The Annals of Statistics
2015, Vol. 43, No. 6, 2653–2675
DOI: 10.1214/15-AOS1355

BOOTSTRAP CONFIDENCE SETS UNDER MODEL MISSPECIFICATION

BY VLADIMIR SPOKOINY[*,†,‡,§,¶,1] AND MAYYA ZHILOVA[*,2]

Weierstrass Institute for Applied Analysis and Stochastics[*], *Humboldt University Berlin*[†], *Moscow Institute of Physics and Technology*[‡], *Institute for Information Transmission Problems RAS*[§] *and Higher School of Economics, Moscow*[¶]

A multiplier bootstrap procedure for construction of likelihood-based confidence sets is considered for finite samples and a possible model misspecification. Theoretical results justify the bootstrap validity for a small or moderate sample size and allow to control the impact of the parameter dimension p: the bootstrap approximation works if p^3/n is small. The main result about bootstrap validity continues to apply even if the underlying parametric model is misspecified under the so-called small modelling bias condition. In the case when the true model deviates significantly from the considered parametric family, the bootstrap procedure is still applicable but it becomes a bit conservative: the size of the constructed confidence sets is increased by the modelling bias. We illustrate the results with numerical examples for misspecified linear and logistic regressions.

1. Introduction.

Since introducing in 1979 by Efron (1979), the bootstrap procedure became one of the most powerful and common tools in statistical confidence estimation and hypothesis testing. Many versions and extensions of the original bootstrap method have been proposed in the literature; see, for example, Barbe and Bertail (1995), Bücher and Dette (2013), Chatterjee and Bose (2005), Chen and Pouzo (2009, 2015), Horowitz (2001), Janssen (1994), Lavergne and Patilea (2013), Ma and Kosorok (2005), Mammen (1993), Newton and Raftery (1994), Wu (1986) among many others. This paper focuses on the multiplier bootstrap procedure which attracted a lot of attention last time due to its nice theoretical properties and numerical performance. We mention the papers of Chatterjee and Bose (2005), Arlot, Blanchard and Roquain (2010) and Chernozhukov, Chetverikov and

Received November 2014; revised June 2015.

[1]Supported in part by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government Grant, ag. 11.G34.31.0073, by the German Research Foundation (DFG) through the Research Unit 1735 and supported in part by the Russian Science Foundation grant (project 14-50-00150).

[2]Supported by the German Research Foundation (DFG) through the Collaborative Research Center 649 "Economic Risk" and supported in part by the Russian Science Foundation grant (project 14-50-00150).

MSC2010 subject classifications. Primary 62F25, 62F40; secondary 62E17.

Key words and phrases. Likelihood-based bootstrap confidence set, finite sample size, multiplier/weighted bootstrap, Gaussian approximation, Pinsker's inequality.

Kato (2013) for the most advanced recent results. Chatterjee and Bose (2005) showed some results on asymptotic bootstrap consistency in a very general framework for estimators obtained by solving estimating equations. Chernozhukov, Chetverikov and Kato (2013) presented a number of nonasymptotic results on bootstrap validity with applications to special problems like testing many moment restrictions or parameter choice for a LASSO procedure. Arlot, Blanchard and Roquain (2010) constructed a nonasymptotical confidence bound in ℓ_s-norm ($s \in [1, \infty]$) for the mean of a sample of high dimensional i.i.d. Gaussian vectors (or with a symmetric and bounded distribution), using the generalized weighted bootstrap for resampling of the quantiles.

This paper makes a further step in studying the multiplier bootstrap method in the problem of confidence estimation by a quasi maximum likelihood method. For a rather general parametric model, we consider likelihood-based confidence sets with the radius determined by a multiplier bootstrap. The aim of the study is to check the validity of the bootstrap procedure in situations with a growing parameter dimension, a limited sample size, and a possible misspecification of the parametric assumption. The main result of the paper explicitly describes the error term of the bootstrap approximation. This particularly allows to track the impact of the parameter dimension p and of the sample size n in the quality of the bootstrap procedure. As one of the corollaries, we show bootstrap validity under the constraint "p^3/n-small." Chatterjee and Bose (2005) stated results under the condition "p/n-small" but their results only apply to low dimensional projections of the MLE vector. In the likelihood-based approach, the construction involves the Euclidean norm of the MLE which leads to completely different tools and results. Chernozhukov, Chetverikov and Kato (2013) allowed a huge parameter dimension with "$\log(p)/n$ small" but they essentially work with a family of univariate tests which again differs essentially from the maximum likelihood approach.

Another interesting and important issue is the impact of the model misspecification on the accuracy of bootstrap approximation. A surprising corollary of our error bounds is that the bootstrap confidence set can be used even if the underlying parametric model is slightly misspecified under the so-called *small modelling bias* (SmB) condition. If the modelling bias becomes large, the bootstrap confidence sets are still applicable, but they become more and more conservative. (SmB) condition is given in Section 4 and it is consistent with classical bias–variance relation in nonparametric estimation.

Our theoretical study uses the square-root Wilks (sq-Wilks) expansion from Spokoiny (2012, 2013) which approximates the square root likelihood ratio statistic by the norm of the standardized score vector. Further, we extend the sq-Wilks expansion to the bootstrap log-likelihood and adopt the Gaussian approximation theory (GAR) to the special case when the distribution of the Euclidean norm of a non-Gaussian vector is approximated by the distribution of the norm of a Gaussian one with the same first and second moments. The Gaussian comparison technique based on the Pinsker inequality completes the study and allows to bridge the real unknown coverage probability and the conditional bootstrap coverage probability under (SmB) condition. In the case of a large modelling bias, we state a one-sided bound: the bootstrap quantiles are uniformly larger than the real ones. This effect is nicely confirmