=1}^K \left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}^*_k)} \geq \mathfrak{z}^\circ_k\left(\mathfrak{c}^\circ(\alpha) + 2\Delta_{\text{full},\max}\right)\right\}\right) - \alpha \geq -\Delta_{\mathfrak{z},\text{total}},$$

where $\Delta_{\text{full},\max} \leq \mathtt{C}\{(p_{\max} + \mathtt{x})^3/n\}^{1/8}$ *in the case of i.i.d. observations (see Section 5.3), and* $\Delta_{\mathfrak{z},\text{total}} \leq 3\Delta_{\text{total}}$; *their explicit definitions are given in* (C.11) *and* (C.14).

Moreover

$$\mathfrak{c}^{\circ}(\alpha) \le \mathfrak{c}(\alpha + \Delta_{\mathfrak{c}}) + \Delta_{\text{full, max}},$$
$$\mathfrak{c}^{\circ}(\alpha) \ge \mathfrak{c}(\alpha - \Delta_{\mathfrak{c}}) - \Delta_{\text{full, max}},$$

for $0 \le \Delta_{\mathfrak{c}} \le 2\Delta_{\text{total}}$ *, defined in* (C.15).

The following theorem does not assume the **(SmB)** condition to be fulfilled. It turns out that in this case the bootstrap procedure becomes conservative, and the bootstrap critical values corrected for the multiplicity $\mathfrak{z}_k^{\circ}(\mathfrak{c}^{\circ}(\alpha))$ are increased with the modelling bias $\sqrt{\text{tr}\{D_k^{-1}H_k^2D_k^{-1}\}} - \sqrt{\text{tr}\{D_k^{-1}(H_k^2 - B_k^2)D_k^{-1}\}}$, therefore, the confidence set based on the bootstrap estimates can be conservative.

Theorem 3.3 (Bootstrap conservativeness for a large modeling bias)**.** *Under the conditions of Section 5 except for* **(SmB)** *it holds with probability* $\ge 1 - 14\mathrm{e}^{-\mathtt{x}}$ *for* $z_k \ge C\sqrt{p_k}$, $1 \le C < 2$

$$\mathbb{P}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > z_k\right\}\right)$$
$$\le \mathbb{P}^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} > z_k\right\}\right) + \Delta_{\text{b, total}}.$$

The deterministic value $\Delta_{\text{b, total}} \in [0, \Delta_{\text{total}}]$ *(see* (3.6) *in the case 5.3). Moreover, the bootstrap-corrected for multiplicity confidence level* $1 - \mathfrak{c}^{\circ}(\alpha)$ *is conservative in comparison with the true corrected confidence level:*

$$1 - \mathfrak{c}^{\circ}(\alpha) \ge 1 - \mathfrak{c}(\alpha + \Delta_{\text{b},\mathfrak{c}}) - \Delta_{\text{full, max}},$$

and it holds for all $k = 1, \ldots, K$ *and* $\alpha \le 1 - 8\mathrm{e}^{-\mathtt{x}}$

$$\mathfrak{z}_k^{\circ}(\mathfrak{c}^{\circ}(\alpha)) \ge \mathfrak{z}_k(\mathfrak{c}(\alpha + \Delta_{\text{b},\mathfrak{c}}) + \Delta_{\text{full, max}})$$
$$+ \sqrt{\text{tr}\{D_k^{-1}H_k^2D_k^{-1}\}} - \sqrt{\text{tr}\{D_k^{-1}(H_k^2 - B_k^2)D_k^{-1}\}} - \Delta_{\mathtt{qf},1,k},$$

for $0 \le \Delta_{\text{b},\mathfrak{c}} \le 2\Delta_{\text{total}}$ *, defined in* (C.18)*, and the positive value* $\Delta_{\mathtt{qf},1,k}$ *is bounded from above with* $(\mathfrak{a}_k^2 + \mathfrak{a}_{B,k}^2)(\sqrt{8\mathtt{x}p_k} + 6\mathtt{x})$ *for the constants* $\mathfrak{a}_k^2 > 0$, $\mathfrak{a}_{B,k}^2 \ge 0$ *from conditions* $(\boldsymbol{\mathcal{I}})$ *,* $(\boldsymbol{\mathcal{I}_B})$ *.*

The **(SmB)** condition is automatically fulfilled if all the parametric models are correct or in the case of i.i.d. observations. This condition is checked for generalised linear model and linear quantile regression in Spokoiny and Zhilova (2014) (the version of 2015).

4 Numerical experiments

Here we check the performance of the bootstrap procedure by constructing simultaneous confidence sets based on the local constant and local quadratic estimates, the former one is also known as Nadaraya-Watson estimate Nadaraya (1964); Watson (1964). Let $Y_1, \ldots, Y_n$ be independent random scalar observations and $X_1, \ldots, X_n$ some deterministic design points. In Sections 4.1-4.3 below we introduce the models and the data, Sections 4.4-4.6 present the results of the experiments.

4.1 Local constant regression

Consider the following quadratic likelihood function reweighted with the kernel functions $K(\cdot)$:

$$L(\boldsymbol{\theta}, x, h) \stackrel{\text{def}}{=} -\frac{1}{2}\sum\nolimits_{i=1}^{n}(Y_i - \boldsymbol{\theta})^2 w_i(x,h),$$
$$w_i(x,h) \stackrel{\text{def}}{=} K(\{x - X_i\}/h),$$
$$K(x) \in [0,1], \ \int_{I\!R} K(x)dx = 1, \ K(x) = K(-x).$$

Here $h > 0$ denotes bandwidth, the local smoothing parameter. The target point and the local MLE read as:

$$\boldsymbol{\theta}^*(x,h) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n w_i(x,h) I\!E Y_i}{\sum_{i=1}^n w_i(x,h)}, \qquad \widetilde{\boldsymbol{\theta}}(x,h) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n w_i(x,h) Y_i}{\sum_{i=1}^n w_i(x,h)}.$$

Let us fix a bandwidth h and consider the range of points $x_1, \ldots, x_K$. They yield K local constant models with the target parameters $\boldsymbol{\theta}_k^* \stackrel{\text{def}}{=} \boldsymbol{\theta}^*(x_k, h)$ and the likelihood functions $L_k(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}, x_k, h)$ for $k = 1, \ldots, K$.

The bootstrap local likelihood function is defined similarly to the global one (2.2), by reweighting $L(\boldsymbol{\theta}, x, h)$ with the bootstrap multipliers $u_1, \ldots, u_n$:

$$L_k^\circ(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L^\circ(\boldsymbol{\theta}, x_k, h) \stackrel{\text{def}}{=} -\frac{1}{2}\sum\nolimits_{i=1}^{n}(Y_i - \boldsymbol{\theta})^2 w_i(x_k,h) u_i,$$
$$\widetilde{\boldsymbol{\theta}}_k^\circ \stackrel{\text{def}}{=} \widetilde{\boldsymbol{\theta}}^\circ(x_k,h) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n w_i(x_k,h) u_i Y_i}{\sum_{i=1}^n w_i(x_k,h) u_i}.$$

4.2 Local quadratic regression

Here the local likelihood function reads as

$$L(\boldsymbol{\theta}, x, h) \stackrel{\text{def}}{=} -\frac{1}{2}\sum\nolimits_{i=1}^{n}(Y_i - \Psi_i^\top \boldsymbol{\theta})^2 w_i(x,h),$$
$$\boldsymbol{\theta}, \Psi_i \in I\!R^3, \quad \Psi_i \stackrel{\text{def}}{=} \left(1, X_i, X_i^2\right)^\top,$$

and

$$\boldsymbol{\theta}^*(x,h) \stackrel{\text{def}}{=} \left(\Psi W(x,h)\Psi^\top\right)^{-1}\Psi W(x,h)I\!\!E\boldsymbol{Y},$$
$$\widetilde{\boldsymbol{\theta}}(x,h) \stackrel{\text{def}}{=} \left(\Psi W(x,h)\Psi^\top\right)^{-1}\Psi W(x,h)\boldsymbol{Y},$$

where

$$\boldsymbol{Y} \stackrel{\text{def}}{=} (Y_1,\ldots,Y_n)^\top, \quad \Psi \stackrel{\text{def}}{=} (\Psi_1,\ldots,\Psi_n) \in I\!R^{3\times n},$$
$$W(x,h) \stackrel{\text{def}}{=} \operatorname{diag}\{w_1(x,h),\ldots,w_n(x,h)\}.$$

And similarly for the bootstrap objects

$$L^\bullet(\boldsymbol{\theta},x,h) \stackrel{\text{def}}{=} -\frac{1}{2}\sum\nolimits_{i=1}^n (Y_i-\Psi_i^\top\boldsymbol{\theta})^2 w_i(x,h)u_i,$$
$$\widetilde{\boldsymbol{\theta}}^\bullet(x,h) \stackrel{\text{def}}{=} \left(\Psi UW(x,h)\Psi^\top\right)^{-1}\Psi UW(x,h)\boldsymbol{Y},$$

for $U \stackrel{\text{def}}{=} \operatorname{diag}\{u_1,\ldots,u_n\}$.

4.3 Simulated data

In the numerical experiments we constructed two 90% simultaneous confidence bands: using Monte Carlo (MC) samples and bootstrap procedure with Gaussian weights ($u_i \sim \mathcal{N}(1,1)$), in each case we used 10^4 $\{Y_i\}$ and 10^4 $\{u_i\}$ independent samples. The sample size $n = 400$. $K(x)$ is Epanechnikov's kernel function. The independent random observations Y_i are generated as follows:

$$Y_i = f(X_i) + \mathcal{N}(0,1), \quad X_i \text{ are equidistant on } [0,1], \tag{4.1}$$

$$f(x) = \begin{cases} 5, & x \in [0,0.25]\cup[0.65,1]; \\ 5+3.8\{1-100(x-0.35)^2\}, & x\in[0.25,0.45]; \\ 5-3.8\{1-100(x-0.55)^2\}, & x\in[0.45,0.65]. \end{cases} \tag{4.2}$$

The number of local models $K = 71$, the points $x_1,\ldots,x_{71}$ are equidistant on $[0,1]$. For the bandwidth we considered two cases: $h = 0.12$ and $h = 0.3$.

4.4 Effect of the modeling bias on a width of a bootstrap confidence band

The function $f(x)$ defined in (4.2) should yield a considerable modeling bias for both mean constant and mean quadratic estimators. Figures 4.1, 4.2 demonstrate that the bootstrap confidence bands become conservative (i.e. wider than the MC confidence

band) when the local model is misspecified. The top graphs on Figures 4.1, 4.2 show the 90% confidence bands, the middle graphs show their width, and the bottom graphs show the value of the modelling bias for $K = 71$ local models (see formulas (4.3) and (4.4) below). For the local constant estimate (Figure 4.1) the width of the bootstrap confidence sets is considerably increased by the modeling bias when $x \in [0.25, 0.65]$. In this case case the expression for the modeling bias term for the k-th model (see also **(SmB)** condition) reads as:

$$\begin{aligned} \left|H_k^{-1} B_k^2 H_k^{-1}\right| &= \frac{\sum_{i=1}^n \left\{I\!\!E Y_i - \boldsymbol{\theta}^*(x_k)\right\}^2 w_i^2(x_k, h)}{\sum_{i=1}^n I\!\!E \left\{Y_i - \boldsymbol{\theta}^*(x_k)\right\}^2 w_i^2(x_k, h)} \\ &= 1 - \left(1 + \frac{\sum_{i=1}^n w_i^2(x_k, h) \left\{f(X_i) - \boldsymbol{\theta}^*(x_k)\right\}^2}{\sum_{i=1}^n w_i^2(x_k, h)}\right)^{-1}. \end{aligned} \tag{4.3}$$

And for the local quadratic estimate it holds:

$$\left\|H_k^{-1} B_k^2 H_k^{-1}\right\| = \left\|\boldsymbol{I}_p - H_k^{-1} \left\{\sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top w_i^2(x_k, h)\right\} H_k^{-1}\right\|, \tag{4.4}$$

where $\boldsymbol{I}_p$ is the identity matrix of dimension $p \times p$ (here $p = 3$), and

$$\begin{aligned} H_k^2 &= \sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top w_i^2(x_k, h) I\!\!E \left\{Y_i - \boldsymbol{\theta}^*(x_k)\right\}^2 \\ &= \sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top w_i^2(x_k, h) \left\{f(X_i) - \boldsymbol{\theta}^*(x_k)\right\}^2 + \sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top w_i^2(x_k, h). \end{aligned} \tag{4.5}$$

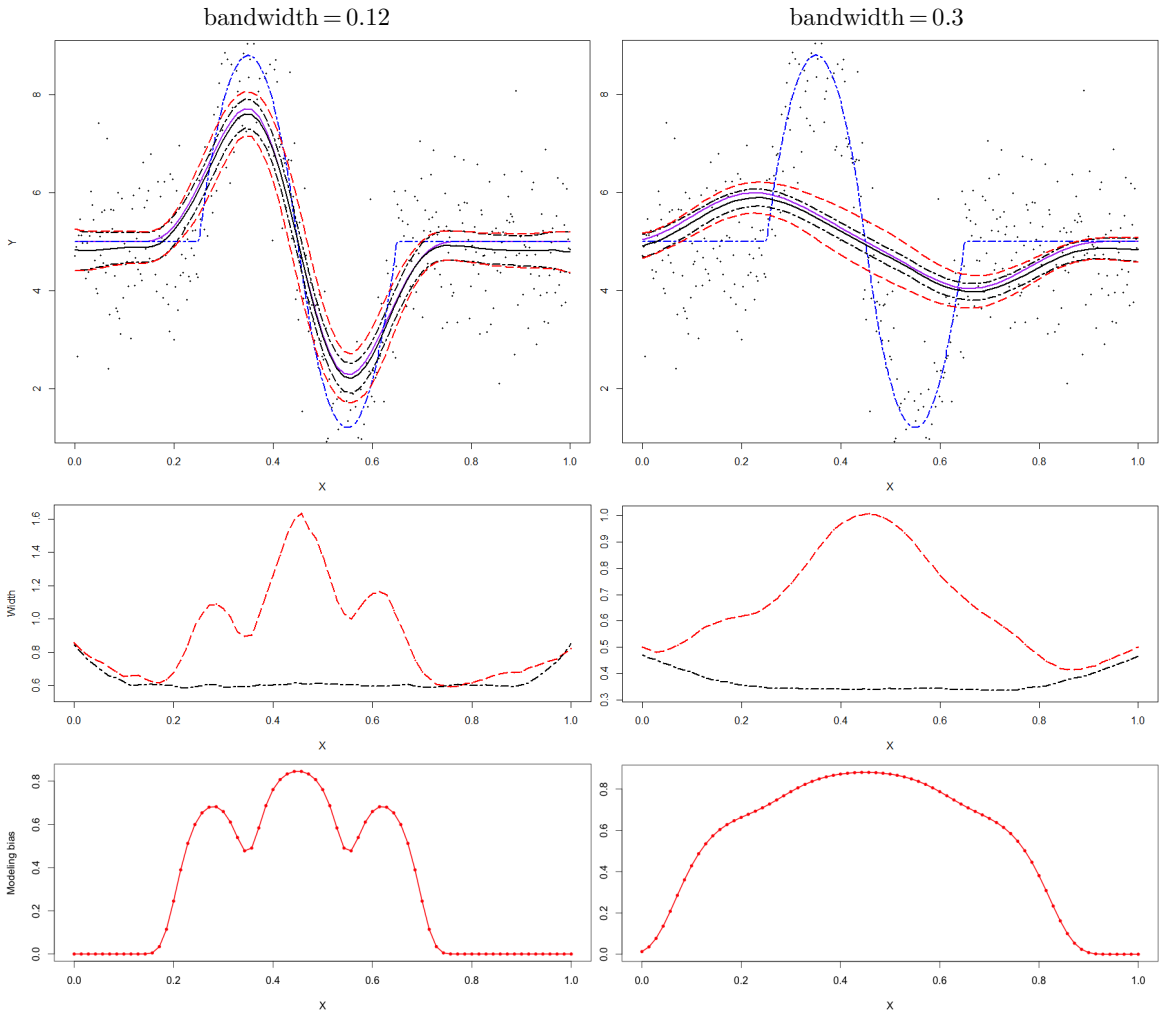
Therefore, if $\max_{1 \le k \le K} \left\{f(X_i) - \boldsymbol{\theta}^*(x_k)\right\}^2 = 0$, then $\left\|H_k^{-1} B_k^2 H_k^{-1}\right\| = 0$. On the Figure 4.1 both the modelling bias and the difference between the widths of the bootstrap and MC confidence bands are close to zero in the regions where the true function $f(x)$ is constant. On Figure 4.2 the modelling bias for $h = 0.12$ is overall smaller than the corresponding value on Figure 4.1. For the bigger bandwidth $h = 0.3$ the modelling biases on Figures 4.1 and 4.2 are comparable with each other.

Thus the numerical experiment is consistent with the theoretical results from Section 3.2, and confirm that in the case when a (local) parametric model is close to the true distribution the simultaneous bootstrap confidence set is valid. Otherwise the bootstrap procedure is conservative: the modelling bias widens the simultaneous bootstrap confidence set.

4.5 Effective coverage probability (local constant estimate)

In this part of the experiment we check the bootstrap validity by computing the effective coverage probability values. This requires to perform many independent experiments: for each of independent 5000 $\{Y_i\} \sim$ (4.1) samples we took 10^4 independent bootstrap samples $\{u_i\} \sim \mathcal{N}(1, 1)$, and constructed simultaneous bootstrap confidence sets for a range of confidence levels. The second row of Table 4.1 contains this range $(1 - \alpha) =$

Figure 4.1: **Local constant regression:**
Confidence bands, their widths, and the modeling bias



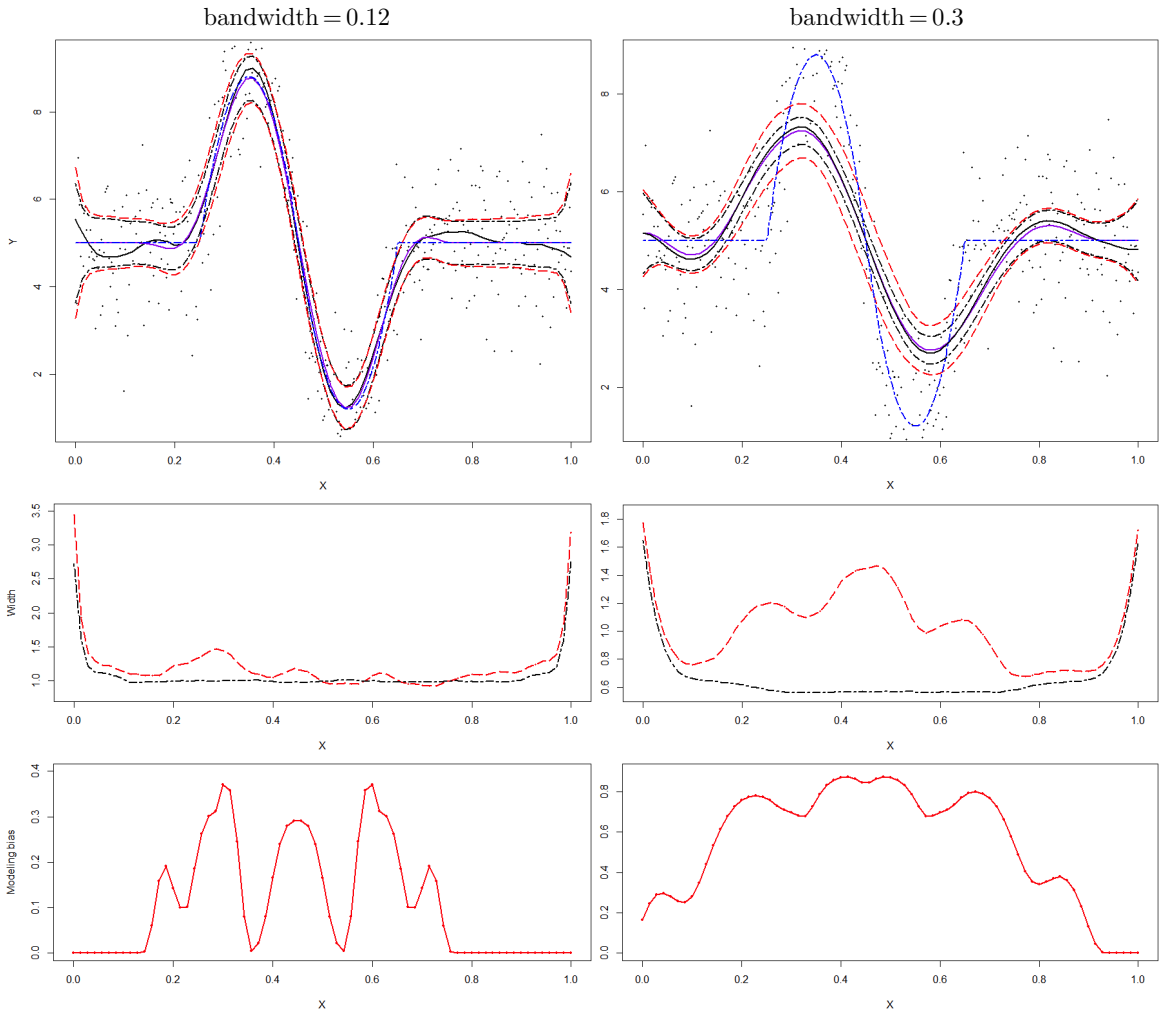
Legend for the top graphs:

- 90% bootstrap simultaneous confidence band
- the true function $f(x)$
- 90% MC simultaneous confidence band
- local constant MLE
- smoothed target function

Legend for the middle and the bottom graphs:

- width of the 90% bootstrap confidence bands from the upper graphs
- width of the 90% MC confidence bands from the upper graphs
- modeling bias from the expression (4.3)

Figure 4.2: **Local quadratic regression:**
Confidence bands, their widths, and the modeling bias



Legend for the top graphs:

- 90% bootstrap simultaneous confidence band
- the true function $f(x)$
- 90% MC simultaneous confidence band
- local constant MLE
- smoothed target function

Legend for the middle and the bottom graphs:

- width of the 90% bootstrap confidence bands from the upper graphs
- width of the 90% MC confidence bands from the upper graphs
- modeling bias from the expression (4.4)

$0.95, 0.9, \dots, 0.5$. The third and the fourth rows of Table 4.1 show the frequencies of the event

$$\max_{1\le k\le K}\left\{L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}^*_k) - \mathfrak{z}^{\circ}_k(\mathfrak{c}^{\circ}(\alpha))\right\} \le 0$$

among 5000 data samples, for the bandwidths $h = 0.12, 0.3$, and for the range of $(1-\alpha)$. The results show that the bootstrap procedure is rather conservative for both $h = 0.12$ and $h = 0.3$, however, the larger bandwidth yields bigger coverage probabilities.

Table 1: Effective coverage probabilities for the local constant regression

	Confidence levels									
h	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.65**	**0.60**	**0.55**	**0.50**
0.12	0.971	0.947	0.917	0.888	0.863	0.830	0.800	0.769	0.738	0.702
0.3	0.982	0.963	0.942	0.918	0.895	0.868	0.842	0.815	0.784	0.750

4.6 Correction for multiplicity

Here we compare the $\boldsymbol{Y}$ and the bootstrap corrections for multiplicity, i.e. the values $\mathfrak{c}(\alpha)$ and $\mathfrak{c}^{\circ}(\alpha)$ defined in (1.8) and (2.4). The numerical results in Tables 2, 3 are based on 10^4 $\{Y_i\} \sim$ (4.1) independent samples and 10^4 independent bootstrap samples $\{u_i\} \sim \mathcal{N}(1,1)$. The second line in Tables 2, 3 contains the range of the nominal confidence levels $(1-\alpha) = 0.95, 0.9, \dots, 0.5$ (similarly to the Table 1). The first column contains the values of the bandwidth $h = 0.12, 0.3$, and the second column – the resampling scheme: Monte Carlo (MC) or bootstrap (B). The Monte Carlo experiment yields the corrected confidence levels $1 - \mathfrak{c}(\alpha)$, and the bootstrap yields $1 - \mathfrak{c}^{\circ}(\alpha)$. The lines 3–6 contain the average values of $1 - \mathfrak{c}(\alpha)$ and $1 - \mathfrak{c}^{\circ}(\alpha)$ over all the experiments. The results show that for the smaller bandwidth both the MC and bootstrap corrections are bigger than the ones for the larger bandwidth. In the case of a smaller bandwidth the local models have less intersections with each other, and hence, the corrections for multiplicity are closer to the Bonferroni's bound.

Remark 4.1. The theoretical results of this paper can be extended to the case when a set of considered local models has cardinality of the continuum, and the confidence bands are uniform w.r.t. the local parameter. This extension would require some uniform statements such as locally uniform square-root Wilks approximation (see e.g. Spokoiny and Zhilova (2013)).

Remark 4.2. The use of the bootstrap procedure in the problem of choosing an optimal bandwidth is considered in Spokoiny and Willrich (2015).

Table 2: **Local constant regression:**
MC vs Bootstrap confidence levels corrected for multiplicity

		Confidence levels									
h	r.m.	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.65**	**0.60**	**0.55**	**0.50**
0.12	MC	0.997	0.994	0.989	0.985	0.980	0.975	0.969	0.963	0.956	0.949
	B	0.998	0.995	0.991	0.988	0.984	0.979	0.975	0.969	0.963	0.957
0.3	MC	0.993	0.983	0.973	0.962	0.949	0.936	0.922	0.906	0.891	0.873
	B	0.994	0.986	0.977	0.968	0.958	0.947	0.935	0.922	0.908	0.893

Table 3: **Local quadratic regression:**
MC vs Bootstrap confidence levels corrected for multiplicity

		Confidence levels									
h	r.m.	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.65**	**0.60**	**0.55**	**0.50**
0.12	MC	0.997	0.993	0.989	0.985	0.979	0.974	0.968	0.961	0.954	0.946
	B	0.998	0.995	0.991	0.988	0.984	0.979	0.974	0.969	0.963	0.956
0.3	MC	0.993	0.983	0.973	0.961	0.949	0.936	0.921	0.904	0.887	0.868
	B	0.996	0.991	0.985	0.978	0.971	0.963	0.954	0.944	0.934	0.923

5 Conditions

Here we show necessary conditions for the main results. The conditions in Section 5.1 come from the general finite sample theory by Spokoiny (2012a), they are required for the results of Sections B.1 and B.2. The conditions in Section 5.2 are necessary to prove the statements on multiplier bootstrap validity.

5.1 Basic conditions

Introduce the stochastic part of the k-th likelihood process: $\zeta_k(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L_k(\boldsymbol{\theta}) - I\!\!E L_k(\boldsymbol{\theta})$, and its marginal summand: $\zeta_{i,k}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell_{i,k}(\boldsymbol{\theta}) - I\!\!E \ell_{i,k}(\boldsymbol{\theta})$ for $\ell_{i,k}(\boldsymbol{\theta})$ defined in (2.1).

(ED_0) *For each $k = 1, \ldots, K$ there exist a positive-definite $p_k \times p_k$ symmetric matrix V_k^2 and constants $\mathtt{g}_k > 0, \nu_k \geq 1$ such that* $\operatorname{Var}\{\nabla_{\boldsymbol{\theta}} \zeta_k(\boldsymbol{\theta}_k^*)\} \leq V_k^2$ *and*

$$\sup_{\boldsymbol{\gamma} \in I\!R^{p_k}} \log I\!\!E \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}} \zeta_k(\boldsymbol{\theta}_k^*)}{\|V_k \boldsymbol{\gamma}\|} \right\} \leq \nu_k^2 \lambda^2 / 2, \qquad |\lambda| \leq \mathtt{g}_k.$$

(ED_2) *For each $k = 1, \ldots, K$ there exist a constant $\omega_k > 0$ and for each $\mathtt{r} > 0$ a*

constant $\mathtt{g}_{2,k}(\mathtt{r})$ *such that it holds for all* $\boldsymbol{\theta} \in \Theta_{0,k}(\mathtt{r})$ *and for* $j = 1, 2$

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^{p_k} \\ \|\boldsymbol{\gamma}_j\| \le 1}} \log I\!E \exp\left\{ \frac{\lambda}{\omega_k} \boldsymbol{\gamma}_1^\top D_k^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta_k(\boldsymbol{\theta}) D_k^{-1} \boldsymbol{\gamma}_2 \right\} \le \nu_k^2 \lambda^2/2, \qquad |\lambda| \le \mathtt{g}_{2,k}(\mathtt{r}).$$

$(\mathcal{L}_0)$ *For each* $k = 1, \ldots, K$ *and for each* $\mathtt{r} > 0$ *there exists a constant* $\delta_k(\mathtt{r}) \ge 0$ *such that for* $\mathtt{r} \le \mathtt{r}_{0,k}$ *(*$\mathtt{r}_{0,k}$ *come from condition* (B.1) *of Theorem B.1 in Section B.1)* $\delta(\mathtt{r}) \le 1/2$ *, and for all* $\boldsymbol{\theta} \in \Theta_{0,k}(\mathtt{r})$ *it holds*

$$\| D_k^{-1} \check{D}_k^2(\boldsymbol{\theta}) D_k^{-1} - \boldsymbol{I}_{p_k} \| \le \delta_k(\mathtt{r}),$$

where $\check{D}_k^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E L_k(\boldsymbol{\theta})$ *and* $\Theta_{0,k}(\mathtt{r}) \stackrel{\text{def}}{=} \{ \boldsymbol{\theta} \in \Theta_k : \| D_k(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*) \| \le \mathtt{r} \}$ *.*

$(\mathcal{I})$ *There exist constants* $\mathfrak{a}_k > 0$ *for all* $k = 1, \ldots, K$ *s.t.*

$$\mathfrak{a}_k^2 D_k^2 \ge V_k^2.$$

Denote $\widehat{\mathfrak{a}}^2 \stackrel{\text{def}}{=} \max_{1 \le k \le K} \mathfrak{a}_k^2$ *.*

$(\mathcal{L}\mathtt{r})$ *For each* $k = 1, \ldots, K$ *and* $\mathtt{r} \ge \mathtt{r}_{0,k}$ *there exists a value* $\mathtt{b}_k(\mathtt{r}) > 0$ *s.t.* $\mathtt{r}\mathtt{b}_k(\mathtt{r}) \to \infty$ *for* $\mathtt{r} \to \infty$ *and* $\forall \boldsymbol{\theta} \in \Theta_k : \| D_k(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*) \| = \mathtt{r}$ *it holds*

$$-2 \{ I\!E L_k(\boldsymbol{\theta}) - I\!E L_k(\boldsymbol{\theta}_k^*) \} \ge \mathtt{r}^2 \mathtt{b}_k(\mathtt{r}).$$

5.2 Conditions required for the bootstrap validity

$(\widehat{\mathbf{SmB}})$ *There exists a constant* $\widehat{\delta}_{\text{smb}} \ge 0$ *such that it holds for the matrices* B_k^2 *and* H_k^2 *defined in* (3.5)*:*

$$\max_{1 \le k \le K} \left\| H_k^{-1} B_k^2 H_k^{-1} \right\| \le \widehat{\delta}^2_{\text{smb}},$$

$$\widehat{\delta}^2_{\text{smb}} \le \mathtt{C} \left(\frac{n}{p^{13}_{\max}} \right)^{1/8} \log^{-7/8}(K) \log^{-3/8}(n p_{\text{sum}}).$$

$(\boldsymbol{ED_{2m}})$ *For each* $k = 1, \ldots, K$ *,* $\mathtt{r} > 0$ *,* $i = 1, \ldots, n$ *,* $j = 1, 2$ *and for all* $\boldsymbol{\theta} \in \Theta_{0,k}(\mathtt{r})$ *it holds for the values* $\omega_k \ge 0$ *and* $\mathtt{g}_{2,k}(\mathtt{r})$ *from the condition* **$(\boldsymbol{ED_2})$** *:*

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^{p_k} \\ \|\boldsymbol{\gamma}_j\| \le 1}} \log I\!E \exp\left\{ \frac{\lambda}{\omega_k} \boldsymbol{\gamma}_1^\top D_k^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta_{i,k}(\boldsymbol{\theta}) D_k^{-1} \boldsymbol{\gamma}_2 \right\} \le \frac{\nu_0^2 \lambda^2}{2n}, \qquad |\lambda| \le \mathtt{g}_{2,k}(\mathtt{r}),$$

$(\mathcal{L}_{0m})$ *For each* $k = 1, \ldots, K$ *,* $\mathtt{r} > 0$ *,* $i = 1, \ldots, n$ *and for all* $\boldsymbol{\theta} \in \Theta_{0,k}(\mathtt{r})$ *there exists a value* $\mathtt{C}_{m,k}(\mathtt{r}) \ge 0$ *such that*

$$\| D_k^{-1} \nabla^2_{\boldsymbol{\theta}} I\!E \ell_{i,k}(\boldsymbol{\theta}) D_k^{-1} \| \le \mathtt{C}_{m,k}(\mathtt{r}) n^{-1}.$$

($\boldsymbol{\mathcal{I}_B}$) *For each* $k = 1, \ldots, K$ *there exists a constant* $\mathfrak{a}^2_{B,k} > 0$ *s.t.*

$$\mathfrak{a}^2_{B,k} D^2_k \geq B^2_k.$$

Denote $\widehat{\mathfrak{a}}^2_B \stackrel{\text{def}}{=} \max_{1 \leq k \leq K} \mathfrak{a}^2_{B,k}$.

($\widehat{\boldsymbol{SD_1}}$) *There exists a constant* $0 \leq \delta^2_{v^*} \leq \mathtt{C} p_{\text{sum}}/n$ *such that it holds for all* $i = 1, \ldots, n$ *with exponentially high probability*

$$\left\| \widehat{H}^{-1} \left\{ \boldsymbol{g}_i \boldsymbol{g}_i^\top - I\!\!E \left[\boldsymbol{g}_i \boldsymbol{g}_i^\top \right] \right\} \widehat{H}^{-1} \right\| \leq \delta^2_{v^*},$$

where

$$\begin{aligned} \boldsymbol{g}_i &\stackrel{\text{def}}{=} \left(\nabla_{\boldsymbol{\theta}} \ell_{i,1}(\boldsymbol{\theta}^*_1)^\top, \ldots, \nabla_{\boldsymbol{\theta}} \ell_{i,K}(\boldsymbol{\theta}^*_K)^\top \right)^\top \in I\!R^{p_{\text{sum}}}, \\ \widehat{H}^2 &\stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n I\!\!E \left\{ \boldsymbol{g}_i \boldsymbol{g}_i^\top \right\}, \\ p_{\text{sum}} &\stackrel{\text{def}}{=} p_1 + \cdots + p_K. \end{aligned}$$

($\boldsymbol{Eb}$) *The i.i.d. bootstrap weights* u_i *are independent of* $\boldsymbol{Y}$ *, and for all* $i = 1, \ldots, n$ *it holds for some constants* $\mathtt{g}_k > 0, \nu_k \geq 1$

$$I\!\!E u_i = 1, \quad \operatorname{Var} u_i = 1,$$

$$\log I\!\!E \exp\{\lambda(u_i - 1)\} \leq \nu_0^2 \lambda^2/2, \qquad |\lambda| \leq \mathtt{g}.$$

5.3 Dependence of the involved terms on the sample size and cardinality of the parameters' set

Here we consider the case of the i.i.d. observations $Y_1, \ldots, Y_n$ and $\mathtt{x} = \mathtt{C} \log n$ in order to specify the dependence of the non-asymptotic bounds on n and p. In the paper by Spokoiny and Zhilova (2014) (the version of 2015) this is done in detail for the i.i.d. case, generalized linear model and quantile regression.

Example 5.1 in Spokoiny (2012a) demonstrates that in this situation $\mathtt{g}_k = \mathtt{C}\sqrt{n}$ and $\omega_k = \mathtt{C}/\sqrt{n}$. then $\mathfrak{z}_k(\mathtt{x}) = \mathtt{C}\sqrt{p_k + \mathtt{x}}$ for some constant $\mathtt{C} \geq 1.85$, for the function $\mathfrak{z}_k(\mathtt{x})$ given in (B.3) in Section B.1. Similarly it can be checked that $\mathtt{g}_{2,k}(\mathtt{r})$ from condition **($\boldsymbol{ED_2}$)** is proportional to $\sqrt{n}$: due to independence of the observations

$$\begin{aligned} &\log I\!\!E \exp\left\{ \frac{\lambda}{\omega_k} \boldsymbol{\gamma}_1^\top D_k^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta_k(\boldsymbol{\theta}) D_k^{-1} \boldsymbol{\gamma}_2 \right\} \\ &\quad = \sum\nolimits_{i=1}^n \log I\!\!E \exp\left\{ \frac{\lambda}{\sqrt{n}} \frac{1}{\omega_k \sqrt{n}} \boldsymbol{\gamma}_1^\top d_k^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta_{i,k}(\boldsymbol{\theta}) d_k^{-1} \boldsymbol{\gamma}_2 \right\} \\ &\quad \leq n \frac{\lambda^2}{n} \mathtt{C} \quad \text{for } |\lambda| \leq \overline{\mathtt{g}}_{2,k}(\mathtt{r}) \sqrt{n}, \end{aligned}$$

where $\zeta_{i,k}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell_{i,k}(\boldsymbol{\theta}) - I\!\!E\ell_{i,k}(\boldsymbol{\theta})$, $d_k^2 \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!\!E\ell_{i,k}(\boldsymbol{\theta}_k^*)$ and $D_k^2 = nd_k^2$ in the i.i.d. case. Function $\overline{\mathtt{g}}_{2,k}(\mathtt{r})$ denotes the marginal analog of $\mathtt{g}_{2,k}(\mathtt{r})$.

Let us show, that for the value $\delta_k(\mathtt{r})$ from the condition $(\mathcal{L}_0)$ it holds $\delta_k(\mathtt{r}) = \mathtt{Cr}/\sqrt{n}$. Suppose for all $\boldsymbol{\theta} \in \Theta_{0,k}(\mathtt{r})$ and $\boldsymbol{\gamma} \in I\!R^{p_k} : \|\boldsymbol{\gamma}\| = 1$ $\|D_k^{-1}\boldsymbol{\gamma}^\top \nabla^3_{\boldsymbol{\theta}} I\!\!E L_k(\boldsymbol{\theta}) D_k^{-1}\| \le \mathtt{C}$, then it holds for some $\overline{\boldsymbol{\theta}} \in \Theta_{0,k}(\mathtt{r})$:

$$\begin{aligned}\|D_k^{-1}D^2(\boldsymbol{\theta})D_k^{-1} - \boldsymbol{I}_{p_k}\| &= \|D_k^{-1}(\boldsymbol{\theta}_k^* - \boldsymbol{\theta})^\top \nabla^3_{\boldsymbol{\theta}} I\!\!E L_k(\overline{\boldsymbol{\theta}}) D_k^{-1}\| \\ &= \|D_k^{-1}(\boldsymbol{\theta}_k^* - \boldsymbol{\theta})^\top D_k D_k^{-1} \nabla^3_{\boldsymbol{\theta}} I\!\!E L_k(\overline{\boldsymbol{\theta}}) D_k^{-1}\| \\ &\le \mathtt{r}\|D_k^{-1}\|\|D_k^{-1}\boldsymbol{\gamma}^\top \nabla^3_{\boldsymbol{\theta}} I\!\!E L_k(\overline{\boldsymbol{\theta}}) D_k^{-1}\| \le \mathtt{Cr}/\sqrt{n}.\end{aligned}$$

Similarly $\mathtt{C}_{m,k}(\mathtt{r}) \le \mathtt{Cr}/\sqrt{n} + \mathtt{C}$ in condition $(\mathcal{L}_{0m})$.

The next remark helps to check the global identifiability condition $(\mathcal{L}\mathtt{r})$ in many situations. Suppose that the parameter domain Θ_k is compact and n is sufficiently large, then the value $\mathtt{b}_k(\mathtt{r})$ from condition $(\mathcal{L}\mathtt{r})$ can be taken as $\mathtt{C}\{1 - \mathtt{r}/\sqrt{n}\} \approx \mathtt{C}$. Indeed, for $\boldsymbol{\theta} : \|D_k(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*)\| = \mathtt{r}$

$$\begin{aligned}-2\left\{I\!\!E L_k(\boldsymbol{\theta}) - I\!\!E L_k(\boldsymbol{\theta}_k^*)\right\} &\ge \mathtt{r}^2\left\{1 - \mathtt{r}\|D_k^{-1}\|\|D_k^{-1}\boldsymbol{\gamma}^\top \nabla^3_{\boldsymbol{\theta}} I\!\!E L_k(\overline{\boldsymbol{\theta}}) D_k^{-1}\|\right\} \\ &\ge \mathtt{r}^2(1 - \mathtt{Cr}/\sqrt{n}).\end{aligned}$$

Due to the obtained orders, the conditions (B.1) and (B.9) of Theorems B.1 and B.5 on concentration of the MLEs $\widetilde{\boldsymbol{\theta}}_k$, $\widetilde{\boldsymbol{\theta}}_k^{\circ}$ require $\mathtt{r}_{0,k} \ge \mathtt{C}\sqrt{p_k + \mathtt{x}}$.

A Approximation of the joint distributions of ℓ_2-norms

Let us previously introduce some notations:
$\mathbf{1}_K \stackrel{\text{def}}{=} (1, \ldots, 1)^\top \in I\!R^K$;
$\|\cdot\|$ is the Euclidean norm for a vector and spectral norm for a matrix;
$\|\cdot\|_{\max}$ is the maximum of absolute values of elements of a vector or of a matrix;
$\|\cdot\|_1$ is the sum of absolute values of elements of a vector or of a matrix.

Consider K random centered vectors $\boldsymbol{\phi}_k \in I\!R^{p_k}$ for $k = 1, \ldots, K$. Each vector equals to a sum of n centered independent vectors:

$$\begin{aligned}\boldsymbol{\phi}_k &= \boldsymbol{\phi}_{k,1} + \cdots + \boldsymbol{\phi}_{k,n}, \\ I\!\!E\boldsymbol{\phi}_k &= I\!\!E\boldsymbol{\phi}_{k,i} = 0 \quad \forall\, 1 \le i \le n.\end{aligned} \tag{A.1}$$

Introduce similarly the vectors $\boldsymbol{\psi}_k \in I\!R^{p_k}$ for $k = 1, \ldots, K$:

$$\begin{aligned}\boldsymbol{\psi}_k &= \boldsymbol{\psi}_{k,1} + \cdots + \boldsymbol{\psi}_{k,n}, \\ I\!\!E\boldsymbol{\psi}_k &= I\!\!E\boldsymbol{\psi}_{k,i} = 0 \quad \forall\, 1 \le i \le n,\end{aligned} \tag{A.2}$$

with the same independence properties as $\boldsymbol{\phi}_{k,i}$, and also independent of all $\boldsymbol{\phi}_{k,i}$.

The goal of this section is to compare the joint distributions of the ℓ_2-norms of the sets of vectors $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$, $k = 1, \ldots, K$ (i.e. the probability laws $\mathcal{L}(\|\boldsymbol{\phi}_1\|, \ldots, \|\boldsymbol{\phi}_K\|)$ and $\mathcal{L}(\|\boldsymbol{\psi}_1\|, \ldots, \|\boldsymbol{\psi}_K\|)$), assuming that their correlation structures are close to each other.

Denote

$$
\begin{aligned}
p_{\max} &\stackrel{\text{def}}{=} \max_{1 \le k \le K} p_k, \quad p_{\text{sum}} \stackrel{\text{def}}{=} p_1 + \cdots + p_K, \\
\lambda^2_{\phi,\max} &\stackrel{\text{def}}{=} \max_{1 \le k \le K} \| \operatorname{Var}(\boldsymbol{\phi}_j) \|, \quad \lambda^2_{\psi,\max} \stackrel{\text{def}}{=} \max_{1 \le k \le K} \| \operatorname{Var}(\boldsymbol{\psi}_j) \|, \\
z_{\max} &\stackrel{\text{def}}{=} \max_{1 \le k \le K} z_k, \quad z_{\min} \stackrel{\text{def}}{=} \min_{1 \le k \le K} z_k, \\
\delta_{z,\max} &\stackrel{\text{def}}{=} \max_{1 \le k \le K} \delta_{z_k}, \quad \delta_{z,\min} \stackrel{\text{def}}{=} \min_{1 \le k \le K} \delta_{z_k},
\end{aligned}
$$

let also

$$
\begin{aligned}
\Delta_\varepsilon \stackrel{\text{def}}{=} \left(\frac{p^3_{\max}}{n} \right)^{1/8} & \log^{9/16}(K) \log^{3/8}(n p_{\text{sum}}) z^{1/8}_{\min} \\
& \times \max\{\lambda_{\phi,\max}, \lambda_{\psi,\max}\}^{3/4} \log^{-1/8}(5n^{1/2}).
\end{aligned} \tag{A.3}
$$

The following conditions are necessary for the Proposition A.1

(C1) *For some* $\mathtt{g}_k, \nu_k, \mathfrak{c}_\phi, \mathfrak{c}_\psi > 0$ *and for all* $i = 1, \ldots, n$, $k = 1, \ldots, K$

$$
\sup_{\substack{\boldsymbol{\gamma}_k \in \mathbb{R}^{p_k}, \\ \|\boldsymbol{\gamma}_k\| = 1}} \log \mathbb{E} \exp\left\{ \lambda \sqrt{n} \boldsymbol{\gamma}_k^\top \boldsymbol{\phi}_{k,i} / \mathfrak{c}_\phi \right\} \le \lambda^2 \nu_k^2 / 2, \quad |\lambda| < \mathtt{g}_k,
$$

$$
\sup_{\substack{\boldsymbol{\gamma}_k \in \mathbb{R}^{p_k}, \\ \|\boldsymbol{\gamma}_k\| = 1}} \log \mathbb{E} \exp\left\{ \lambda \sqrt{n} \boldsymbol{\gamma}_k^\top \boldsymbol{\psi}_{k,i} / \mathfrak{c}_\psi \right\} \le \lambda^2 \nu_k^2 / 2, \quad |\lambda| < \mathtt{g}_k,
$$

where $\mathfrak{c}_\phi \ge \mathtt{C} \lambda_{\phi,\max}$ *and* $\mathfrak{c}_\psi \ge \mathtt{C} \lambda_{\phi,\max}$.

(C2) *For some* $\delta^2_\Sigma \ge 0$

$$
\max_{1 \le k_1, k_2 \le K} \left\| \operatorname{Cov}(\boldsymbol{\phi}_{k_1}, \boldsymbol{\phi}_{k_2}) - \operatorname{Cov}(\boldsymbol{\psi}_{k_1}, \boldsymbol{\psi}_{k_2}) \right\|_{\max} \le \delta^2_\Sigma. \tag{A.4}
$$

Proposition A.1 (Approximation of the joint distributions of ℓ_2-norms)**.** *Consider the centered random vectors* $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K$ *and* $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K$ *given in* (A.1), (A.2). *Let the conditions (C1) and (C2) be fulfilled, and the values* $z_k \ge \sqrt{p_k + \Delta_\varepsilon}$ *and* $\delta_{z_k} \ge 0$ *be s.t.*

$\mathtt{C}\max\{n^{-1/2}, \delta_{z,\max}\} \leq \Delta_\varepsilon \leq \mathtt{C} z_{\max}^{-1}$, *then it holds with dominating probability*

$$\mathbb{P}\left(\bigcup\nolimits_{k=1}^{K}\{\|\boldsymbol{\phi}_k\| > z_k\}\right) - \mathbb{P}\left(\bigcup\nolimits_{k=1}^{K}\{\|\boldsymbol{\psi}_k\| > z_k - \delta_{z_k}\}\right) \geq -\Delta_{\ell_2},$$
$$\mathbb{P}\left(\bigcup\nolimits_{k=1}^{K}\{\|\boldsymbol{\phi}_k\| > z_k\}\right) - \mathbb{P}\left(\bigcup\nolimits_{k=1}^{K}\{\|\boldsymbol{\psi}_k\| > z_k + \delta_{z_k}\}\right) \leq \Delta_{\ell_2}$$

for the deterministic non-negative value

$$\begin{aligned}\Delta_{\ell_2} \leq\ & 12.5\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8} \log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\max\left\{\lambda_{\phi,\max}, \lambda_{\psi,\max}\right\}^{3/4} \\ & + 3.2\mathtt{C}\delta_\Sigma^2\left(\frac{p_{\max}^3}{n}\right)^{1/4} p_{\max} z_{\min}^{1/2}\log^2(K)\log^{3/4}(np_{\text{sum}})\max\left\{\lambda_{\phi,\max}, \lambda_{\psi,\max}\right\}^{7/2} \\ \leq\ & 25\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8} \log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\max\left\{\lambda_{\phi,\max}, \lambda_{\psi,\max}\right\}^{3/4},\end{aligned}$$

where the last inequality holds for

$$\delta_\Sigma^2 \leq 4\mathtt{C}\left(\frac{n}{p_{\max}^{13}}\right)^{1/8} \log^{-7/8}(K)\log^{-3/8}(np_{\text{sum}})\left(\max\left\{\lambda_{\phi,\max}, \lambda_{\psi,\max}\right\}\right)^{-11/4}.$$

Remark A.1. The approximating error term Δ_{ℓ_2} consists of three errors, which correspond to: the Gaussian approximation result (Lemma A.2), Gaussian comparison (Lemma A.7), and anti-concentration inequality (Lemma A.8). The bound on Δ_{ℓ_2} above implies that the number K of the random vectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K$ should satisfy $\log K \ll (n/p_{\max}^3)^{1/12}$ in order to keep the approximating error term Δ_{ℓ_2} small. This condition can be relaxed by using a sharper Gaussian approximation result. For instance, using in Lemma A.2 the Slepian-Stein technique plus induction argument from the recent paper by Chernozhukov et al. (2014b) instead of the Lindeberg's approach, would lead to the improved bound: $\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/6}$ multiplied by a logarithmic term.

A.1 Joint Gaussian approximation of ℓ_2-norm of sums of independent vectors by Lindeberg's method

Introduce the following random vectors from $\mathbb{R}^{p_{\text{sum}}}$:

$$\begin{aligned}&\Phi \stackrel{\text{def}}{=} \left(\boldsymbol{\phi}_1^\top, \ldots, \boldsymbol{\phi}_K^\top\right)^\top, \quad \Phi_i \stackrel{\text{def}}{=} \left(\boldsymbol{\phi}_{1,i}^\top, \ldots, \boldsymbol{\phi}_{K,i}^\top\right)^\top,\ i = 1, \ldots, n, \\ &\Phi = \sum\nolimits_{i=1}^{n}\Phi_i, \qquad\qquad \mathbb{E}\Phi = \mathbb{E}\Phi_i = 0.\end{aligned} \tag{A.5}$$

Define their Gaussian analogs as follows:

$$\overline{\Phi}_i \stackrel{\text{def}}{=} \left(\overline{\phi}_{1,i}^\top, \ldots, \overline{\phi}_{K,i}^\top\right)^\top, \qquad \overline{\Phi} \stackrel{\text{def}}{=} \left(\overline{\phi}_1^\top, \ldots, \overline{\phi}_K^\top\right)^\top = \sum\nolimits_{i=1}^n \overline{\Phi}_i, \tag{A.6}$$

$$\overline{\Phi}_i \sim \mathcal{N}(0, \operatorname{Var}\Phi_i), \qquad \overline{\Phi} \sim \mathcal{N}(0, \operatorname{Var}\Phi), \tag{A.7}$$

$$\overline{\phi}_{k,i} \sim \mathcal{N}(0, \operatorname{Var}\phi_{k,i}), \qquad \overline{\phi}_k \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \overline{\phi}_{k,i} \sim \mathcal{N}(0, \operatorname{Var}\phi_k). \tag{A.8}$$

Lemma A.2 (Joint GAR with equal covariance matrices)**.** *Consider the sets of random vectors* ϕ_j *and* $\overline{\phi}_j$, $j = 1, \ldots, K$ *defined in* (A.1)*, and* (A.5)– (A.8)*. If the conditions of Lemmas A.4 are A.5 are fulfilled, then it holds for all* $\Delta, \beta > 0$, $z_j \geq \max\left\{\Delta + \sqrt{p_j}, 2.25\log(K)/\beta\right\}$ *with dominating probability*

$$I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\phi_j\| > z_j\right\}\right) \leq I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\overline{\phi}_j\| > z_j - \Delta - \frac{3\log(K)}{2\beta}\right\}\right) + \delta_{3,\phi}(\Delta, \beta),$$

$$I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\phi_j\| > z_j\right\}\right) \geq I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\overline{\phi}_j\| > z_j + \Delta + \frac{3\log(K)}{2\beta}\right\}\right) - \delta_{3,\phi}(\Delta, \beta)$$

for $\delta_{3,\phi}(\Delta, \beta) \leq \mathtt{C}\left(\frac{1}{\Delta^3} + \frac{\beta}{\Delta^2} + \frac{\beta^2}{\Delta}\right)\left\{\frac{p_{\max}^3}{n}\log(K)\log^3(np_{\text{sum}})\right\}^{1/2}$ *given in* (A.15)*.*

Proof of Lemma A.2.

$$I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\phi_j\| > z_j\right\}\right) = I\!E\, 1\!\!1\big(\max\nolimits_{1\leq j\leq K}\left\{\|\phi_j\|^2 - z_j^2\right\} > 0\big).$$

Let us approximate the $\max_{1\leq j\leq K}$ function using the smooth maximum:

$$h_\beta\left(\{x_j\}\right) \stackrel{\text{def}}{=} \beta^{-1}\log\left(\sum\nolimits_{j=1}^K \mathrm{e}^{\beta x_j}\right) \text{ for } \beta > 0,\, x_j \in I\!R,$$

$$h_\beta\left(\{x_j\}\right) - \beta^{-1}\log(K) \leq \max_{1\leq j\leq K}\{x_j\} \leq h_\beta\left(\{x_j\}\right). \tag{A.9}$$

The indicator function $1\!\!1\{x > 0\}$ is approximated with the three times differentiable function $g(x)$ growing monotonously from 0 to 1:

$$g(x) \stackrel{\text{def}}{=} \begin{cases} 0, & x \leq 0, \\ 16x^3/3, & x \in [0, 1/4], \\ 0.5 + 2(x - 0.5) - 16(x - 0.5)^3/3, & x \in [1/4, 3/4], \\ 1 + 16(x-1)^3/3, & x \in [3/4, 1], \\ 1, & x \geq 1. \end{cases}$$

It holds for all $x \in I\!R$ and $\Delta > 0$

$$1\!\!1\left\{x > \Delta\right\} \leq g(x/\Delta) \leq 1\!\!1\left\{x/\Delta > 0\right\}.$$

Therefore

$$
\begin{aligned}
&I\!P\left(\max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\| - z_j\right\} > \Delta\right) \\
&\quad \le I\!E\, 1\!\!1\left(\max_{1\le j\le K}\left\{\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j}\right\} > \Delta\right) \\
&\quad \le I\!E g\left(\max_{1\le j\le K}\left\{\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j\Delta}\right\}\right) \\
&\quad \le I\!E g\left(\frac{1}{\Delta\beta}\log\left\{\sum\nolimits_{j=1}^{K}\exp\left[\beta\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j}\right]\right\}\right) \qquad \text{(A.10)} \\
&\quad \le I\!E g\left(\max_{1\le j\le K}\left\{\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j\Delta}\right\} + \frac{\log(K)}{\beta\Delta}\right) \\
&\quad \le I\!E\, 1\!\!1\left(\max_{1\le j\le K}\left\{\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j}\right\} > -\frac{\log(K)}{\beta}\right) \\
&\quad \le I\!P\left(\max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\| - z_j\right\} > -1.5\frac{\log(K)}{\beta}\right), \qquad \text{(A.11)}
\end{aligned}
$$

where the last inequality holds for $z_j \ge 2.25\log(K)/\beta$. Denote

$$
\boldsymbol{z} \stackrel{\text{def}}{=} (z_1, \ldots, z_K)^\top \in I\!R^K,\ z_j > 0.
$$

Introduce the function $F_{\Delta,\beta}(\Phi, \boldsymbol{z}) : I\!R^{p_{\text{sum}}} \times I\!R^K \mapsto I\!R$:

$$
F_{\Delta,\beta}(\Phi, \boldsymbol{z}) \stackrel{\text{def}}{=} g\left(\frac{1}{\Delta\beta}\log\left\{\sum\nolimits_{j=1}^{K}\exp\left[\beta\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j}\right]\right\}\right) \qquad \text{(A.12)}
$$

Then by (A.10) and (A.11)

$$
\begin{aligned}
&I\!P\left(\max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\| - z_j\right\} > \Delta\right) \\
&\qquad \le I\!E F_{\Delta,\beta}(\Phi, \boldsymbol{z}) \qquad \text{(A.13)} \\
&\qquad \le I\!P\left(\max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\| - z_j\right\} > -\frac{3\log(K)}{2\beta}\right). \qquad \text{(A.14)}
\end{aligned}
$$

Lemma A.6 checks that $F_{\Delta,\beta}(\cdot, \boldsymbol{z})$ admits applying the Lindeberg's telescopic sum device (see Lindeberg (1922)) in order to approximate $I\!E F_{\Delta,\beta}(\Phi, \boldsymbol{z})$ with $I\!E F_{\Delta,\beta}(\overline{\Phi}, \boldsymbol{z})$. Define for $q = 2, \ldots, n-1$ the following $I\!R^{p_{\text{sum}}}$-valued random sums:

$$
S_q \stackrel{\text{def}}{=} \sum_{i=1}^{q-1}\overline{\Phi}_i + \sum_{i=q+1}^{n}\Phi_i, \qquad S_1 \stackrel{\text{def}}{=} \sum_{i=2}^{n}\Phi_i, \qquad S_n \stackrel{\text{def}}{=} \sum_{i=1}^{n-1}\overline{\Phi}_i.
$$

The difference $F_{\Delta,\beta}(\Phi,\boldsymbol{z}) - F_{\Delta,\beta}(\overline{\Phi},\boldsymbol{z})$ can be represented as the telescopic sum:

$$F_{\Delta,\beta}(\Phi,\boldsymbol{z}) - F_{\Delta,\beta}(\overline{\Phi},\boldsymbol{z}) = \sum\nolimits_{i=1}^{n} \{F_{\Delta,\beta}(S_i+\Phi_i,\boldsymbol{z}) - F_{\Delta,\beta}(S_i+\overline{\Phi}_i,\boldsymbol{z})\}.$$

The third order Taylor expansions of $F_{\Delta,\beta}(S_i+\Phi_i,\boldsymbol{z})$ and $F_{\Delta,\beta}(S_i+\overline{\Phi}_i,\boldsymbol{z})$ w.r.t. the first argument at S_i, and Lemma A.6 imply for each $i=1,\ldots,n$:

$$\begin{aligned}
&\Big| F_{\Delta,\beta}(S_i+\Phi_i,\boldsymbol{z}) - F_{\Delta,\beta}(S_i+\overline{\Phi}_i,\boldsymbol{z}) - \nabla_{\Phi} F_{\Delta,\beta}(S_i,\boldsymbol{z})^{\top}(\Phi_i-\overline{\Phi}_i) \\
&\quad - \frac{1}{2}(\Phi_i-\overline{\Phi}_i)^{\top}\nabla^2_{\Phi} F_{\Delta,\beta}(S_i,\boldsymbol{z})(\Phi_i+\overline{\Phi}_i)\Big| \\
&\leq \frac{\mathtt{C}_3(\Delta,\beta)}{6}\left(\max_{1\leq j\leq K} \left\{ \|S_{j,i}+\boldsymbol{\phi}_{j,i}\|^3 \right\} \|\Phi_i\|^3_{\max} + \max_{1\leq j\leq K} \left\{ \|S_{j,i}+\overline{\boldsymbol{\phi}}_{j,i}\|^3 \right\} \|\overline{\Phi}_i\|^3_{\max} \right),
\end{aligned}$$

where the value $\mathtt{C}_3(\Delta,\beta)$ is defined in Lemma A.6, and the random vectors $S_{j,i} \in I\!R^{p_j}$ for $j=1,\ldots,K$ are s.t. for all $i=1,\ldots,n$

$$S_i = \left(S_{1,i}^{\top}, S_{2,i}^{\top}, \ldots, S_{K,i}^{\top} \right)^{\top}.$$

By their construction S_i and $\Phi_i - \overline{\Phi}_i$ are independent, $I\!E\Phi_i = I\!E\overline{\Phi}_i = 0$ and $\operatorname{Var}\Phi_i = \operatorname{Var}\overline{\Phi}_i$, therefore

$$\begin{aligned}
&\left| I\!E F_{\Delta,\beta}(\Phi,\boldsymbol{z}) - I\!E F_{\Delta,\beta}(\overline{\Phi},\boldsymbol{z}) \right| \\
&\quad = \left| \sum\nolimits_{i=1}^{n} \{ I\!E H_{\Delta}(S_i+\Phi_i,\boldsymbol{z}) - I\!E H_{\Delta}(S_i+\overline{\Phi}_i,\boldsymbol{z}) \} \right| \\
&\quad \leq \frac{\mathtt{C}_3(\Delta,\beta)}{6} \sum_{i=1}^{n} I\!E \left(\max_{1\leq j\leq K} \left\{ \|S_{j,i}+\boldsymbol{\phi}_{j,i}\|^3 \right\} \|\Phi_i\|^3_{\max} + \max_{1\leq j\leq K} \left\{ \|S_{j,i}+\overline{\boldsymbol{\phi}}_{j,i}\|^3 \right\} \|\overline{\Phi}_i\|^3_{\max} \right).
\end{aligned}$$

Lemma A.5 implies for all $i=1,\ldots,n$ with probability $\geq 1-2\mathrm{e}^{-\mathtt{x}}$

$$\left(I\!E \max_{1\leq j\leq K} \left\{ \|S_{j,i}+\boldsymbol{\phi}_{j,i}\|^6 \right\} \right)^{1/2} \leq \mathtt{C}\nu_0 \max_{1\leq j\leq K} \| \operatorname{Var}^{1/2}(\boldsymbol{\phi}_j) \|^3 \sqrt{p_{\max}\log(K)}(p_{\max}+6\mathtt{x}),$$

and the same bound holds for $\left(I\!E \max_{1\leq j\leq K} \left\{ \|S_{j,i}+\overline{\boldsymbol{\phi}}_{j,i}\|^6 \right\} \right)^{1/2}$. Denote

$$\delta_{\max,\phi} \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^{n} \left\{ I\!E \left(\|\Phi_i\|^6_{\max} \right) \right\}^{1/2} + \left\{ I\!E \left(\|\overline{\Phi}_i\|^6_{\max} \right) \right\}^{1/2}.$$

By Lemma A.4 it holds for $t = (\mathtt{x}+\log(p_{\text{sum}}))^3 \left(\sqrt{2}\mathfrak{c}_{\phi}\nu_0\right)^6 n^{-3}$ with probability $\geq 1-\mathrm{e}^{-\mathtt{x}}$

$$\|\Phi_i\|^6_{\max} \leq t, \quad \|\overline{\Phi}_i\|^6_{\max} \leq t.$$

If $\mathtt{x} = \mathtt{C}\log n$, then the last bound on $\left| I\!E F_{\Delta,\beta}(\Phi,\boldsymbol{z}) - I\!E F_{\Delta,\beta}(\overline{\Phi},\boldsymbol{z}) \right|$ continues with

probability $\geq 1 - 6\exp(-\mathtt{x})$ as follows

$$\begin{aligned}
&\left|I\!\!E F_{\Delta,\beta}(\Phi, \boldsymbol{z}) - I\!\!E F_{\Delta,\beta}(\overline{\Phi}, \boldsymbol{z})\right| \\
&\quad \leq \mathtt{C}\frac{\mathtt{C}_3(\Delta,\beta)}{3}\sqrt{p_{\max}^3 \log(K)}\delta_{\max,\phi} \max_{1\leq j\leq K} \|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|^3 \\
&\quad \leq \frac{\mathtt{C}}{3}\left(\frac{1}{\Delta^3} + \frac{\beta}{\Delta^2} + \frac{\beta^2}{\Delta}\right)\frac{p_{\max}^{3/2}}{n^{1/2}} \log^{1/2}(K) \log^{3/2}(np_{\mathrm{sum}}) \max_{1\leq j\leq K} \|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|^3 \left(2\nu_0^2\mathfrak{c}_\phi^2\right)^{3/2} \\
&\quad \stackrel{\mathrm{def}}{=} \delta_{3,\phi}(\Delta,\beta). \qquad \text{(A.15)}
\end{aligned}$$

The derived bounds imply:

$$\begin{aligned}
&I\!\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) \\
&\quad \stackrel{\text{by (A.13)}}{\leq} I\!\!E F_{\Delta,\beta}\left(\Phi, \boldsymbol{z} - \Delta\mathbf{1}_K\right) \\
&\quad \stackrel{\text{by (A.15)}}{\leq} I\!\!E F_{\Delta,\beta}\left(\overline{\Phi}, \boldsymbol{z} - \Delta\mathbf{1}_K\right) + \delta_{3,\phi}(\Delta,\beta) \qquad \text{(A.16)} \\
&\quad \stackrel{\text{by (A.14)}}{\leq} I\!\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\overline{\boldsymbol{\phi}}_j\| > z_j - \Delta - \frac{3\log(K)}{2\beta}\right\}\right) + \delta_{3,\phi}(\Delta,\beta),
\end{aligned}$$

and similarly

$$\begin{aligned}
&I\!\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) \\
&\quad \geq I\!\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\overline{\boldsymbol{\phi}}_j\| > z_j + \frac{3\log(K)}{2\beta} + \Delta\right\}\right) - \delta_{3,\phi}(\Delta,\beta).
\end{aligned}$$

□

The next lemma is formulated separately, since it is used for a proof of another result.

Lemma A.3 (Smooth uniform GAR)**.** *Under the conditions of Lemma A.2 it holds with dominating probability for the function* $F_{\Delta,\beta}(\cdot, \boldsymbol{z})$ *given in* (A.12)*:*

1.1. $I\!\!P\left(\bigcup_{j=1}^K \left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) \leq I\!\!E F_{\Delta,\beta}\left(\overline{\Phi}, \boldsymbol{z} - \Delta\mathbf{1}_K\right) + \delta_{3,\phi}(\Delta,\beta)$,

1.2. $I\!\!P\left(\bigcup_{j=1}^K \left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) \geq I\!\!E H_{\Delta,\beta}\left(\overline{\Phi}, \boldsymbol{z} + \frac{3\log(K)}{2\beta}\mathbf{1}_K\right) - \delta_{3,\phi}(\Delta,\beta)$;

2.1. $I\!\!E F_{\Delta,\beta}(\Phi, \boldsymbol{z}) \leq I\!\!P\left(\bigcup_{j=1}^K \left\{\|\boldsymbol{\phi}_j\| > z_j - \frac{3\log(K)}{2\beta}\right\}\right)$,

2.2. $I\!\!E F_{\Delta,\beta}(\Phi, \boldsymbol{z}) \geq I\!\!P\left(\bigcup_{j=1}^K \left\{\|\boldsymbol{\phi}_j\| > z_j + \Delta\right\}\right)$.

Proof of Lemma A.3. The first inequality 1.1 is obtained in (A.16), the second inequality 1.2 follows similarly from (A.14) and (A.15). The inequalities 2.1 and 2.2 are given in (A.13) and (A.14). □

Lemma A.4. *Let for some* $\mathfrak{c}_\phi, \mathtt{g}_1, \nu_0 > 0$ *and for all* $i = 1, \ldots, n\,,\ j = 1, \ldots, p_{\mathrm{sum}}$

$$\log I\!\!E \exp\left\{\lambda\sqrt{n}|\phi_i^j|/\mathfrak{c}_\phi\right\} \le \lambda^2\nu_0^2/2, \quad |\lambda| < \mathtt{g}_1,$$

here ϕ_i^j *denotes the* j*-th coordinate of vector* $\boldsymbol{\phi}_i$*. Then it holds for all* $i = 1, \ldots, n$ *and* $m, t > 0$

$$I\!\!P\left(\max_{1\le j\le p_{\mathrm{sum}}} |\phi_i^j|^m > t\right) \le \exp\left\{-\frac{nt^{2/m}}{2\mathfrak{c}_\phi^2\nu_0^2} + \log(p_{\mathrm{sum}})\right\}.$$

Proof of Lemma A.4. Let us bound the $\max_j |\phi_i^j|$ using the following bound for the maximum:

$$\max_{1\le j\le p_{\mathrm{sum}}} |\phi_i^j| \le \log\left\{\sum\nolimits_{j=1}^{p_{\mathrm{sum}}} \exp\bigl(|\phi_i^j|\bigr)\right\}.$$

By the Lemma's condition

$$I\!\!E \exp\left\{\max_{1\le j\le p} \frac{\lambda\sqrt{n}}{\mathfrak{c}_\phi}|\phi_i^j|\right\} \le \exp\left(\lambda^2\nu_0^2/2 + \log p_{\mathrm{sum}}\right).$$

Thus, the statement follows from the exponential Chebyshev's inequality. □

Lemma A.5. *If for the centered random vectors* $\boldsymbol{\phi}_j \in I\!R^{p_j}\quad j = 1, \ldots, K$

$$\sup_{\substack{\boldsymbol{\gamma}\in I\!R^{p_j},\\ \|\boldsymbol{\gamma}\|\neq 0}} \log I\!\!E \exp\left\{\lambda\frac{\boldsymbol{\gamma}^\top\boldsymbol{\phi}_j}{\|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\boldsymbol{\gamma}\|}\right\} \le \nu_0^2\lambda^2/2, \qquad |\lambda| \le \mathtt{g}$$

for some constants $\nu_0 > 0$ *and* $\mathtt{g} \ge \nu_0^{-1}\max_{1\le j\le K}\sqrt{2p_j\log(K)}$*, then*

$$I\!\!E \max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\|\right\} \le \mathtt{C}\nu_0 \max_{1\le j\le K}\|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|\sqrt{2p_{\max}\log(K)},$$

$$\left(I\!\!E \max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\|^6\right\}\right)^{1/2} \le \mathtt{C}\nu_0 \max_{1\le j\le K}\|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|^3\sqrt{2p_{\max}\log(K)}(p_{\max} + 6\mathtt{x}),$$

The second bound holds with probability $\ge 1 - 2\mathrm{e}^{-\mathtt{x}}$*.*

Proof of Lemma A.5. Let us take for each $j = 1, \ldots, K$ finite ε_j-grids $\mathrm{G}_j(\varepsilon) \subset I\!R^{p_j}$ on the $(p_j - 1)$-spheres of radius 1 s.t

$$\forall\boldsymbol{\gamma} \in I\!R^{p_j} \text{ s.t. } \|\boldsymbol{\gamma}\| = 1 \ \ \exists\boldsymbol{\gamma}_0 \in \mathrm{G}_j(\varepsilon): \ \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \le \varepsilon,\ \|\boldsymbol{\gamma}_0\| = 1.$$

Then

$$\|\boldsymbol{\phi}_j\| \leq (1-\varepsilon_j)^{-1} \max_{\boldsymbol{\gamma} \in \mathrm{G}_j(\varepsilon_j)} \left\{ \boldsymbol{\gamma}^\top \boldsymbol{\phi}_j \right\}.$$

Hence, by inequality (A.9) and the imposed condition it holds for all $0 < \mu < \mathtt{g} / \max_{1 \leq j \leq K} \| \operatorname{Var}^{1/2}(\boldsymbol{\phi}_j) \|$:

$$\begin{aligned}
I\!\!E \max_{1 \leq j \leq K} \left\{ \|\boldsymbol{\phi}_j\| \right\} &\leq \max_{1 \leq j \leq K} \frac{1}{1-\varepsilon_j} I\!\!E \max_{1 \leq j \leq K} \max_{\boldsymbol{\gamma} \in \mathrm{G}_j(\varepsilon_j)} \left\{ \boldsymbol{\gamma}^\top \boldsymbol{\phi}_j \right\} \\
&\leq \mathtt{C} \frac{1}{\mu} I\!\!E \log \left\{ \sum_{1 \leq j \leq K} \sum_{\boldsymbol{\gamma} \in \mathrm{G}_j(\varepsilon_j)} \exp\left(\mu \boldsymbol{\gamma}^\top \boldsymbol{\phi}_j \right) \right\} \\
&\leq \mathtt{C} \frac{1}{\mu} \log \left\{ \sum_{1 \leq j \leq K} \sum_{\boldsymbol{\gamma} \in \mathrm{G}_j(\varepsilon_j)} I\!\!E \exp\left(\mu \boldsymbol{\gamma}^\top \boldsymbol{\phi}_j \right) \right\} \\
&\leq \mathtt{C} \max_{1 \leq j \leq K} \frac{\log(K \operatorname{card}\{\mathrm{G}_j(\varepsilon_j)\})}{\mu} + \mathtt{C} \frac{\mu \nu_0^2}{2} \max_{1 \leq j \leq K} \| \operatorname{Var}(\boldsymbol{\phi}_j) \| \\
&\leq \mathtt{C} \max_{1 \leq j \leq K} \{p_j\} \frac{\log(K)}{\mu} + \mathtt{C} \frac{\mu \nu_0^2}{2} \max_{1 \leq j \leq K} \| \operatorname{Var}(\boldsymbol{\phi}_j) \| \\
&= \mathtt{C} \nu_0 \max_{1 \leq j \leq K} \{\sqrt{p_j}\} \max_{1 \leq j \leq K} \| \operatorname{Var}^{1/2}(\boldsymbol{\phi}_j) \| \sqrt{2 \log(K)} \\
&\quad \text{for} \quad \mu = \mathtt{C} \nu_0^{-1} \max_{1 \leq j \leq K} \{\sqrt{p_j}\} \sqrt{2 \log(K)} / \max_{1 \leq j \leq K} \| \operatorname{Var}^{1/2}(\boldsymbol{\phi}_j) \|.
\end{aligned}$$

For the second part of the statement we combine the first part with the result of Theorem B.3 on deviation of a random quadratic form: it holds with dominating probability for $V_{\phi_j}^2 \stackrel{\text{def}}{=} \operatorname{Var} \boldsymbol{\phi}_j$

$$\begin{aligned}
\|\boldsymbol{\phi}_j\|^2 &\leq \mathfrak{Z}_{\mathtt{qf}}^2(\mathtt{x}, V_{\phi_j}) \\
&\leq \operatorname{tr}(V_{\phi_j}^2) + 6\mathtt{x} \|V_{\phi_j}^2\| \leq \|V_{\phi_j}^2\| (p_j + 6\mathtt{x}).
\end{aligned}$$

□

Lemma A.6. *Let $\Gamma \in I\!R^{p_{\mathrm{sum}}}$, $\boldsymbol{\gamma}_j \in I\!R^{p_j}$ for $j = 1, \ldots, K$ are s.t. $\Gamma = \left(\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_K^\top \right)^\top$, and $\boldsymbol{z} \stackrel{\text{def}}{=} (z_1, \ldots, z_K)^\top$ s.t. $z_j \geq \sqrt{p_j}$, then it holds for the function $F_{\Delta, \beta}(\cdot, \boldsymbol{z})$ defined in (A.12):*

$$\left\| \nabla_\Gamma^2 F_{\Delta, \beta}(\Gamma, \boldsymbol{z}) \right\|_1 \leq \mathtt{C}_2(\Delta, \beta) \max_{1 \leq j \leq K} \left\{ \|\boldsymbol{\gamma}_j\|^2 \right\}, \quad \mathtt{C}_2(\Delta, \beta) \stackrel{\text{def}}{=} \mathtt{C} \left(\frac{1}{\Delta^2} + \frac{\beta}{\Delta} \right),$$

$$\left\| \nabla_\Gamma^3 F_{\Delta, \beta}(\Gamma, \boldsymbol{z}) \right\|_1 \leq \mathtt{C}_3(\Delta, \beta) \max_{1 \leq j \leq K} \left\{ \|\boldsymbol{\gamma}_j\|^3 \right\}, \quad \mathtt{C}_3(\Delta, \beta) \stackrel{\text{def}}{=} \mathtt{C} \left(\frac{1}{\Delta^3} + \frac{\beta}{\Delta^2} + \frac{\beta^2}{\Delta} \right).$$

Proof of Lemma A.6. Denote

$$s(\Gamma) \stackrel{\text{def}}{=} \sum\nolimits_{j=1}^{K} \exp\left(\beta \frac{\|\boldsymbol{\gamma}_j\|^2 - z_j^2}{2z_j}\right), \quad h_\beta(s(\Gamma)) \stackrel{\text{def}}{=} \beta^{-1} \log\{s(\Gamma)\}, \tag{A.17}$$

then $F_{\beta,\Delta}(\Gamma, \boldsymbol{z}) = g\left(\Delta^{-1} h_\beta\left(s(\Gamma)\right)\right)$. Let γ^q denote the q-th coordinate of the vector $\Gamma \in I\!R^{p_{\text{sum}}}$. It holds for $q, l, b, r = 1, \ldots, p_{\text{sum}}$:

$$\begin{aligned}
\frac{d}{d\gamma^q} F_{\beta,\Delta}(\Gamma, \boldsymbol{z}) &= \frac{1}{\Delta} g'\left\{\Delta^{-1} h_\beta(s(\Gamma))\right\} \frac{d}{d\gamma^q} h_\beta(s(\Gamma)), \\
\frac{d^2}{d\gamma^q d\gamma^l} F_{\beta,\Delta}(\Gamma, \boldsymbol{z}) &= \frac{1}{\Delta^2} g''\left\{\Delta^{-1} h_\beta(s(\Gamma))\right\} \frac{d}{d\gamma^q} h_\beta(s(\Gamma)) \frac{d}{d\gamma^l} h_\beta(s(\Gamma)) \\
&\quad + \frac{1}{\Delta} g'\left\{\Delta^{-1} h_\beta(s(\Gamma))\right\} \frac{d^2}{d\gamma^q d\gamma^l} h_\beta(s(\Gamma)), \\
\frac{d^3}{d\gamma^q d\gamma^l d\gamma^b} F_{\beta,\Delta}(\Gamma, \boldsymbol{z}) &= \frac{1}{\Delta^3} g'''\left\{\Delta^{-1} h_\beta(s(\Gamma))\right\} \frac{d}{d\gamma^q} h_\beta(s(\Gamma)) \frac{d}{d\gamma^l} h_\beta(s(\Gamma)) \frac{d}{d\gamma^b} h_\beta(s(\Gamma)) \\
&\quad + \frac{1}{\Delta^2} g''\left\{\Delta^{-1} h_\beta(s(\Gamma))\right\} \left\{ \frac{d^2}{d\gamma^q d\gamma^b} h_\beta(s(\Gamma)) \frac{d}{d\gamma^l} h_\beta(s(\Gamma)) \right. \\
&\quad \left. + \frac{d}{d\gamma^q} h_\beta(s(\Gamma)) \frac{d^2}{d\gamma^l d\gamma^b} h_\beta(s(\Gamma)) + \frac{d}{d\gamma^b} h_\beta(s(\Gamma)) \frac{d^2}{d\gamma^q d\gamma^l} h_\beta(s(\Gamma)) \right\} \\
&\quad + \frac{1}{\Delta} g'\left\{\Delta^{-1} h_\beta(s(\Gamma))\right\} \frac{d^3}{d\gamma^q d\gamma^l d\gamma^b} h_\beta(s(\Gamma)).
\end{aligned}$$

Let for $1 \le q \le p_{\text{sum}}$ $j(q)$ denote an index from 1 to K s.t. the coordinate γ^q of the vector $\Gamma = \left(\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_K^\top\right)^\top$ belongs to its sub-vector $\boldsymbol{\gamma}_{j(q)}$.

$$\frac{d}{d\gamma^q} h_\beta(s(\Gamma)) = \frac{1}{\beta} \frac{1}{s(\Gamma)} \frac{d}{d\gamma^q} s(\Gamma) = \frac{1}{s(\Gamma)} \frac{\gamma^q}{z_{j(q)}} \exp\left(\beta \frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2z_{j(q)}}\right),$$

$$\frac{d^2}{d\gamma^q d\gamma^l} h_\beta(s(\Gamma)) = \frac{1}{\beta}\frac{1}{s(\Gamma)}\frac{d^2}{d\gamma^q d\gamma^l}s(\Gamma) - \frac{1}{\beta}\frac{1}{s^2(\Gamma)}\frac{d}{d\gamma^q}s(\Gamma)\frac{d}{d\gamma^l}s(\Gamma)$$

$$= \begin{cases} \left\{\frac{1}{z_{j(q)}} + \beta\left(\frac{\gamma^q}{z_{j(q)}}\right)^2\right\}\frac{1}{s(\Gamma)}\exp\left(\beta\frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2z_{j(q)}}\right) \\ \qquad -\frac{\beta}{s^2(\Gamma)}\left\{\frac{\gamma^q}{z_{j(q)}}\right\}^2\exp\left(2\beta\frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2z_{j(q)}}\right), & q = l; \\ \frac{\beta}{s(\Gamma)}\frac{\gamma^q\gamma^l}{z_{j(q)}^2}\exp\left(\beta\frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2z_{j(q)}}\right) \\ \qquad -\frac{\beta}{s^2(\Gamma)}\frac{\gamma^q\gamma^l}{z_{j(q)}^2}\exp\left(2\beta\frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2z_{j(q)}}\right), & j(q) = j(l), q \neq l; \\ -\frac{\beta}{s^2(\Gamma)}\frac{\gamma^q\gamma^l}{z_{j(q)}z_{j(l)}}\exp\left(\beta\frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2z_{j(q)}} + \beta\frac{\|\boldsymbol{\gamma}_{j(l)}\|^2 - z_{j(l)}^2}{2z_{j(l)}}\right), & j(q) \neq j(l). \end{cases}$$

By definition (A.17) of $s(\Gamma)$ it holds for all $\Gamma \in I\!R^{p_{\mathrm{sum}}}$:

$$\frac{1}{s(\Gamma)}\exp\left(\beta\frac{\|\boldsymbol{\gamma}_j\|^2 - z_j^2}{2z_j}\right) \leq 1, \quad \sum_{j=1}^{K}\frac{1}{s(\Gamma)}\exp\left(\beta\frac{\|\boldsymbol{\gamma}_j\|^2 - z_j^2}{2z_j}\right) = 1.$$

Therefore,

$$\sum_{q,l=1}^{p_{\mathrm{sum}}}\left|\frac{d}{d\gamma^q}h_\beta(s(\Gamma))\frac{d}{d\gamma^l}h_\beta(s(\Gamma))\right| \leq \left\{\sum_{j=1}^{K}\frac{1}{s(\Gamma)z_j}\exp\left(\beta\frac{\|\boldsymbol{\gamma}_j\|^2 - z_j^2}{2z_j}\right)\sum_{q=1}^{p_j}\gamma^q\right\}^2$$

$$\leq \left|\max_{1\leq j\leq K}\|\boldsymbol{\gamma}_j\|\frac{\sqrt{p_j}}{z_j}\right|^2$$

$$\leq \max_{1\leq j\leq K}\|\boldsymbol{\gamma}_j\|^2 \quad \text{for } z_j \geq \sqrt{p_j}.$$

Similarly

$$\sum_{q,l=1}^{p_{\mathrm{sum}}}\left|\frac{d^2}{d\gamma^q d\gamma^l}h_\beta(s(\Gamma))\right| \leq \mathtt{C}\beta\max_{1\leq j\leq K}\|\boldsymbol{\gamma}_j\|^2,$$

$$\sum_{q,l,b=1}^{p_{\mathrm{sum}}}\left|\frac{d^2}{d\gamma^q d\gamma^l}h_\beta(s(\Gamma))\frac{d}{d\gamma^b}h_\beta(s(\Gamma)) + \frac{d^3}{d\gamma^q d\gamma^l d\gamma^b}h_\beta(s(\Gamma))\right| \leq \mathtt{C}\left(\beta + \beta^2\right)\max_{1\leq j\leq K}\|\boldsymbol{\gamma}_j\|^3.$$

□

A.2 Gaussian comparison

The following Lemma shows how to compare the expected values of a twice differentiable function evaluated at the independent centered Gaussian vectors. This statement is used

for the Gaussian comparison step in the scheme (3.1). The proof of the result is based on the Gaussian interpolation method introduced by Stein (1981) and Slepian (1962) (see also Röllin (2013) and Chernozhukov et al. (2013b) and references therein). The proof is given here in order to keep the text self-contained.

Lemma A.7 (Gaussian comparison using Slepian interpolation)**.** *Let the $\mathbb{R}^{p_{\text{sum}}}$-dimensional random centered vectors $\overline{\Phi}$ and $\overline{\Psi}$ be independent and normally distributed, $f(Z) : \mathbb{R}^{p_{\text{sum}}} \mapsto \mathbb{R}$ is any twice differentiable function s.t. the expected values in the expression below are bounded. Then it holds*

$$\left|\mathbb{E} f(\overline{\Phi}) - \mathbb{E} f(\overline{\Psi})\right| \leq \frac{1}{2} \left\|\operatorname{Var}\overline{\Phi} - \operatorname{Var}\overline{\Psi}\right\|_{\max} \sup_{t\in[0,1]} \left\|\mathbb{E}\nabla^2 f\left(\overline{\Phi}\sqrt{t} + \overline{\Psi}\sqrt{1-t}\right)\right\|_1 .$$

Proof of Lemma A.7. Introduce for $t \in [0,1]$ the Gaussian vector process Z_t and the deterministic scalar-valued function $\varkappa(t)$:

$$\begin{aligned} Z_t &\stackrel{\text{def}}{=} \overline{\Phi}\sqrt{t} + \overline{\Psi}\sqrt{1-t} \in \mathbb{R}^{p_{\text{sum}}}, \\ \varkappa(t) &\stackrel{\text{def}}{=} \mathbb{E} f(Z(t)), \end{aligned}$$

then $\mathbb{E} f(\overline{\Phi}) = \varkappa(1)$, $\mathbb{E} f(\overline{\Psi}) = \varkappa(0)$ and

$$\left|\mathbb{E} f(\overline{\Phi}) - \mathbb{E} f(\overline{\Psi})\right| = |\varkappa(1) - \varkappa(0)| \leq \int_0^1 \left|\varkappa'(t)\right| dt.$$

Let us consider $\varkappa'(t)$:

$$\begin{aligned} \varkappa'(t) &= \frac{d}{dt}\mathbb{E} f(Z_t) = \mathbb{E}\left[\{\nabla f(Z_t)\}^\top \frac{d}{dt} Z_t\right] \\ &= \frac{1}{2\sqrt{t}}\mathbb{E}\left\{\overline{\Phi}^\top \nabla f(Z_t)\right\} - \frac{1}{2\sqrt{1-t}}\mathbb{E}\left\{\overline{\Psi}^\top \nabla f(Z_t)\right\}. \end{aligned} \tag{A.18}$$

Further we use the Gaussian integration by parts formula (see e.g Section A.6 in Talagrand (2003)): if $(x_1, \ldots, x_{p_{\text{sum}}})^\top$ is a centered Gaussian vector and $f(x_1, \ldots, x_{p_{\text{sum}}})$ is s.t. the integrals below exist, then it holds for all $j = 1, \ldots, p_{\text{sum}}$:

$$\mathbb{E}\left\{x_j f(x_1, \ldots, x_{p_{\text{sum}}})\right\} = \sum_{k=1}^{p_{\text{sum}}} \mathbb{E}(x_j x_k)\mathbb{E}\left\{\frac{d}{dx_k} f(x_1, \ldots, x_{p_{\text{sum}}})\right\}. \tag{A.19}$$

Let $\overline{\Phi}^j, \overline{\Psi}^j$ denote the j-th coordinates of $\overline{\Phi}$ and $\overline{\Psi}$. Let also $\frac{d}{d_j} f(Z_t)$ denote the partial derivative of the vectors $f(Z_t)$ w.r.t. the j-th coordinate of Z_t. Then it holds

due to (A.19):

$$\mathbb{E}\left\{\overline{\Phi}^\top \nabla f(Z_t)\right\} = \sum_{j=1}^{p_{\text{sum}}} \mathbb{E}\left\{\overline{\Phi}^j \frac{d}{d_j} f(Z_t)\right\} = \sum_{j,q=1}^{p_{\text{sum}}} \mathbb{E}\left(\overline{\Phi}^j \overline{\Phi}^q\right) \mathbb{E}\left\{\frac{d}{d\overline{\Phi}^q}\frac{d}{d_j} f(Z_t)\right\}$$
$$= \sqrt{t} \sum_{j,q=1}^{p_{\text{sum}}} \mathbb{E}\left(\overline{\Phi}^j \overline{\Phi}^q\right) \mathbb{E}\left\{\frac{d^2}{d_q d_j} f(Z_t)\right\}.$$

Similarly for the second term in (A.18):

$$\mathbb{E}\left\{\overline{\Psi}^\top \nabla f(Z_t)\right\} = \sqrt{1-t} \sum_{j,q=1}^{p_{\text{sum}}} \mathbb{E}\left(\overline{\Psi}^j \overline{\Psi}^q\right) \mathbb{E}\left\{\frac{d^2}{d_q d_j} f(Z_t)\right\},$$

therefore

$$\varkappa'(t) = \frac{1}{2} \sum_{j=1}^{p_{\text{sum}}} \sum_{q=1}^{p_{\text{sum}}} \left\{\mathbb{E}\left(\overline{\Phi}^j \overline{\Phi}^q\right) - \mathbb{E}\left(\overline{\Psi}^j \overline{\Psi}^q\right)\right\} \mathbb{E}\left\{\frac{d^2}{d_q d_j} f(Z_t)\right\}$$
$$\leq \frac{1}{2} \left\|\operatorname{Var} \overline{\Phi} - \operatorname{Var} \overline{\Psi}\right\|_{\max} \sup_{t\in[0,1]} \left\|\mathbb{E}\nabla^2 f(Z_t)\right\|_1.$$

□

A.3 Simultaneous anti-concentration for ℓ_2-norms of Gaussian vectors

Lemma A.8 (Simultaneous Gaussian anti-concentration). *Let* $\left(\overline{\phi}_1^\top, \ldots, \overline{\phi}_K^\top\right)^\top \in \mathbb{R}^{p_{\text{sum}}}$ *be centered normally distributed random vector, and* $\overline{\phi}_j \in \mathbb{R}^{p_j}$, $j = 1, \ldots, K$. *It holds for all* $z_j \geq \sqrt{p_j}$ *and* $0 < \Delta_j \leq z_j$, $j = 1, \ldots, K$:

$$\mathbb{P}\left(\bigcup_{j=1}^K \{\|\overline{\phi}_j\| > z_j\}\right) - \mathbb{P}\left(\bigcup_{j=1}^K \{\|\overline{\phi}_j\| > z_j + \Delta_j\}\right) \leq \Delta_{\text{ac}}(\{\Delta_j\}),$$

where

$$\Delta_{\text{ac}}(\{\Delta_j\}) \leq \mathtt{C}\left\{\varkappa\sqrt{1 \vee \log(K/2)} + \mathtt{C} \max_{1\leq j\leq K}\{\Delta_j\} \sqrt{\max_{1\leq j\leq K} \log(2z_j/\Delta_j)}\right\},$$

and $\varkappa \stackrel{\text{def}}{=} \max_{1\leq j\leq K}\{\Delta_j/z_j\} \leq 1$ *is a deterministic positive constant. An explicit definition of* $\Delta_{\text{ac}}(\{\Delta_j\})$ *is given in* (A.22).

Proof of Lemma A.8.

$$
\begin{aligned}
&I\!P\left(\bigcup\nolimits_{j=1}^{K}\{\|\overline{\phi}_j\| > z_j\}\right) - I\!P\left(\bigcup\nolimits_{j=1}^{K}\{\|\overline{\phi}_j\| > z_j + \Delta_j\}\right) \\
&\quad \le I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\overline{\phi}_j\| z_j^{-1} - 1 > 0\right\}\right) - I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\overline{\phi}_j\| z_j^{-1} - 1 > \varkappa\right\}\right) \\
&\quad = I\!P\left(\max_{1\le j\le K}\left\{\|\overline{\phi}_j\| z_j^{-1} - 1\right\} > 0\right) - I\!P\left(\max_{1\le j\le K}\left\{\|\overline{\phi}_j\| z_j^{-1} - 1\right\} > \varkappa\right) \\
&\quad \le I\!P\left(0 \le \max_{1\le j\le K}\left\{\|\overline{\phi}_j\| z_j^{-1} - 1\right\} \le \varkappa\right). \qquad \text{(A.20)}
\end{aligned}
$$

It holds

$$
\|\overline{\phi}_j\| = \sup_{\substack{\gamma\in I\!R^{p_j},\\ \|\gamma\|=1}} \left\{\gamma^\top \overline{\phi}_j\right\}.
$$

Let $\mathrm{G}_j(\varepsilon_j) \subset I\!R^{p_j}$ (for $1 \le j \le K$) denote a finite ε_j-net on $(p_j - 1)$-sphere of radius 1:

$$
\forall \gamma \in I\!R^{p_j} \text{ s.t. } \|\gamma\| = 1 \;\; \exists \gamma_0 \in \mathrm{G}_j(\varepsilon_j): \;\; \|\gamma - \gamma_0\| \le \varepsilon_j, \;\; \|\gamma_0\| = 1.
$$

This implies for all $j = 1, \ldots, K$

$$
(1 - \varepsilon_j)\|\overline{\phi}_j\| \le \max_{\gamma\in \mathrm{G}_j(\varepsilon_j)} \left\{\gamma^\top \overline{\phi}_j\right\} \le \|\overline{\phi}_j\|.
$$

Let us take $\varepsilon_1, \ldots, \varepsilon_K > 0$ s.t. $\forall\, j = 1, \ldots, K$

$$
\varepsilon_j \|\overline{\phi}_j\| z_j^{-1} \le \varkappa, \qquad \text{(A.21)}
$$

then

$$
0 \le \max_{1\le j\le K}\left\{\frac{\|\overline{\phi}_j\|}{z_j}\right\} - \max_{1\le j\le K}\max_{\gamma\in \mathrm{G}_j(\varepsilon_j)}\left\{\frac{\gamma^\top \overline{\phi}_j}{z_j}\right\} \le \varkappa,
$$

and the inequality (A.20) continues as

$$
\begin{aligned}
&I\!P\left(0 \le \max_{1\le j\le K}\left\{\|\overline{\phi}_j\| z_j^{-1} - 1\right\} \le \varkappa\right) \\
&\quad \le I\!P\left(\left|\max_{1\le j\le K}\sup_{\gamma\in \mathrm{G}_j(\varepsilon_j)}\left\{\frac{\gamma^\top \overline{\phi}_j}{z_j}\right\} - 1\right| \le \varkappa\right).
\end{aligned}
$$

The random values $\gamma^\top \overline{\phi}_j z_j^{-1} \sim \mathcal{N}(0, z_j^{-2}\,\mathrm{Var}\{\gamma^\top \overline{\phi}_j\})$. The anti-concentration inequality by Chernozhukov et al. (2014c) for the maximum of a centered high-dimensional

Gaussian vector (see Theorem A.9 below), applied to $\max_{1\le j\le K}\sup_{\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j)}\left\{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j z_j^{-1}\right\}$, implies

$$\mathbb{P}\left(\left|\max_{1\le j\le K}\sup_{\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j)}\left\{\frac{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j}{z_j}\right\}-1\right|\le \varkappa\right)$$
$$\le \Delta_{\mathrm{ac}} \stackrel{\mathrm{def}}{=} \mathtt{C}_{\mathrm{ac}}\varkappa\sqrt{1\vee\log\left(\varkappa^{-1}\sum\nolimits_{j=1}^{K}\{2/\varepsilon_j\}^{p_j}\right)}, \tag{A.22}$$

where the constant $\mathtt{C}_{\mathrm{ac}}$ depends on min and max of $\mathrm{Var}\{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j z_j^{-1}\}\le \mathbb{E}\|\overline{\boldsymbol{\phi}}_j\|^2 z_j^{-2}\le 1$; the sum $\sum_{j=1}^{K}\{2/\varepsilon_j\}^{p_j}$ is proportional to cardinality of the set $\{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j z_j^{-1},\,\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j),\,j=1,\ldots,K\}$. If one takes $\varepsilon_j = 2\mathtt{C}\{\Delta_j/(2z_j)\}^{\frac{p_{\min}+1}{p_j+1}}$, then (A.21) holds with exponentially high probability due to Gaussianity of the vectors $\overline{\boldsymbol{\phi}}_j$ and Theorem 1.2 in Spokoiny (2012b), hence

$$\Delta_{\mathrm{ac}} \le \mathtt{C}_{\mathrm{ac}}\varkappa\sqrt{1\vee\mathtt{C}\log\left(\frac{1}{2}\sum\nolimits_{j=1}^{K}\{2/\varepsilon_j\}^{p_j+1}\right)}$$
$$\le \mathtt{C}_{\mathrm{ac}}\left\{\varkappa\sqrt{1\vee\log(K/2)}+\mathtt{C}\max_{1\le j\le K}\{\Delta_j\}\sqrt{\max_{1\le j\le K}\log(2z_j/\Delta_j)}\right\}. \tag{A.23}$$

□

Theorem A.9 (Anti-concentration inequality for maxima of a Gaussian random vector, Chernozhukov et al. (2014c)). *Let $(X_1,\ldots,X_p)^\top$ be a centered Gaussian random vector with $\sigma_j^2 \stackrel{\mathrm{def}}{=} \mathbb{E}X_j^2>0$ for all $1\le j\le p$. Let $\underline{\sigma}\stackrel{\mathrm{def}}{=}\min_{1\le j\le p}\sigma_j$, $\overline{\sigma}\stackrel{\mathrm{def}}{=}\max_{1\le j\le p}\sigma_j$. Then for every $\epsilon>0$*

$$\sup_{x\in\mathbb{R}}\mathbb{P}\left(\left|\max_{1\le j\le p}X_j - x\right|\le\epsilon\right)\le \mathtt{C}_{\mathrm{ac}}\epsilon\sqrt{1\vee\log(p/\epsilon)},$$

where $\mathtt{C}_{\mathrm{ac}}$ depends only on $\underline{\sigma}$ and $\overline{\sigma}$. When the variances are all equal, namely $\underline{\sigma}=\overline{\sigma}=\sigma$, $\log(p/\epsilon)$ on the right side can be replaced by $\log p$.

A.4 Proof of Proposition A.1

Proof of Proposition A.1. Let $\Phi\stackrel{\mathrm{def}}{=}\left(\boldsymbol{\phi}_1^\top,\ldots,\boldsymbol{\phi}_K^\top\right)^\top\in\mathbb{R}^{p_{\mathrm{sum}}}$ for $p_{\mathrm{sum}}\stackrel{\mathrm{def}}{=}p_1+\cdots+p_K$ (as in (A.5)), and similarly $\Psi\stackrel{\mathrm{def}}{=}\left(\boldsymbol{\psi}_1^\top,\ldots,\boldsymbol{\psi}_K^\top\right)^\top\in\mathbb{R}^{p_{\mathrm{sum}}}$. Let also $\overline{\Phi}\sim\mathcal{N}(0,\mathrm{Var}\,\Phi)$ and $\overline{\Psi}\sim\mathcal{N}(0,\mathrm{Var}\,\Psi)$. Introduce the following value, which comes from Lemma A.7 on Gaussian comparison:

$$\delta_2(\Delta,\beta)\stackrel{\mathrm{def}}{=}\mathtt{C}_2(\Delta,\beta)\max_{1\le j\le K}\sup_{t\in[0,1]}\left\{\mathbb{E}\|\overline{\boldsymbol{\phi}}_j\sqrt{t}+\overline{\boldsymbol{\psi}}_j\sqrt{1-t}\|^2\right\}$$
$$\le\mathtt{C}_2(\Delta,\beta)\max_{1\le j\le K}\max\left\{\mathrm{tr}\,\mathrm{Var}(\overline{\boldsymbol{\phi}}_j),\mathrm{tr}\,\mathrm{Var}(\overline{\boldsymbol{\psi}}_j)\right\}. \tag{A.24}$$

It holds

$$
\begin{aligned}
&I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) \\
&\overset{\text{by L. A.3}}{\geq} I\!E H_{\Delta,\beta}\left(\overline{\Phi}, \boldsymbol{z} + \frac{3\log(K)}{2\beta}\mathbf{1}_K\right) - \delta_{3,\phi}(\Delta,\beta) \\
&\overset{\text{by L. A.7, A.6}}{\geq} I\!E H_{\Delta,\beta}\left(\overline{\Psi}, \boldsymbol{z} + \frac{3\log(K)}{2\beta}\mathbf{1}_K\right) - \frac{1}{2}\delta_{\Sigma}^2\delta_2(\Delta,\beta) - \delta_{3,\phi}(\Delta,\beta) \\
&\overset{\text{by L. A.3}}{\geq} I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\overline{\boldsymbol{\psi}}_j\| > z_j + \Delta + \frac{3\log(K)}{2\beta}\right\}\right) - \frac{1}{2}\delta_{\Sigma}^2\delta_2(\Delta,\beta) - \delta_{3,\phi}(\Delta,\beta) \\
&\overset{\text{by L. A.8}}{\geq} I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\overline{\boldsymbol{\psi}}_j\| > z_j - \delta_{z_j} - \Delta\right\}\right) - \frac{1}{2}\delta_{\Sigma}^2\delta_2(\Delta,\beta) - \delta_{3,\phi}(\Delta,\beta) \\
&\qquad - 2\Delta_{\mathrm{ac}}\left(\left\{\delta_{z_j}\right\} + 2\Delta + \frac{3\log(K)}{\beta}\right) \qquad \text{(A.25)} \\
&\overset{\text{by L. A.2}}{\geq} I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\boldsymbol{\psi}_j\| > z_j - \delta_{z_j}\right\}\right) - \frac{1}{2}\delta_{\Sigma}^2\delta_2(\Delta,\beta) \qquad \text{(A.26)} \\
&\qquad - \delta_{3,\phi}(\Delta,\beta) - \delta_{3,\psi}(\Delta,\beta) - 2\Delta_{\mathrm{ac}}\left(\left\{\delta_{z_j}\right\} + 2\Delta + \frac{3\log(K)}{\beta}\right),
\end{aligned}
$$

where $\delta_{3,\psi}(\Delta,\beta)$ is defined similarly to $\delta_{3,\phi}(\Delta,\beta)$ in (A.15):

$$
\delta_{3,\psi}(\Delta,\beta) \overset{\text{def}}{=} \frac{\mathtt{C}_3(\Delta,\beta)}{3}\frac{p_{\max}^{3/2}}{n^{1/2}}\log^{1/2}(K)\log^{3/2}(np_{\mathrm{sum}})\left(2\nu_0^2\mathfrak{c}_{\psi}^2\lambda_{\psi,\max}^2\right)^{3/2}. \qquad \text{(A.27)}
$$

By Lemma A.8 inequality (A.25) requires the following: $\delta_{z_j} + 2\Delta + \frac{3\log(K)}{\beta} \leq z_j$ .The bound in the inverse direction is derived similarly. Denote the approximating error term obtained in (A.26) as

$$
\Delta_{\ell_2} \overset{\text{def}}{=} \frac{1}{2}\delta_{\Sigma}^2\delta_2(\Delta,\beta) + \delta_{3,\phi}(\Delta,\beta) + \delta_{3,\psi}(\Delta,\beta) + 2\Delta_{\mathrm{ac}}\left(\left\{\delta_{z_j}\right\} + 2\Delta + \frac{3\log(K)}{\beta}\right).
$$

Consider this term in more details, by inequality (A.23)

$$
\begin{aligned}
\Delta_{\mathrm{ac}}\left(\left\{\delta_{z_j}\right\} + 2\Delta + \frac{3\log(K)}{\beta}\right) &\leq \max_{1\leq j\leq K}\left(\delta_{z_j} + 2\Delta + \frac{3\log(K)}{\beta}\right) \\
&\times \left\{\mathtt{C}\frac{\log^{1/2}(K)}{z_j} + \log^{1/2}(2z_{\max}) - \log^{1/2}\left(\delta_{z_j} + 2\Delta + \frac{3\log(K)}{\beta}\right)\right\}.
\end{aligned}
$$

Let us take $\beta = \frac{\log(K)}{\Delta}$, then

$$\begin{aligned}
\Delta_{\text{ac}} &\leq 5\mathtt{C}\Delta\frac{\log^{1/2}(K)}{z_{\min}} + \mathtt{C}\max_{1\leq j\leq K}\frac{\delta_{z_j}}{z_j}\log^{1/2}(K)\\
&\quad + \mathtt{C}\left(5\Delta+\delta_{z,\max}\right)\left(\log^{1/2}\left(2z_{\max}\right) + \sqrt{-\log\left(\delta_{z,\min}+5\Delta\right)}\right),\\
&\leq 5\mathtt{C}\Delta\frac{\log^{1/2}(K)}{z_{\min}} + \mathtt{C}\max_{1\leq j\leq K}\frac{\delta_{z_j}}{z_j}\log^{1/2}(K)\\
&\quad + 2\mathtt{C}\left(5\Delta+\delta_{z,\max}\right)\sqrt{-\log\left(\delta_{z,\min}+5\Delta\right)}\\
&\leq 5\mathtt{C}\Delta\frac{\log^{1/2}(K)}{z_{\min}} + \mathtt{C}\max_{1\leq j\leq K}\frac{\delta_{z_j}}{z_j}\log^{1/2}(K) + 2\mathtt{C}\left(5\Delta+\delta_{z,\max}\right)\sqrt{-\log\left(5\Delta\right)}\\
&\leq 5\mathtt{C}\Delta\Big\{\frac{\log^{1/2}(K)}{z_{\min}} + 2.4\log^{1/2}\big(5n^{1/2}\big)\Big\} + \mathtt{C}\max_{1\leq j\leq K}\frac{\delta_{z_j}}{z_j}\log^{1/2}(K)\\
&\leq 6\mathtt{C}\Delta\Big\{\frac{\log^{1/2}(K)}{z_{\min}} + 0.4\log^{1/2}\big(5n^{1/2}\big)\Big\}, \qquad \text{(A.28)}
\end{aligned}$$

where the second inequality holds for $\delta_{z,\min} + 5\Delta \leq 1/(2z_{\max})$, and the last one holds for $\delta_{z,\max} \leq \Delta$ and $\Delta \geq n^{-1/2}$.

$$\delta_{3,\phi}(\Delta,\beta) + \delta_{3,\psi}(\Delta,\beta) \overset{\text{by (A.27)}}{\leq} \mathtt{C}\frac{\log^{5/2}(K)}{\Delta^3}\frac{p_{\max}^{3/2}}{n^{1/2}}\log^{3/2}(np_{\text{sum}})\left(\lambda_{\phi,\max}^3 + \lambda_{\psi,\max}^3\right), \quad \text{(A.29)}$$

$$\begin{aligned}
\delta_\Sigma\delta_2(\Delta,\beta) &\overset{\text{by (A.24)}}{\leq} \mathtt{C}\delta_\Sigma^2\frac{\log(K)}{\Delta^2}\max_{1\leq j\leq K}\max\left\{\operatorname{tr}\operatorname{Var}(\overline{\boldsymbol{\phi}}_j), \operatorname{tr}\operatorname{Var}(\overline{\boldsymbol{\psi}}_j)\right\}\\
&\leq \mathtt{C}\delta_\Sigma^2\frac{\log(K)}{\Delta^2}p_{\max}\max\left\{\lambda_{\phi,\max}^2, \lambda_{\psi,\max}^2\right\}.
\end{aligned}$$

After minimizing the sum of the expressions (A.28) and (A.29) w.r.t Δ, we have

$$\begin{aligned}
\Delta_{\ell_2} &\leq 12.5\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8}\log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}^{3/4}\\
&\quad + 3.2\mathtt{C}\delta_\Sigma^2 p_{\max}z_{\min}^{1/2}\left(\frac{p_{\max}^3}{n}\right)^{1/4}\log^2(K)\log^{3/4}(np_{\text{sum}})\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}^{7/2}\\
&\leq 25\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8}\log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}^{3/4},
\end{aligned}$$

where the last inequality holds for

$$\delta_\Sigma^2 \leq 4\mathtt{C}p_{\max}^{-1}z_{\min}^{-1/2}\left(\frac{p_{\max}^3}{n}\right)^{-1/8}\log^{-7/8}(K)\log^{-3/8}(np_{\text{sum}})\left(\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}\right)^{-11/4}.$$

□

B Square-root Wilks approximations

This section's goal is to derive square root Wilks approximations simultaneously for K parametric models, for the $\boldsymbol{Y}$ and bootstrap worlds. This is done in Section B.3 below. Both of the results are used in the approximating scheme (3.1) for the bootstrap justification. In order to make the text self-contained we recall in Section B.1 some results from the general finite sample theory by Spokoiny (2012a,b, 2013). In Section B.2 we recall similar finite sample results for the bootstrap world for a single parametric model, obtained in Spokoiny and Zhilova (2014).

B.1 Finite sample theory

Let us use the notations given in the introduction: $L_k(\boldsymbol{\theta})$, $k = 1, \ldots, K$ are the log-likelihood processes, which depend on the data $\boldsymbol{Y}$ and correspond to the regular parametric families of probability distributions $\{I\!P_k(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_k \subset I\!R^{p_k}\}$. The general finite sample approach by Spokoiny (2012a) does not require that the true distribution $I\!P$ of the data $\boldsymbol{Y}$ belongs to any of the parametric families $\{I\!P_k(\boldsymbol{\theta})\}$. The target parameters $\boldsymbol{\theta}_k^*$ are defined as in (1.3) by projection of the true measure $I\!P$ on $\{I\!P_k(\boldsymbol{\theta})\}$. Let D_k^2 denote the full Fisher information $p_k \times p_k$ matrices, which are deterministic, symmetric and positive-definite:

$$D_k^2 \stackrel{\text{def}}{=} -\nabla_{\boldsymbol{\theta}}^2 I\!E L_k(\boldsymbol{\theta}_k^*).$$

Centered p_k-dimensional random vectors $\boldsymbol{\xi}_k$ denote the normalised scores:

$$\boldsymbol{\xi}_k \stackrel{\text{def}}{=} D_k^{-1} \nabla_{\boldsymbol{\theta}} L_k(\boldsymbol{\theta}_k^*).$$

Introduce the following elliptic vicinities around the true points $\boldsymbol{\theta}_k^*$:

$$\Theta_{0,k}(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \Theta_k : \|D_k(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*)\| \le \mathtt{r}\}.$$

Let $1 \le k \le K$ be fixed. The non-asymptotic Wilks approximating bound by Spokoiny (2012a, 2013) requires that the maximum likelihood estimate $\widetilde{\boldsymbol{\theta}}_k$ gets into the local vicinity $\Theta_{0,k}(\mathtt{r}_{0,k})$ of some radius $\mathtt{r}_{0,k} > 0$ with probability $\ge 1 - 3\mathrm{e}^{-\mathtt{x}}$, $\mathtt{x} > 0$. This is guaranteed by the following concentration result:

Theorem B.1 (Concentration of the MLE, Spokoiny (2013)). *Let the conditions* $(\boldsymbol{ED_0})$, $(\boldsymbol{ED_2})$, $(\boldsymbol{\mathcal{L}_0})$, $(\boldsymbol{\mathcal{I}})$ *and* $(\boldsymbol{\mathcal{L}\mathtt{r}})$ *be fulfilled. If for each* $k = 1, \ldots, K$ *for the constants* $\mathtt{r}_{0,k} > 0$ *and for the functions* $\mathtt{b}_k(\mathtt{r})$ *from* $(\boldsymbol{\mathcal{L}\mathtt{r}})$ *holds:*

$$\mathtt{b}_k(\mathtt{r})\mathtt{r} \ge 2\left\{\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B_k) + 6\omega_k\nu_k\,\mathfrak{Z}_k(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_{0,k}))\right\}, \quad \mathtt{r} > \mathtt{r}_{0,k} \tag{B.1}$$

where the functions $\mathfrak{z}_k(\mathtt{x})$ *and* $\mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, I\!B_k)$ *are defined in* (B.3) *and* (B.4) *respectively, then it holds for all* $k = 1, \dots, K$

$$I\!P\left(\widetilde{\boldsymbol{\theta}}_k \notin \Theta_{0,k}(\mathtt{r}_{0,k})\right) \leq 3\mathrm{e}^{-\mathtt{x}}.$$

The constants ω_k, ν_k *and* $\mathfrak{a}_k$ *come from the imposed conditions* $(\boldsymbol{ED_0})$–$(\boldsymbol{\mathcal{I}})$ *(from Section 5). In the case 5.3* $\mathtt{r}_{0,k} \geq \mathtt{C}\sqrt{p_k + \mathtt{x}}$.

Theorem B.2 (Wilks approximation, Spokoiny (2013))**.** *Under the conditions of Theorem B.1 for some* $\mathtt{r}_{0,k} > 0$ *s.t.* (B.1) *is fulfilled, it holds for each* $k = 1, \dots, K$ *with probability* $\geq 1 - 5\mathrm{e}^{-\mathtt{x}}$

$$\left|2\Big\{L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*)\Big\} - \|\boldsymbol{\xi}_k\|^2\right| \leq \Delta_{k,\mathrm{W}^2}(\mathtt{r}_{0,k}, \mathtt{x}),$$
$$\left|\sqrt{2\Big\{L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*)\Big\}} - \|\boldsymbol{\xi}_k\|\right| \leq \Delta_{k,\mathrm{W}}(\mathtt{r}_{0,k}, \mathtt{x})$$

for

$$\Delta_{k,\mathrm{W}}(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} 3\mathtt{r}\left\{\delta(\mathtt{r}) + 6\nu_k\, \mathfrak{z}_k(\mathtt{x})\omega_k\right\}, \tag{B.2}$$
$$\Delta_{k,\mathrm{W}^2}(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} \frac{2}{3}\left\{2\mathtt{r} + \mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, I\!B_k)\right\}\Delta_{k,\mathrm{W}}(\mathtt{r}, \mathtt{x}),$$
$$\mathfrak{z}_k(\mathtt{x}) \stackrel{\mathrm{def}}{=} 2\sqrt{p_k} + \sqrt{2\mathtt{x}} + 4p_k(\mathtt{x}\mathtt{g}_k^{-2} + 1)\mathtt{g}_k^{-1}. \tag{B.3}$$

In the case 5.3 it holds for $\mathtt{r} \leq \mathtt{r}_{0,k}$ *:*

$$\Delta_{k,\mathrm{W}}(\mathtt{r}, \mathtt{x}) \leq \mathtt{C}\,\frac{p_k + \mathtt{x}}{\sqrt{n}}, \quad \Delta_{k,\mathrm{W}^2}(\mathtt{r}, \mathtt{x}) \leq \mathtt{C}\sqrt{\frac{(p_k + \mathtt{x})^3}{n}}.$$

The constants $\mathtt{g}_k$ *and* $\delta_k(\mathtt{r})$ *come from the imposed conditions* $(\boldsymbol{ED_0})$, $(\boldsymbol{\mathcal{L}_0})$ *(from Section 5). The function* $\mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, I\!B_k)$, *defined in* (B.4)*, corresponds to the quantile function of deviations of the approximating random value* $\|\boldsymbol{\xi}_k\|$ *(see Theorem B.3 below).*

The following theorem characterizes the tail behaviour of the approximating terms $\|\boldsymbol{\xi}_k\|^2$. It means that with bounded exponential moments of the vectors $\boldsymbol{\xi}_k$ (conditions $(\boldsymbol{ED_0})$, $(\boldsymbol{\mathcal{I}})$) its squared Euclidean norms $\|\boldsymbol{\xi}_k\|^2$ have three regimes of deviations: sub-Gaussian, Poissonian and large-deviations' zone.

Theorem B.3 (Deviation bound for a random quadratic form, Spokoiny (2012b))**.** *Let condition* $(\boldsymbol{ED_0})$ *be fulfilled, then for* $\mathtt{g}_k \geq \sqrt{2\,\mathrm{tr}(I\!B_k^2)}$ *it holds for each* $k = 1, \dots, K$ *:*

$$I\!P\left(\|\boldsymbol{\xi}_k\|^2 \geq \mathfrak{z}_{\mathtt{qf}}^2(\mathtt{x}, I\!B_k)\right) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_{c,k}},$$

where $I\!B_k^2 \stackrel{\text{def}}{=} D_k^{-1} V_k^2 D_k^{-1}$, $\lambda(I\!B_k)$ *is a maximum eigenvalue of* $I\!B_k^2$,

$$\mathfrak{Z}_{\mathtt{qf}}^2(\mathtt{x}, I\!B_k) \stackrel{\text{def}}{=} \begin{cases} \operatorname{tr}(I\!B_k^2) + \sqrt{8 \operatorname{tr}(I\!B_k^4)\mathtt{x}}, & \mathtt{x} \le \sqrt{2 \operatorname{tr}(I\!B_k^4)}/\{18\lambda(I\!B_k)\}, \\ \operatorname{tr}(I\!B_k^2) + 6\mathtt{x}\lambda(I\!B_k), & \sqrt{2 \operatorname{tr}(I\!B_k^4)}/\{18\lambda(I\!B_k)\} < \mathtt{x} \le \mathtt{x}_{c,k}, \\ |\mathtt{z}_{c,k} + 2(\mathtt{x} - \mathtt{x}_{c,k})/\mathtt{g}_{c,k}|^2 \lambda(I\!B_k), & \mathtt{x} > \mathtt{x}_{c,k}, \end{cases} \tag{B.4}$$

$$\begin{aligned} 2\mathtt{x}_{c,k} &\stackrel{\text{def}}{=} 2\mathtt{x}_{c,k}(I\!B_k) \stackrel{\text{def}}{=} \mu_c \mathtt{z}_{c,k}^2 + \log\det\left(\boldsymbol{I}_{p_k} - \mu_c I\!B_k^2/\lambda(I\!B_k)\right), \\ \mathtt{z}_{c,k}^2 &\stackrel{\text{def}}{=} \left\{\mathtt{g}_k^2/\mu_c^2 - \operatorname{tr}\left(I\!B_k^2\right)/\mu_c\right\}/\lambda(I\!B_k), \\ \mathtt{g}_{c,k} &\stackrel{\text{def}}{=} \sqrt{\mathtt{g}_k^2 - \mu_c \operatorname{tr}\left(I\!B_k^2\right)}/\sqrt{\lambda(I\!B_k)}, \\ \mu_c &\stackrel{\text{def}}{=} 2/3. \end{aligned} \tag{B.5}$$

The matrices V_k^2 *come from condition* **(ED$_0$)** *and can be defined as*

$$V_k^2 \stackrel{\text{def}}{=} \operatorname{Var}\left\{\nabla_{\boldsymbol{\theta}} L_k(\boldsymbol{\theta}_k^*)\right\}.$$

By condition **($\mathcal{I}$)** $\operatorname{tr}(I\!B_k^2) \le \mathfrak{a}_k^2 p_k$, $\operatorname{tr}(I\!B^4) \le \mathfrak{a}_k^4 p_k$ *and* $\lambda(I\!B_k) \le \mathfrak{a}_k^2$. *In the case 5.3* $\mathtt{g}_k = \mathtt{C}\sqrt{n}$, *hence* $\mathtt{x}_{c,k} = \mathtt{C}n$, *and for* $\mathtt{x} \le \mathtt{x}_{c,k}$ *it holds:*

$$\mathfrak{Z}_{\mathtt{qf}}^2(\mathtt{x}, I\!B_k) \le \mathfrak{a}_k^2(p_k + 6\mathtt{x}). \tag{B.6}$$

B.2 Finite sample theory for the bootstrap world

Introduce for each $k = 1, \ldots, K$ the bootstrap score vectors at the point $\boldsymbol{\theta} \in \Theta_k$:

$$\begin{aligned} \boldsymbol{\xi}_k^\circ(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} D_k^{-1} \nabla_{\boldsymbol{\theta}} \boldsymbol{\zeta}_k^\circ(\boldsymbol{\theta}) \\ &= \sum_{i=1}^n D_k^{-1} \nabla_{\boldsymbol{\theta}} \ell_{i,k}(\boldsymbol{\theta})(u_i - 1). \end{aligned}$$

Theorem B.4 (Bootstrap Wilks approximation, Spokoiny and Zhilova (2014))**.** *Under the conditions of Theorems B.1 and B.5 for each* $k = 1, \ldots, K$ *and some* $\mathtt{r}_{0,k}^2 \ge 0$ *s.t.* (B.1) *and* (B.9) *are fulfilled, it holds for each* k *with* $I\!P$*-probability* $\ge 1 - 5\mathrm{e}^{-\mathtt{x}}$

$$I\!P^\circ\left(\left|\sup_{\boldsymbol{\theta} \in \Theta_k} 2\left\{L_k^\circ(\boldsymbol{\theta}) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)\right\} - \|\boldsymbol{\xi}_k^\circ(\widetilde{\boldsymbol{\theta}}_k)\|^2\right| \le \Delta_{k,\mathrm{W}^2}^\circ(\mathtt{r}_{0,k}, \mathtt{x})\right) \ge 1 - 4\mathrm{e}^{-\mathtt{x}},$$

$$I\!P^\circ\left(\left|\sqrt{\sup_{\boldsymbol{\theta} \in \Theta_k} 2\left\{L_k^\circ(\boldsymbol{\theta}) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)\right\}} - \|\boldsymbol{\xi}_k^\circ(\widetilde{\boldsymbol{\theta}}_k)\|\right| \le \Delta_{k,\mathrm{W}}^\circ(\mathtt{r}_{0,k}, \mathtt{x})\right) \ge 1 - 4\mathrm{e}^{-\mathtt{x}}.$$

where the error terms $\Delta^{\circ}_{k,\mathrm{W}}(\mathtt{r},\mathtt{x}), \Delta^{\circ}_{k,\mathrm{W}^2}(\mathtt{r},\mathtt{x})$ *are deterministic and*

$$\begin{aligned}\Delta^{\circ}_{k,\mathrm{W}}(\mathtt{r},\mathtt{x}) &\stackrel{\text{def}}{=} 2\Delta_{k,\mathrm{W}}(\mathtt{r},\mathtt{x}) + 36\nu_k \mathtt{r}\omega_{1,k}(\mathtt{r},\mathtt{x})\,\mathfrak{Z}_k(\mathtt{x}),\\ \Delta^{\circ}_{k,\mathrm{W}^2}(\mathtt{r},\mathtt{x}) &\stackrel{\text{def}}{=} \frac{1}{18}\left\{12\mathtt{r}\Delta^{\circ}_{k,\mathrm{W}}(\mathtt{r},\mathtt{x}) + \Delta^{\circ}_{k,\mathrm{W}}(\mathtt{r},\mathtt{x})^2\right\}.\end{aligned} \tag{B.7}$$

$\Delta_{k,\mathrm{W}}(\mathtt{r},\mathtt{x})$ *and* $\mathfrak{Z}_k(\mathtt{x})$ *are defined in* (B.2) *and* (B.3) *respectively and*

$$\omega_{1,k}(\mathtt{r},\mathtt{x}) = \omega_{1,k} \stackrel{\text{def}}{=} \frac{\mathtt{C}_{m,k}(\mathtt{r})}{\sqrt{n}} + 2\omega_k\nu_k\sqrt{2\mathtt{x}}, \tag{B.8}$$

where $\mathtt{C}_{m,k}(\mathtt{r}), \omega_k, \nu_k$ *come from the imposed conditions* $(\mathcal{L}_{0m})$, (ED_2) *and* (ED_0). *For the case 5.3 and* $\mathtt{r} \le \mathtt{r}_{0,k}$ *it holds:*

$$\Delta^{\circ}_{k,\mathrm{W}}(\mathtt{r},\mathtt{x}) \le \mathtt{C}\,\frac{p_k+\mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}}, \quad \Delta^{\circ}_{k,\mathrm{W}^2}(\mathtt{r},\mathtt{x}) \le \mathtt{C}\sqrt{\frac{(p_k+\mathtt{x})^3}{n}}\sqrt{\mathtt{x}}.$$

and $\omega_{1,k}(\mathtt{r}) \le \mathtt{C}\mathtt{r}/n + \mathtt{C}\sqrt{\mathtt{x}/n}$.

Theorem B.5 (Concentration of the bootstrap MLE, Spokoiny and Zhilova (2014))**.** *Let the conditions of Theorems B.1 and B.7,* $(\mathcal{L}_{0m})$ *and* (ED_{2m}) *be fulfilled. If the following holds for each* $k = 1,\dots,K$, $\omega_{1,k}(\mathtt{r},\mathtt{x})$ *defined in* (B.8) *and the* $I\!P$*-random matrices* $\mathcal{B}_k^2 \stackrel{\text{def}}{=} D_k^{-1}\,\mathrm{Var}^{\circ}\left\{\nabla_{\boldsymbol{\theta}} L^{\circ}_k(\boldsymbol{\theta}^*_k)\right\} D_k^{-1}$ *:*

$$\begin{aligned}\mathtt{b}_k(\mathtt{r})\mathtt{r} \ge 2\big\{ &\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B_k) + \mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}_k) + 6\nu_k\,\mathfrak{Z}_k(\mathtt{x})\omega_{1,k}(\mathtt{r}_{0,k})\mathtt{r}_{0,k}\big\}\\ &+ 12\nu_k(\omega_k + \omega_{1,k}(\mathtt{r},\mathtt{x}))\,\mathfrak{Z}_k(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_{0,k})) \quad \textit{for} \quad \mathtt{r} > \mathtt{r}_{0,k},\end{aligned} \tag{B.9}$$

then for each k *it holds with* $I\!P$*-probability* $\ge 1 - 3\mathrm{e}^{-\mathtt{x}}$

$$I\!P^{\circ}\left(\widetilde{\boldsymbol{\theta}}^{\circ}_k \notin \Theta_{0,k}(\mathtt{r}_{0,k})\right) \le 3\mathrm{e}^{-\mathtt{x}}.$$

Lemma B.6 below is implied straightforwardly by Lemma B.7 in Spokoiny and Zhilova (2014).

Lemma B.6. *Let the conditions of* (Eb), $(\mathcal{L}_{0m})$ *and* (ED_{2m}) *be fulfilled, then for each* $k = 1,\dots,K$ *it holds for* $\mathtt{r} \le \mathtt{r}_{0,k}$ *with* $I\!P$*-probability* $\ge 1 - \mathrm{e}^{-\mathtt{x}}$

$$I\!P^{\circ}\left(\sup_{\boldsymbol{\theta}\in\Theta_{0,k}(\mathtt{r})} \|\boldsymbol{\xi}^{\circ}_k(\boldsymbol{\theta}) - \boldsymbol{\xi}^{\circ}_k(\boldsymbol{\theta}^*_k)\| \le \Delta^{\circ}_{\xi,k}(\mathtt{r},\mathtt{x})\right) \ge 1 - \mathrm{e}^{-\mathtt{x}},$$

where

$$\Delta^{\circ}_{\xi,k}(\mathtt{r},\mathtt{x}) \stackrel{\text{def}}{=} 6\nu_k\,\mathfrak{Z}_k(\mathtt{x})\omega_{1,k}(\mathtt{r},\mathtt{x})\mathtt{r}$$

In the case 5.3 it holds for the bounding term

$$\Delta^{\circ}_{\xi}(\mathtt{r}_0,\mathtt{x}) \le \mathtt{C}\frac{p_k+\mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}}.$$

Theorem B.7 (Deviation bound for the bootstrap quadratic form, Spokoiny and Zhilova (2014)). *Let conditions* $(\boldsymbol{Eb})$, $(\boldsymbol{\mathcal{I}})$, $(\widehat{\boldsymbol{SD}_1})$, $(\boldsymbol{\mathcal{I}_B})$ *be fulfilled, then for each* $k = 1, \ldots, K$ *and* $\mathtt{g}_k \geq \sqrt{2\,\mathrm{tr}(\mathcal{B}_k^2)}$ *it holds:*

$$I\!P^\circ \left(\|\boldsymbol{\xi}_k^\circ(\boldsymbol{\theta}_k^*)\|^2 \leq \mathfrak{Z}_{\mathtt{qf}}^2(\mathtt{x}, \mathcal{B}_k) \right) \geq 1 - 2\mathrm{e}^{-\mathtt{x}} - 8.4\mathrm{e}^{-\mathtt{x}_{c,k}(\mathcal{B}_k)},$$

where

$$\mathcal{B}_k^2 \stackrel{\mathrm{def}}{=} D_k^{-1}\mathcal{V}^2(\boldsymbol{\theta}_k^*)D_k^{-1}, \quad \mathcal{V}_k^2(\boldsymbol{\theta}_k^*) \stackrel{\mathrm{def}}{=} \mathrm{Var}^\circ \nabla_{\boldsymbol{\theta}} L_k^\circ(\boldsymbol{\theta}_k^*),$$

$\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, \cdot)$ *and* $\mathtt{x}_{c,k}(\cdot)$ *are defined respectively in* (B.4) *and* (B.5). *Similarly to* (B.6) *it holds for* $\mathtt{x} \leq \mathtt{x}_{c,k}(\mathcal{B}_k)$ *:*

$$\mathfrak{Z}_{\mathtt{qf}}^2(\mathtt{x}, \mathcal{B}_k) \leq \mathfrak{a}_k^{\circ 2}(p_k + 6\mathtt{x})$$
$$\text{for } \mathfrak{a}_k^{\circ 2} \stackrel{\mathrm{def}}{=} (1 + \delta_{\mathcal{V},k}^2(\mathtt{x}))(\mathfrak{a}_k^2 + \mathfrak{a}_{B,k}^2)$$

and $\delta_{\mathcal{V},k}^2(\mathtt{x})$ *defined in* (C.1) *(see Section C.1 on Bernstein matrix inequalities).*

B.3 Simultaneous square-root Wilks approximations

The statements below follow from the results from Sections B.1 and B.2 by probability union bound.

Lemma B.8 (Simultaneous concentration bounds).

1. *Let conditions of Theorem B.1 be fulfilled and* (B.1) *hold for each* $k = 1, \ldots, K$ *with* $\mathtt{x} = \mathtt{x}_1 + \log(K)$ *for some* $\mathtt{x}_1 > 0$, *then*

$$I\!P\left(\bigcup_{k=1}^K \left\{ \widetilde{\boldsymbol{\theta}}_k \notin \Theta_{0,k}(\mathtt{r}_{0,k}) \right\} \right) \leq 3\mathrm{e}^{-\mathtt{x}_1}.$$

2. *Let conditions of Theorem B.5 be fulfilled and* (B.9) *hold for each* $k = 1, \ldots, K$ *with* $\mathtt{x} = \mathtt{x}_1 + \log(K)$ *for some* $\mathtt{x}_1 > 0$, *then it holds with* $I\!P$*-probability* $\geq 1 - 3\mathrm{e}^{-\mathtt{x}_1}$

$$I\!P^\circ\left(\bigcup_{k=1}^K \left\{ \widetilde{\boldsymbol{\theta}}_k^\circ \notin \Theta_{0,k}(\mathtt{r}_{0,k}) \right\} \right) \leq 3\mathrm{e}^{-\mathtt{x}_1}.$$

Lemma B.9 (Simultaneous Wilks approximations).

1. *Let the conditions of part 1 of Lemma B.8 be fulfilled for some* $\mathtt{r}_{0,k} > 0$ *and* $\mathtt{x} = \mathtt{x}_1 + \log(K)$, *then it holds*

$$I\!P\left(\bigcap_{k=1}^K \left\{ \left| 2\left\{ L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*) \right\} - \|\boldsymbol{\xi}_k\|^2 \right| \leq \Delta_{k,\mathrm{W}^2}(\mathtt{r}_{0,k}, \mathtt{x}_1 + \log(K)) \right\} \right) \geq 1 - 5\mathrm{e}^{-\mathtt{x}_1},$$

$$I\!P\left(\bigcap_{k=1}^K \left\{ \left| \sqrt{2\left\{ L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*) \right\}} - \|\boldsymbol{\xi}_k\| \right| \leq \Delta_{k,\mathrm{W}}(\mathtt{r}_{0,k}, \mathtt{x}_1 + \log(K)) \right\} \right) \geq 1 - 5\mathrm{e}^{-\mathtt{x}_1}.$$

2. *Let the conditions of parts 1,2 of Lemma B.8 be fulfilled for some* $\mathtt{r}_{0,k} > 0$ *and* $\mathtt{x} = \mathtt{x}_1 + \log(K)$*, then it holds with* $I\!P$*-probability* $\geq 1 - 5\mathrm{e}^{-\mathtt{x}_1}$

$$I\!P^{\circ}\left(\bigcap_{k=1}^{K}\left\{\left|\sup_{\boldsymbol{\theta}\in\Theta_k} 2\Big\{L_k^{\circ}(\boldsymbol{\theta}) - L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)\Big\} - \|\boldsymbol{\xi}_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)\|^2\right| \leq \Delta_{k,\mathrm{W}^2}^{\circ}(\mathtt{r}_{0,k}, \mathtt{x}_1 + \log(K))\right\}\right) \geq 1 - 4\mathrm{e}^{-\mathtt{x}_1},$$

$$I\!P^{\circ}\left(\bigcap_{k=1}^{K}\left\{\left|\sqrt{\sup_{\boldsymbol{\theta}\in\Theta_k} 2\Big\{L_k^{\circ}(\boldsymbol{\theta}) - L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)\Big\}} - \|\boldsymbol{\xi}_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)\|\right| \leq \Delta_{k,\mathrm{W}}^{\circ}(\mathtt{r}_{0,k}, \mathtt{x}_1 + \log(K))\right\}\right) \geq 1 - 4\mathrm{e}^{-\mathtt{x}_1}.$$

Lemma B.10. *Let the conditions of Lemma B.6 be fulfilled, then it holds with* $I\!P$*-probability* $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$I\!P^{\circ}\left(\bigcap_{k=1}^{K}\left\{\sup_{\substack{\boldsymbol{\theta}\in\Theta_{0,k}(\mathtt{r}),\\ \mathtt{r}\leq\mathtt{r}_{0,k}}} \|\boldsymbol{\xi}_k^{\circ}(\boldsymbol{\theta}) - \boldsymbol{\xi}_k^{\circ}(\boldsymbol{\theta}_k^*)\| \leq \Delta_{\xi,k}^{\circ}(\mathtt{r}, \mathtt{x} + \log(K))\right\}\right) \geq 1 - \mathrm{e}^{-\mathtt{x}}.$$

C Proofs of the main results

Before proving the statements from Section 3.2 we formulate below the Bernstein matrix inequality, which is necessary for the further proofs.

C.1 Bernstein matrix inequality

Here we restate the Theorem 1.4 by Tropp (2012) for the random $p_{\mathrm{sum}} \times p_{\mathrm{sum}}$ matrix $\widehat{\mathcal{V}}^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}^{\circ}\big(\nabla_{\boldsymbol{\theta}} L_1^{\circ}(\boldsymbol{\theta}_1^*)^{\top}, \ldots, \nabla_{\boldsymbol{\theta}} L_K^{\circ}(\boldsymbol{\theta}_K^*)^{\top}\big)^{\top}$ from the bootstrap world. Matrix $\widehat{\mathcal{V}}^2$ equals to the sum of independent matrices $\mathrm{Var}^{\circ}\big(\nabla_{\boldsymbol{\theta}}\ell_{i,1}(\boldsymbol{\theta}_1^*)^{\top}u_i, \ldots, \nabla_{\boldsymbol{\theta}}\ell_{i,K}(\boldsymbol{\theta}_K^*)^{\top}u_i\big)^{\top}$. Let us denote

$$\begin{aligned}
\boldsymbol{g}_i &\stackrel{\mathrm{def}}{=} \Big(\nabla_{\boldsymbol{\theta}}\ell_{i,1}(\boldsymbol{\theta}_1^*)^{\top}, \ldots, \nabla_{\boldsymbol{\theta}}\ell_{i,K}(\boldsymbol{\theta}_K^*)^{\top}\Big)^{\top} \in I\!R^{p_{\mathrm{sum}}},\\
\widehat{H}^2 &\stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^{n} I\!E\left\{\boldsymbol{g}_i\boldsymbol{g}_i^{\top}\right\},\\
\widehat{v}_i &\stackrel{\mathrm{def}}{=} \widehat{H}^{-1}\left\{\boldsymbol{g}_i\boldsymbol{g}_i^{\top} - I\!E\left[\boldsymbol{g}_i\boldsymbol{g}_i^{\top}\right]\right\}\widehat{H}^{-1},
\end{aligned}$$

then

$$\widehat{H}^2 = I\!E\widehat{\mathcal{V}}^2, \qquad \sum\nolimits_{i=1}^{n}\widehat{v}_i^2 = \widehat{H}^{-1}\widehat{\mathcal{V}}^2\widehat{H}^{-1} - \boldsymbol{I}_{p_{\mathrm{sum}}}.$$

Define also the deterministic scalar value

$$\widehat{\varkappa}_v^2 \stackrel{\mathrm{def}}{=} \Big\|\sum\nolimits_{i=1}^{n} I\!E\widehat{v}_i^4\Big\|.$$

Theorem C.1 (Bernstein inequality for $\widehat{\mathcal{V}}^2$)**.** *Let the condition* **($\widehat{SD_1}$)** *be fulfilled, then it holds with probability* $\geq 1 - \mathrm{e}^{-\mathtt{x}}$ *:*

$$\|\widehat{H}^{-1}\widehat{\mathcal{V}}^2\widehat{H}^{-1} - \boldsymbol{I}_{p_{\mathrm{sum}}}\| \leq \delta^2_{\widehat{\mathcal{V}}}(\mathtt{x}),$$

where the error term is defined as

$$\delta^2_{\widehat{\mathcal{V}}}(\mathtt{x}) \stackrel{\mathrm{def}}{=} \sqrt{2\widehat{\varkappa}^2_v\{\log(p_{\mathrm{sum}}) + \mathtt{x}\}} + \frac{2}{3}\delta^2_{v^*}\{\log(p_{\mathrm{sum}}) + \mathtt{x}\} \tag{C.1}$$

and is proportional to $\sqrt{\{\log(p_{\mathrm{sum}}) + \mathtt{x}\}/n}$ *in the case 5.3.*

We omit here the proof of Theorem C.1, since it follows straightforwardly from Theorem 1.4 by Tropp (2012), and is already given in Spokoiny and Zhilova (2014).

C.2 Bootstrap validity for the case of one parametric model

Here we state the results on bootstrap validity from Spokoiny and Zhilova (2014), they will be used for some of the further proofs.

Theorem C.2. *Let the conditions of Section 5 be fulfilled, then it holds for each* $k = 1, \ldots, K$ *,* $z_k \geq \max\{2, \sqrt{p_k}\} + \mathtt{C}(p_k + \mathtt{x})/\sqrt{n}$ *with probability* $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$ *:*

$$\left|I\!\!P\left(L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}^*_k) > z_k^2/2\right) - I\!\!P^\circ\left(L^\circ_k(\widetilde{\boldsymbol{\theta}}^\circ_k) - L^\circ_k(\widetilde{\boldsymbol{\theta}}_k) > z_k^2/2\right)\right| \leq \Delta_{\mathrm{full},\,k}\,.$$

The error term $\Delta_{\mathrm{full},k} \leq \mathtt{C}\{(p_k + \mathtt{x})^3/n\}^{1/8}$ *in the case of i.i.d. model; see Section 5.3.*

Theorem C.3 (Validity of the bootstrap under a small modeling bias)**.** *Assume the conditions of Theorem C.2. Then for* $\alpha \leq 1 - 8\mathrm{e}^{-\mathtt{x}}$ *, it holds*

$$\left|I\!\!P\left(L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}^*_k) > (\mathfrak{z}^\circ_k(\alpha))^2/2\right) - \alpha\right| \leq \Delta_{\mathfrak{z},\,\mathrm{full},\,k}\,.$$

The error term $\Delta_{\mathfrak{z},\,\mathrm{full},\,k} \leq \mathtt{C}\{(p_k + \mathtt{x})^3/n\}^{1/8}$ *in the case of i.i.d. model; see Section 5.3.*

Theorem C.4 (Performance of the bootstrap for a large modeling bias)**.** *Under the conditions of Section 5 except for* **($\widehat{\mathrm{SmB}}$)** *it holds for* $z_k \geq \max\{2, \sqrt{p_k}\} + \mathtt{C}(p_k + \mathtt{x})/\sqrt{n}$ *with probability* $\geq 1 - 14\mathrm{e}^{-\mathtt{x}}$

1. $I\!\!P\left(L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}^*_k) > z_k^2/2\right) \leq I\!\!P^\circ\left(L^\circ_k(\widetilde{\boldsymbol{\theta}}^\circ_k) - L^\circ(\widetilde{\boldsymbol{\theta}}_k) > z_k^2/2\right) + \Delta_{\mathrm{b},\,\mathrm{full},\,k}.$

2. $$\mathfrak{z}^\circ_k(\alpha) \geq \mathfrak{z}_k(\alpha + \Delta_{\mathrm{b},\,\mathrm{full},\,k}) + \sqrt{\mathrm{tr}\{D_k^{-1}H_k^2D_k^{-1}\}} - \sqrt{\mathrm{tr}\{D_k^{-1}(H_k^2 - B_k^2)D_k^{-1}\}} - \Delta_{\mathtt{qf},1,k},$$

   $$\mathfrak{z}^\circ_k(\alpha) \leq \mathfrak{z}_k(\alpha - \Delta_{\mathrm{b},\,\mathrm{full},\,k}) + \sqrt{\mathrm{tr}\{D_k^{-1}H_k^2D_k^{-1}\}} - \sqrt{\mathrm{tr}\{D_k^{-1}(H_k^2 - B_k^2)D_k^{-1}\}} + \Delta_{\mathtt{qf},2,k}.$$

The term $\Delta_{\mathrm{b,full},k} \le \mathtt{C}\{(p_k+\mathtt{x})^3/n\}^{1/8}$ *in the case of i.i.d. model; see Section 5.3. The positive values* $\Delta_{\mathtt{qf},1,k}, \Delta_{\mathtt{qf},2,k}$ *are bounded from above with* $(\mathfrak{a}_k^2+\mathfrak{a}_{B,k}^2)(\sqrt{8\mathtt{x}p_k}+6\mathtt{x})$ *for the constants* $\mathfrak{a}_k^2>0$, $\mathfrak{a}_{B,k}^2\ge 0$ *from conditions* $(\boldsymbol{\mathcal{I}})$, $(\boldsymbol{\mathcal{I}_B})$.

C.3 Proof of Theorem 3.1

Lemma C.5 (Closeness of $\mathcal{L}(\|\boldsymbol{\xi}_1\|,\dots,\|\boldsymbol{\xi}_K\|)$ and $\mathcal{L}^\circ(\|\boldsymbol{\xi}_1^\circ\|,\dots,\|\boldsymbol{\xi}_K^\circ\|)$)**.** *If the conditions* $(\boldsymbol{ED_0})$, $(\boldsymbol{\mathcal{I}})$, $(\boldsymbol{\widehat{SmB}})$, $(\boldsymbol{\mathcal{I}_B})$, $(\boldsymbol{\widehat{SD_1}})$ *and* $(\boldsymbol{Eb})$ *are fulfilled, then it holds with probability* $\ge 1-6\mathrm{e}^{-\mathtt{x}}$ *for all* $\delta_{z_k}\ge 0$ *and* $z_k \ge \sqrt{p_k}+\Delta_\varepsilon$ *s.t.* $\mathtt{C}\max\limits_{1\le k\le K}\{n^{-1/2},\delta_{z_k}\} \le \Delta_\varepsilon \le \mathtt{C}\min\limits_{1\le k\le K}\{1/z_k\}$ *(Δ_ε is given in (A.3)):*

$$I\!\!P\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\xi}_k\|>z_k\}\right) - I\!\!P^\circ\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\xi}_k^\circ\|>z_k-\delta_{z_k}\}\right) \ge -\Delta_{\ell_2},$$

$$I\!\!P\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\xi}_k\|>z_k\}\right) - I\!\!P^\circ\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\xi}_k^\circ\|>z_k+\delta_{z_k}\}\right) \le \Delta_{\ell_2}.$$

for the deterministic nonnegative value

$$\Delta_{\ell_2} \le 25\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8}\log^{9/8}(K)\log^{3/8}(np_{\mathrm{sum}})\left\{\left(\widehat{\mathfrak{a}}^2+\widehat{\mathfrak{a}}_B^2\right)\left(1+\delta_{\hat{\nu}}^2(\mathtt{x})\right)\right\}^{3/8}.$$

A more explicit bound on Δ_{ℓ_2} *is given in Proposition A.1, see also Remark A.1.*

Proof of Lemma C.5. The statement follows from Proposition A.1 and Theorem C.1. Let us take $\boldsymbol{\phi}_k := \boldsymbol{\xi}_k$ and $\boldsymbol{\psi}_k := \boldsymbol{\xi}_k^\circ$. Define similarly to Φ in (A.5)

$$\Xi \stackrel{\mathrm{def}}{=} \left(\boldsymbol{\xi}_1^\top,\dots,\boldsymbol{\xi}_K^\top\right)^\top \quad \Xi^\circ \stackrel{\mathrm{def}}{=} \left(\boldsymbol{\xi}_1^{\circ\top},\dots,\boldsymbol{\xi}_K^{\circ\top}\right)^\top. \tag{C.2}$$

Condition (A.4) rewrites for (C.2) as

$$\|\mathrm{Var}\,\Xi - \mathrm{Var}^\circ\,\Xi^\circ\|_{\max} \le \delta_\Sigma^2$$

for some $\delta_\Sigma^2 \ge 0$. Denote

$$\widehat{D}^2 \stackrel{\mathrm{def}}{=} \mathrm{diag}\left\{D_1^2,\dots,D_K^2\right\},$$

$$\widehat{V}^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}\left(\nabla_{\boldsymbol{\theta}}L_1(\boldsymbol{\theta}_1^*)^\top,\dots,\nabla_{\boldsymbol{\theta}}L_K(\boldsymbol{\theta}_K^*)^\top\right)^\top.$$

$\widehat{D}^2$ is a block-diagonal matrix and $\widehat{V}^2$ is a block matrix. Both of them are symmetric, positive definite and have the dimension $p_{\mathrm{sum}}\times p_{\mathrm{sum}}$. Let also

$$\widehat{\mathcal{V}}^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}^\circ\left(\nabla_{\boldsymbol{\theta}}L_1^\circ(\boldsymbol{\theta}_1^*)^\top,\dots,\nabla_{\boldsymbol{\theta}}L_K^\circ(\boldsymbol{\theta}_K^*)^\top\right)^\top,$$

$$\boldsymbol{g}_i \stackrel{\mathrm{def}}{=} \left(\nabla_{\boldsymbol{\theta}}\ell_{i,1}(\boldsymbol{\theta}_1^*)^\top,\dots,\nabla_{\boldsymbol{\theta}}\ell_{i,K}(\boldsymbol{\theta}_K^*)^\top\right)^\top \in I\!R^{p_{\mathrm{sum}}},$$

$$\widehat{H}^2 \stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n I\!\!E\left\{\boldsymbol{g}_i\boldsymbol{g}_i^\top\right\}, \quad \widehat{B}^2 \stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n I\!\!E\left\{\boldsymbol{g}_i\right\}I\!\!E\left\{\boldsymbol{g}_i\right\}^\top.$$

It holds

$$\operatorname{Var}\Xi = \widehat{D}^{-1}\widehat{V}^2\widehat{D}^{-1}, \quad \operatorname{Var}^\circ\Xi^\circ = \widehat{D}^{-1}\widehat{\mathcal{V}}^2\widehat{D}^{-1},$$
$$\widehat{H}^2 = I\!\!E\widehat{\mathcal{V}}^2, \quad \widehat{V}^2 = \widehat{H}^2 - \widehat{B}^2.$$

Therefore

$$\begin{aligned}
\|\operatorname{Var}\Xi - \operatorname{Var}^\circ\Xi^\circ\|_{\max} &= \big\|\widehat{D}^{-1}\big(\widehat{V}^2 - \widehat{\mathcal{V}}^2\big)\widehat{D}^{-1}\big\|_{\max} \\
&\le \big\|\widehat{D}^{-1}\big(\widehat{H}^2 - \widehat{\mathcal{V}}^2\big)\widehat{D}^{-1}\big\|_{\max} + \big\|\widehat{D}^{-1}\widehat{B}^2\widehat{D}^{-1}\big\|_{\max} \\
&\le \delta^2_{\widehat{\mathcal{V}}}(\mathtt{x})\big\|\widehat{D}^{-1}\widehat{H}^2\widehat{D}^{-1}\big\| + \big\|\widehat{D}^{-1}\widehat{B}^2\widehat{D}^{-1}\big\| \qquad \text{(C.3)} \\
&\le \big\{\delta^2_{\widehat{\mathcal{V}}}(\mathtt{x}) + \widehat{\delta}^2_{\mathrm{smb}}\big\}(\widehat{\mathfrak{a}}^2 + \widehat{\mathfrak{a}}^2_B) =: \delta^2_\Sigma\,. \qquad \text{(C.4)}
\end{aligned}$$

Here inequality (C.3) follows from the matrix Bernstein inequality by Tropp (2012) (see Section C.1). Inequality (C.4) is implied by conditions $(\boldsymbol{\mathcal{I}_B})$ and $(\widehat{\mathbf{SmB}})$, and Cauchy-Schwarz inequality.

Condition (C1) of Proposition A.1 is fulfilled for the vectors $\boldsymbol{\xi}_{i,k}$ and $\boldsymbol{\xi}^\circ_{i,k}$ due to conditions $(\boldsymbol{ED_0})$, $(\boldsymbol{\mathcal{I}})$ and $(\widehat{\boldsymbol{SD_1}})$, $(\boldsymbol{Eb})$, $(\widehat{\mathbf{SmB}})$, $(\boldsymbol{\mathcal{I}_B})$ for $\mathfrak{c}_\phi := \widehat{\mathfrak{a}}$ and $\mathfrak{c}^2_\psi := \big(\widehat{\mathfrak{a}}^2 + \widehat{\mathfrak{a}}^2_B\big)\Big\{\delta^2_{v^*} + \max_{1\le i\le n}\|\widehat{H}^{-1}I\!\!E\left[\boldsymbol{g}_i\boldsymbol{g}_i^\top\right]\widehat{H}^{-1}\|^2\Big\}$. □

Proof of Theorem 3.1. Let us denote $\mathtt{x}_2 \stackrel{\mathrm{def}}{=} \mathtt{x} + \log(K)$. It holds with probability $\ge 1 - 12\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
& I\!\!P^\circ\left(\bigcup\nolimits_{k=1}^K\Big\{\sqrt{2L^\circ_k(\widetilde{\boldsymbol{\theta}}^\circ_k) - 2L^\circ_k(\widetilde{\boldsymbol{\theta}}_k)} > z_k\Big\}\right) \\
\overset{\text{L. B.9}}{\ge}\ & I\!\!P^\circ\left(\bigcup\nolimits_{k=1}^K\Big\{\|\boldsymbol{\xi}^\circ_k(\widetilde{\boldsymbol{\theta}}_k)\| \ge z_k + \Delta^\circ_{\mathrm{w},k}(\mathtt{r}_{0,k},\mathtt{x}_2)\Big\}\right) \\
\overset{\text{L. B.10}}{\ge}\ & I\!\!P^\circ\left(\bigcup\nolimits_{k=1}^K\Big\{\|\boldsymbol{\xi}^\circ_k(\boldsymbol{\theta}^*_k)\| > z_k + \Delta^\circ_{\mathrm{w},k}(\mathtt{r}_{0,k},\mathtt{x}_2) + \Delta^\circ_{\xi,k}(\mathtt{r}_{0,k},\mathtt{x}_2)\Big\}\right) \\
\overset{\text{L. C.5}}{\ge}\ & I\!\!P\left(\bigcup\nolimits_{k=1}^K\Big\{\|\boldsymbol{\xi}_k\| > z_k - \Delta_{\mathrm{w},k}(\mathtt{r}_{0,k},\mathtt{x}_2)\Big\}\right) - \Delta_{\mathrm{total}} \\
\overset{\text{L. B.9}}{\ge}\ & I\!\!P\left(\bigcup\nolimits_{k=1}^K\Big\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}^*_k)} > z_k\Big\}\right) - \Delta_{\mathrm{total}},
\end{aligned}$$

for

$$\Delta_{\mathrm{total}} \stackrel{\mathrm{def}}{=} \Delta_{\ell_2}, \qquad \text{(C.5)}$$

$$\begin{aligned}
\delta_{z_k} &:= \Delta_{\mathrm{w},k}(\mathtt{r}_{0,k},\mathtt{x}+\log(K)) + \Delta^\circ_{\mathrm{w},k}(\mathtt{r}_{0,k},\mathtt{x}+\log(K)) \qquad \text{(C.6)} \\
&\quad + \Delta^\circ_{\xi,k}(\mathtt{r}_{0,k},\mathtt{x}+\log(K)) \\
&\le \mathtt{C}\frac{p_k + \mathtt{x} + \log(K)}{\sqrt{n}}\sqrt{\mathtt{x}+\log(K)} \quad \text{in the case 5.3.} \qquad \text{(C.7)}
\end{aligned}$$

Definition of Δ_{ℓ_2} is given in Proposition A.1, see also Remark A.1. The bound from Lemma C.5 says:

$$\Delta_{\ell_2} \leq 25\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8} \log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\left\{\left(\widehat{\mathfrak{a}}^2+\widehat{\mathfrak{a}}_B^2\right)\left(1+\delta_{\widehat{\nu}}^2(\mathtt{x})\right)\right\}^{3/8}.$$

For δ_{z_k} bounded as in (C.7) the conditions $\mathtt{C}\max\limits_{1\leq k\leq K}\{n^{-1/2},\delta_{z_k}\} \leq \Delta_\varepsilon \leq \mathtt{C}\min\limits_{1\leq k\leq K}\{1/z_k\}$ are fulfilled. □

C.4 Proof of Theorem 3.2

Proof of Theorem 3.2. For the pointwise quantile functions $\mathfrak{z}_k(\alpha)$ and $\mathfrak{z}_k^\circ(\alpha)$ it holds for each $k=1,\ldots,K$ with dominating probability:

$$\begin{aligned}\mathfrak{z}_k^\circ\left(\alpha+\Delta_{\text{full},\,k}\right) &\leq \mathfrak{z}_k\left(\alpha\right),\\ \mathfrak{z}_k^\circ\left(\alpha\right) &\geq \mathfrak{z}_k\left(\alpha+\Delta_{\text{full }k}\right)-\varepsilon_k\end{aligned} \tag{C.8}$$

here $\Delta_{\text{full},\,k} \leq \left\{(p_k+\mathtt{x})^3/\sqrt{n}\right\}^{1/8}$, it comes from Theorem C.2, and $\varepsilon_k \leq \mathtt{C}(p_k+\mathtt{x})/\sqrt{n}$,

$$\varepsilon_k \stackrel{\text{def}}{=} \begin{cases}0, & \text{if c.d.f. of } L_k(\widetilde{\boldsymbol{\theta}}_k)-L_k(\boldsymbol{\theta}_k^*) \text{ is continuous in } \mathfrak{z}_k(\alpha+\Delta_{\text{full},\,k});\\ \mathtt{C}(p_k+\mathtt{x})/\sqrt{n} \text{ s.t. (C.9) is fulfilled}, & \text{otherwise.}\end{cases}$$

$$I\!P\left(\sqrt{2\left\{L_k(\widetilde{\boldsymbol{\theta}}_k)-L_k(\boldsymbol{\theta}_k^*)\right\}} > \mathfrak{z}_k(\alpha+\Delta_{\text{full},\,k})-\varepsilon_k\right) \geq \alpha+\Delta_{\text{full},\,k}. \tag{C.9}$$

Indeed, due to Theorem C.2 and definition (1.5)

$$\begin{aligned}&I\!P^\circ\left(\sqrt{2\left\{L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ)-L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)\right\}} > \mathfrak{z}_k(\alpha)\right)\\ &\qquad\leq I\!P\left(\sqrt{2\left\{L_k(\widetilde{\boldsymbol{\theta}}_k)-L_k(\boldsymbol{\theta}_k^*)\right\}} > \mathfrak{z}_k(\alpha)\right)+\Delta_{\text{full},\,k} \leq \alpha+\Delta_{\text{full},\,k},\end{aligned}$$

therefore, by definition (2.3) $\mathfrak{z}_k^\circ(\alpha+\Delta_{\text{full},\,k}) \leq \mathfrak{z}_k(\alpha)$. The lower bound is derived similarly.

If there exist the inverse functions $\mathfrak{c}^{-1}(\cdot)$ and $\mathfrak{c}^{\circ\,-1}(\cdot)$, then it holds for $\beta\in(0,1)$:

$$\begin{aligned}I\!P\left(\bigcup\nolimits_{k=1}^K\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k)-2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k(\beta)\right\}\right) &\leq \mathfrak{c}^{-1}(\beta),\\ I\!P^\circ\left(\bigcup\nolimits_{k=1}^K\left\{\sqrt{2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ)-2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)} \geq \mathfrak{z}_k^\circ\left(\beta\right)\right\}\right) &\leq \mathfrak{c}^{\circ\,-1}(\beta).\end{aligned} \tag{C.10}$$

Therefore, it holds

$$
\begin{aligned}
& \mathfrak{c}^{\circ\,-1}(\beta + \Delta_{\text{full, max}}) \\
\geq \quad & I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} \geq \mathfrak{z}_k^{\circ}\left(\beta + \Delta_{\text{full},\,k}\right)\right\}\right) \\
\overset{\text{by (C.8)}}{\geq} \quad & I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} \geq \mathfrak{z}_k\left(\beta\right)\right\}\right) \\
\overset{\text{by Th. 3.1}}{\geq} \quad & I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k\left(\beta\right)\right\}\right) - \Delta_{\text{total}} \\
\overset{\text{by L. C.6 and (C.10)}}{\geq} \quad & \mathfrak{c}^{-1}(\beta) - \Delta_{\text{total}} - \Delta_{\text{ac,LR}},
\end{aligned}
$$

here $\Delta_{\text{ac,LR}} \leq \Delta_{\text{total}}$ (by Lemma C.6) and

$$
\begin{aligned}
\Delta_{\text{full, max}} & \overset{\text{def}}{=} \max_{1 \leq k \leq K} \Delta_{\text{full},\,k} \qquad (C.11) \\
& \leq \mathtt{C}\{(p_{\max} + \mathtt{x})^3/n\}^{1/8} \text{ in the case 5.3.}
\end{aligned}
$$

Thus

$$
\begin{aligned}
\mathfrak{c}^{\circ\,-1}(\beta + \Delta_{\text{full, max}}) & \geq \mathfrak{c}^{-1}(\beta) - \Delta_{\text{total}} - \Delta_{\text{ac,LR}}, \\
\mathfrak{c}^{\circ}(\alpha) & \leq \mathfrak{c}(\alpha + \Delta_{\text{total}} + \Delta_{\text{ac,LR}}) + \Delta_{\text{full, max}}. \qquad (C.12)
\end{aligned}
$$

Hence it holds

$$
\begin{aligned}
& I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^{\circ}(\beta)\right\}\right) \\
\overset{\text{by (C.8)}}{\leq} \quad & I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k\left(\beta + \Delta_{\text{full},\,k}\right) - \varepsilon_k\right\}\right) \\
\overset{\text{by L. C.6 and (C.10)}}{\leq} \quad & \mathfrak{c}^{-1}(\beta + \Delta_{\text{full, max}}) + \Delta_{\text{ac,LR}}.
\end{aligned}
$$

Therefore, if $\mathfrak{c}(\alpha) \geq \Delta_{\text{full, max}}$, then

$$
I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^{\circ}(\mathfrak{c}(\alpha) - \Delta_{\text{full, max}})\right\}\right) \leq \alpha + \Delta_{\text{ac,LR}}.
$$

And by (C.12) for $\mathfrak{c}^{\circ}(\alpha) \geq 2\Delta_{\text{full, max}}$ it holds

$$
\begin{aligned}
& I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^{\circ}\left(\mathfrak{c}^{\circ}(\alpha) - 2\Delta_{\text{full, max}}\right)\right\}\right) - \alpha \\
& \leq \Delta_{\text{total}} + 2\Delta_{\text{ac,LR}}.
\end{aligned}
$$

Similarly for the inverse direction:

$$\begin{aligned}
\mathfrak{c}^{\circ\,-1}(\beta) &\le I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ})-2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)}\ge \mathfrak{z}_k^{\circ}(\beta)\right\}-\varepsilon_{1,k}\right)\\
&\le I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ})-2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)}\ge \mathfrak{z}_k(\beta+\Delta_{\text{full},k})-\varepsilon_{1,k}-\varepsilon_k\right\}\right)\\
&\le I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k)-2L_k(\boldsymbol{\theta}_k^*)}\ge \mathfrak{z}_k(\beta+\Delta_{\text{full},k})\right\}\right)+\Delta_{\text{total}}+\Delta_{\text{ac,LR}}\\
&\le \mathfrak{c}^{-1}(\beta+\Delta_{\text{full,max}})+\Delta_{\text{total}}+\Delta_{\text{ac,LR}},
\end{aligned}$$

where $0\le\varepsilon_{1,k}\le \mathtt{C}(p_k+\mathtt{x})/\sqrt{n}$. This implies

$$\begin{aligned}
\mathfrak{c}^{\circ\,-1}(\beta) &\le \mathfrak{c}^{-1}(\beta+\Delta_{\text{full,max}})+\Delta_{\text{total}}+\Delta_{\text{ac,LR}},\\
\mathfrak{c}^{\circ}(\alpha) &\ge \mathfrak{c}(\alpha-\Delta_{\text{total}}-\Delta_{\text{ac,LR}})-\Delta_{\text{full,max}}.
\end{aligned}\tag{C.13}$$

$$\begin{aligned}
& I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k)-2L_k(\boldsymbol{\theta}_k^*)}\ge \mathfrak{z}_k^{\circ}(\beta+\Delta_{\text{full},k})\right\}\right)\\
\overset{\text{by (C.8)}}{\ge}\ & I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k)-2L_k(\boldsymbol{\theta}_k^*)}\ge \mathfrak{z}_k(\beta)\right\}\right)\\
\ge\ & \mathfrak{c}^{-1}(\beta)-\Delta_{\text{ac,LR}}.
\end{aligned}$$

$$I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k)-2L_k(\boldsymbol{\theta}_k^*)}\ge \mathfrak{z}_k^{\circ}(\mathfrak{c}(\alpha)+\Delta_{\text{full,max}})\right\}\right)\ge\alpha-\Delta_{\text{ac,LR}}.$$

And by (C.13)

$$\begin{aligned}
& I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k)-2L_k(\boldsymbol{\theta}_k^*)}\ge \mathfrak{z}_k^{\circ}(\mathfrak{c}^{\circ}(\alpha)+2\Delta_{\text{full,max}})\right\}\right)-\alpha\\
&\ge -\Delta_{\text{total}}-2\Delta_{\text{ac,LR}}.
\end{aligned}$$

for

$$\Delta_{\mathfrak{z},\,\text{total}}\overset{\text{def}}{=}\Delta_{\text{total}}+2\Delta_{\text{ac,LR}}\le 3\Delta_{\text{total}}.\tag{C.14}$$

Conditions of Theorem 3.1 include $z_k\ge C\sqrt{p_k}$, therefore, it has to be checked that $\mathfrak{z}_k^{\circ}(\alpha)\ge C\sqrt{p_k}$. It holds by Theorem B.4, Proposition A.1, Lemmas B.6 and C.7 with probability $\ge 1-12\mathrm{e}^{-\mathtt{x}}$:

$$\begin{aligned}
& I\!P^{\circ}\left(\sqrt{2\{L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ})-L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)\}}>\mathtt{C}\sqrt{p_k-\sqrt{2\mathtt{x}p_k}}+\mathtt{C}(p_k+\mathtt{x})/\sqrt{n}\right)\\
&\qquad\ge 1-8\mathrm{e}^{-\mathtt{x}},
\end{aligned}$$

Taking $1 - 8\mathrm{e}^{-\mathtt{x}} \geq \alpha$, we have

$$\mathfrak{z}_k^\circ(\alpha) \geq \mathtt{C}\sqrt{p_k - \sqrt{2\mathtt{x}p_k}} + \mathtt{C}2(p_k + \mathtt{x})/\sqrt{n}.$$

Inequalities for $\mathfrak{c}^\circ(\alpha)$ had been already derived in (C.12) and (C.13) with

$$\Delta_{\mathfrak{c}} \stackrel{\text{def}}{=} \Delta_{\text{total}} + \Delta_{\text{ac,LR}}. \tag{C.15}$$

□

Lemma C.6. *Let the conditions from Section 5.1 be fulfilled, and the values $z_k \geq \sqrt{p_k}$ and $\delta_{z_k} \geq 0$ be s.t. $\mathtt{C} \max\limits_{1\leq k\leq K}\{n^{-1/2}, \delta_{z_k}\} \leq \Delta_\varepsilon \leq \mathtt{C} \min\limits_{1\leq k\leq K}\{1/z_k\}$ (Δ_ε is given in (A.3)), then it holds with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$*

$$\begin{aligned}
&I\!\!P\left(\bigcup\nolimits_{k=1}^K \Big\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq z_k\Big\}\right) \\
&\quad - I\!\!P\left(\bigcup\nolimits_{k=1}^K \Big\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq z_k + \delta_{z_k}\Big\}\right) \leq \Delta_{\text{ac,LR}},
\end{aligned}$$

where

$$\Delta_{\text{ac,LR}} \leq 12.5\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8} \log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\widehat{\mathfrak{a}}^{3/4}.$$

Proof of Lemma C.6. This statement's proof is similar to the one of Theorem 3.1 (see Section C.3). Here instead of the bootstrap statistics we consider only the values from the $\boldsymbol{Y}$-world. Let us denote $\mathtt{x}_2 \stackrel{\text{def}}{=} \mathtt{x} + \log(K)$. It holds with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
& I\!\!P\left(\bigcup\nolimits_{k=1}^K \Big\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > z_k\Big\}\right) \\
\stackrel{\text{L. B.9}}{\leq}\ & I\!\!P\left(\bigcup\nolimits_{k=1}^K \Big\{\|\boldsymbol{\xi}_k\| > z_k - \Delta_{\text{w},k}(\mathtt{r}_{0,k}, \mathtt{x}_2)\Big\}\right) \\
\stackrel{\text{Pr. A.1}}{\leq}\ & I\!\!P\left(\bigcup\nolimits_{k=1}^K \Big\{\|\boldsymbol{\xi}_k\| > z_k + \delta_{z_k} + \Delta_{\text{w},k}(\mathtt{r}_{0,k}, \mathtt{x}_2)\Big\}\right) + \Delta_{\text{ac,LR}} \\
\leq\ & I\!\!P\left(\bigcup\nolimits_{k=1}^K \Big\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > z_k + \delta_{z_k}\Big\}\right) + \Delta_{\text{ac,LR}},
\end{aligned}$$

where

$$\Delta_{\text{ac,LR}} \leq 12.5\mathtt{C}\left(p_{\max}^3/n\right)^{1/8} \log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\widehat{\mathfrak{a}}^{3/4}.$$

Similarly to (C.5) and (C.6) the term $\Delta_{\text{ac,LR}}$ is equal to Δ_{ℓ_2} from Proposition A.1 with $\Delta_\Sigma^2 := 0$, $\delta_{z_k} := \delta_{z_k} + 2\Delta_{\text{w},k}(\mathtt{r}_{0,k}, \mathtt{x} + \log(K))$. □

Lemma C.7 (Lower bound for deviations of a Gaussian quadratic form)**.** *Let* $\boldsymbol{\phi} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ *and* Σ *is any symmetric non-negative definite matrix, then it holds for any* $\mathtt{x} > 0$

$$\mathbb{P}\left(\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \geq 2\sqrt{\mathtt{x}\operatorname{tr}(\Sigma^2)}\right) \leq \exp(-\mathtt{x}).$$

Proof of Lemma C.7. It is sufficient to consider w.l.o.g. only the case of diagonal matrix Σ, since it can be represented as $\Sigma = U^\top \operatorname{diag}\{a_1, \ldots, a_p\}U$ for an orthogonal matrix U and the eigenvalues $a_1 \geq \cdots \geq a_p$; $U\boldsymbol{\phi} \sim \mathcal{N}(0, \boldsymbol{I}_p)$.

By the exponential Chebyshev inequality it holds for $\mu > 0$, $\Delta > 0$

$$\mathbb{P}\left(\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \geq \Delta\right) \leq \exp(-\mu\Delta/2)\mathbb{E}\exp\left(\mu\left\{\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2\right\}/2\right).$$

$$\log\mathbb{E}\exp\left(\mu\left\{\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2\right\}/2\right) \leq \frac{1}{2}\sum_{j=1}^{p}\left\{\mu a_j - \log(1+a_j\mu)\right\},$$

therefore

$$\begin{aligned}
\mathbb{P}\left(\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \geq \Delta\right) &\leq \exp\left(-\frac{1}{2}\left[\mu\Delta + \textstyle\sum_{j=1}^{p}\left\{\log(1+a_j\mu) - \mu a_j\right\}\right]\right) \\
&\leq \exp\left(-\frac{1}{2}\left[\mu\Delta - \mu^2\textstyle\sum_{j=1}^{p} a_j^2/2\right]\right) \\
&\leq \exp\left(-\Delta^2/\left\{4\textstyle\sum_{j=1}^{p} a_j^2\right\}\right).
\end{aligned}$$

If $\mathtt{x} := \Delta^2/\left\{4\sum_{i=1}^{p} a_j^2\right\}$, then $\Delta = 2\sqrt{\mathtt{x}\sum_{j=1}^{p} a_j^2}$. □

C.5 Proof of Theorem 3.3

Proof of Theorem 3.3. Let us denote $\mathtt{x}_2 \stackrel{\text{def}}{=} \mathtt{x} + \log(K)$. By Lemmas B.9, B.10 and C.5 it holds with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
&\mathbb{P}^\circ\left(\bigcup_{k=1}^{K}\left\{\sqrt{2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - 2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)} > z_k\right\}\right) \\
&\geq \mathbb{P}^\circ\left(\bigcup_{k=1}^{K}\left\{\|\boldsymbol{\xi}_k^\circ(\boldsymbol{\theta}_k^*)\| > z_k + \Delta_{\mathrm{w},k}^\circ(\mathtt{r}_{0,k}, \mathtt{x}_2) + \Delta_{\xi,k}^\circ(\mathtt{r}_{0,k}, \mathtt{x}_2)\right\}\right) \\
&\geq \mathbb{P}\left(\bigcup_{k=1}^{K}\left\{\|\widetilde{\boldsymbol{\xi}}_k\| > z_k - \Delta_{\mathrm{w},k}(\mathtt{r}_{0,k}, \mathtt{x}_2)\right\}\right) - \Delta_{\mathrm{b,\,total}} \qquad \text{(C.16)} \\
&\geq \mathbb{P}\left(\bigcup_{k=1}^{K}\left\{\|\boldsymbol{\xi}_k\| > z_k - \Delta_{\mathrm{w},k}(\mathtt{r}_{0,k}, \mathtt{x}_2)\right\}\right) - \Delta_{\mathrm{b,\,total}} \qquad \text{(C.17)} \\
&\geq \mathbb{P}\left(\bigcup_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > z_k\right\}\right) - \Delta_{\mathrm{b,\,total}},
\end{aligned}$$

here $\widetilde{\boldsymbol{\xi}}_k \stackrel{\text{def}}{=} \bigl(D_k^{-1} H_k^2 D_k^{-1}\bigr)^{1/2} (\operatorname{Var} \boldsymbol{\xi}_k)^{-1/2} \boldsymbol{\xi}_k$, and $\Delta_{\text{b,total}}$ is given below. Using the same notations as in the proof of Lemma C.5, we have

$$\begin{aligned} \widetilde{\Xi} &\stackrel{\text{def}}{=} \bigl(\widetilde{\boldsymbol{\xi}}_1^\top, \ldots, \widetilde{\boldsymbol{\xi}}_K^\top\bigr)^\top \\ &= \bigl(\widehat{D}^{-1} \widehat{H}^2 \widehat{D}^{-1}\bigr)^{1/2} (\operatorname{Var} \Xi)^{-1/2} \Xi, \end{aligned}$$

and by Theorem C.1 and by conditions $(\boldsymbol{\mathcal{I}})$, $(\boldsymbol{\mathcal{I}_B})$, it holds with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned} \bigl\| \operatorname{Var} \widetilde{\Xi} - \operatorname{Var}^\circ \Xi^\circ \bigr\|_{\max} &= \bigl\| \widehat{D}^{-1} \bigl(\widehat{H}^2 - \widehat{\mathcal{V}}^2 \bigr) \widehat{D}^{-1} \bigr\|_{\max} \\ &\leq \delta_{\widehat{\mathcal{V}}}^2(\mathtt{x}) \bigl\| \widehat{D}^{-1} \widehat{H}^2 \widehat{D}^{-1} \bigr\|_{\max} \\ &\leq \delta_{\widehat{\mathcal{V}}}^2(\mathtt{x}) (\widehat{\mathfrak{a}}^2 + \widehat{\mathfrak{a}}_B^2). \end{aligned}$$

Thus, inequality (C.16) follows from Proposition A.1 applied to the sets of vectors $\boldsymbol{\xi}_1^\circ(\boldsymbol{\theta}_1^*), \ldots, \boldsymbol{\xi}_K^\circ(\boldsymbol{\theta}_K^*)$ and $\widetilde{\boldsymbol{\xi}}_1, \ldots, \widetilde{\boldsymbol{\xi}}_K$. The error term $\Delta_{\text{b,total}}$ is equal to Δ_{total} from Theorem C.3 (see (C.5), (C.6)) with $\widehat{\delta}_{\text{smb}}^2 := 0$, thus

$$\Delta_{\text{b,total}} \leq 25 \mathtt{C} \left(\frac{p_{\max}^3}{n} \right)^{1/8} \log^{9/8}(K) \log^{3/8}(n p_{\text{sum}}) \left\{ \bigl(\widehat{\mathfrak{a}}^2 + \widehat{\mathfrak{a}}_B^2 \bigr) \bigl(1 + \delta_{\widehat{\mathcal{V}}}^2(\mathtt{x}) \bigr) \right\}^{3/8}.$$

Inequality (C.17) is implied by definitions of $\widetilde{\boldsymbol{\xi}}_k$ and matrices H_k^2, V_k^2, indeed:

$$\begin{aligned} &\Bigl\| \bigl(D_k^{-1} H_k^2 D_k^{-1} \bigr)^{-1/2} \operatorname{Var} \boldsymbol{\xi}_k \bigl(D_k^{-1} H_k^2 D_k^{-1} \bigr)^{-1/2} \Bigr\| \\ &\leq \Bigl\| \bigl(D_k^{-1} H_k^2 D_k^{-1} \bigr)^{1/2} \bigl(D_k H_k^{-2} V_k^2 H_k^{-2} D_k \bigr) \bigl(D_k^{-1} H_k^2 D_k^{-1} \bigr)^{1/2} \Bigr\| \\ &\leq 1, \end{aligned}$$

therefore, $\|\widetilde{\boldsymbol{\xi}}_k\| \geq \|\boldsymbol{\xi}_k\|$.

The second inequality in the statement is proven similarly to (C.12). It implies together with Theorem C.4 the rest part of the statement having

$$\Delta_{\text{b},\mathfrak{c}} \stackrel{\text{def}}{=} \Delta_{\text{b, total}} + \Delta_{\text{ac,LR}}. \tag{C.18}$$

□

References

Arlot, S., Blanchard, G., and Roquain, E. (2010). Some nonasymptotic results on resampling in high dimension, II: Multiple tests. *The Annals of Statistics*, 38(1):83–99.

Benjamini, Y. (2010). Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal*, 52(6):708–721.

Beran, R. (1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83(403):679–686.

Beran, R. (1990). Refining bootstrap simultaneous confidence sets. *Journal of the American Statistical Association*, 85(410):417–426.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, pages 1071–1095.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilita*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Cao, H. and Kosorok, M. R. (2011). Simultaneous critical values for t-tests in very high dimensions. *Bernoulli*, 17(1):347–394.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013a). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013b). Supplement to "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors".

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014b). Central limit theorems and bootstrap in high dimensions. *arXiv preprint arXiv:1412.3661*.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014c). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, 162:47–70.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, pages 1852–1884.

Dickhaus, T. (2014). *Simultaneous Statistical Inference: With Applications in the Life Sciences*. Springer.

Hall, P. (1991). On convergence rates of suprema. *Probability Theory and Related Fields*, 89(4):447–455.

Hall, P. (1993). On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 291–304.

Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics*, 41(4):1892–1921.

Hall, P. and Pittelkow, Y. (1990). Simultaneous bootstrap confidence bands in regression. *Journal of Statistical Computation and Simulation*, 37(1-2):99–113.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis*, 29(2):163–179.

Härdle, W. and Marron, J. (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, pages 778–796.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, 51(2):186–192.

Johnston, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *Journal of Multivariate Analysis*, 12(3):402–414.

Kim, K. I. and van de Wiel, M. A. (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC bioinformatics*, 9(1):114.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723.

Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225.

Liu, W. (2010). *Simultaneous inference in regression*. CRC Press.

Liu, Y. and Wu, Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of nonparametric statistics*, 23(2):415–437.

Manly, B. F. (2006). *Randomization, bootstrap and Monte Carlo methods in biology*, volume 70. CRC Press.

Miller, R. G. (1981). *Simultaneous statistical inference*. Springer.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Neumann, M. H. and Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, 9(4):307–333.

Qu, Z. (2008). Testing for structural change in regression quantiles. *Journal of Econometrics*, 146(1):170–184.

Röllin, A. (2013). Stein's method in high dimensions with applications. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 49, pages 529–549. Institut Henri Poincaré.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Shao, J. and Tu, D. (1995). *The jackknife and bootstrap*. Springer.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41(2):463–501.

Spokoiny, V. (2012a). Parametric estimation. Finite sample theory. *The Annals of Statistics*, 40(6):2877–2909.

Spokoiny, V. (2012b). Supplement to "Parametric estimation. Finite sample theory".

Spokoiny, V. (2013). Bernstein-von Mises Theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430*.

Spokoiny, V. and Willrich, N. (2015). Bootstrap tuning in ordered model selection. *In preparation*.

Spokoiny, V. and Zhilova, M. (2014). Bootstrap confidence sets under model misspecification. *To appear in the Annals of Statistics. arXiv:1410.0347*.

Spokoiny, V. G. and Zhilova, M. M. (2013). Uniform properties of the local maximum likelihood estimate. *Automation and Remote Control*, 74(10):1656–1669.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151.

Talagrand, M. (2003). *Spin glasses: a challenge for mathematicians: cavity and mean field models*, volume 46. Springer.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Westfall, P. H. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Working, H. and Hotelling, H. (1929). Applications of the theory of error to the interpretation of trends. *Journal of the American Statistical Association*, 24(165A):73–85.

SFB 649 Discussion Paper Series 2015

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Pricing Kernel Modeling" by Denis Belomestny, Shujie Ma and Wolfgang Karl Härdle, January 2015.
002 "Estimating the Value of Urban Green Space: A hedonic Pricing Analysis of the Housing Market in Cologne, Germany" by Jens Kolbe and Henry Wüstemann, January 2015.
003 "Identifying Berlin's land value map using Adaptive Weights Smoothing" by Jens Kolbe, Rainer Schulz, Martin Wersing and Axel Werwatz, January 2015.
004 "Efficiency of Wind Power Production and its Determinants" by Simone Pieralli, Matthias Ritter and Martin Odening, January 2015.
005 "Distillation of News Flow into Analysis of Stock Reactions" by Junni L. Zhang, Wolfgang K. Härdle, Cathy Y. Chen and Elisabeth Bommes, January 2015.
006 "Cognitive Bubbles" by Ciril Bosch-Rosay, Thomas Meissnerz and Antoni Bosch-Domènech, February 2015.
007 "Stochastic Population Analysis: A Functional Data Approach" by Lei Fang and Wolfgang K. Härdle, February 2015.
008 "Nonparametric change-point analysis of volatility" by Markus Bibinger, Moritz Jirak and Mathias Vetter, February 2015.
009 "From Galloping Inflation to Price Stability in Steps: Israel 1985–2013" by Rafi Melnick and till Strohsal, February 2015.
010 "Estimation of NAIRU with Inflation Expectation Data" by Wei Cui, Wolfgang K. Härdle and Weining Wang, February 2015.
011 "Competitors In Merger Control: Shall They Be Merely Heard Or Also Listened To?" by Thomas Giebe and Miyu Lee, February 2015.
012 "The Impact of Credit Default Swap Trading on Loan Syndication" by Daniel Streitz, March 2015.
013 "Pitfalls and Perils of Financial Innovation: The Use of CDS by Corporate Bond Funds" by Tim Adam and Andre Guettler, March 2015.
014 "Generalized Exogenous Processes in DSGE: A Bayesian Approach" by Alexander Meyer-Gohde and Daniel Neuhoff, March 2015.
015 "Structural Vector Autoregressions with Heteroskedasticy" by Helmut Lütkepohl and Aleksei Netšunajev, March 2015.
016 "Testing Missing at Random using Instrumental Variables" by Christoph Breunig, March 2015.
017 "Loss Potential and Disclosures Related to Credit Derivatives – A Cross-Country Comparison of Corporate Bond Funds under U.S. and German Regulation" by Dominika Paula Gałkiewicz, March 2015.
018 "Manager Characteristics and Credit Derivative Use by U.S. Corporate Bond Funds" by Dominika Paula Gałkiewicz, March 2015.
019 "Measuring Connectedness of Euro Area Sovereign Risk" by Rebekka Gätjen Melanie Schienle, April 2015.
020 "Is There an Asymmetric Impact of Housing on Output?" by Tsung-Hsien Michael Lee and Wenjuan Chen, April 2015.
021 "Characterizing the Financial Cycle: Evidence from a Frequency Domain Analysis" by Till Strohsal, Christian R. Proaño and Jürgen Wolters, April 2015.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2015

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

022 "Risk Related Brain Regions Detected with 3D Image FPCA" by Ying Chen, Wolfgang K. Härdle, He Qiang and Piotr Majer, April 2015.
023 "An Adaptive Approach to Forecasting Three Key Macroeconomic Variables for Transitional China" by Linlin Niu, Xiu Xu and Ying Chen, April 2015.
024 "How Do Financial Cycles Interact? Evidence from the US and the UK" by Till Strohsal, Christian R. Proaño, Jürgen Wolters, April 2015.
025 "Employment Polarization and Immigrant Employment Opportunities" by Hanna Wielandt, April 2015.
026 "Forecasting volatility of wind power production" by Zhiwei Shen and Matthias Ritter, May 2015.
027 "The Information Content of Monetary Statistics for the Great Recession: Evidence from Germany" by Wenjuan Chen and Dieter Nautz, May 2015.
028 "The Time-Varying Degree of Inflation Expectations Anchoring" by Till Strohsal, Rafi Melnick and Dieter Nautz, May 2015.
029 "Change point and trend analyses of annual expectile curves of tropical storms" by P.Burdejova, W.K.Härdle, P.Kokoszka and Q.Xiong, May 2015.
030 "Testing for Identification in SVAR-GARCH Models" by Helmut Luetkepohl and George Milunovich, June 2015.
031 " Simultaneous likelihood-based bootstrap confidence sets for a large number of models" by Mayya Zhilova, June 2015.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".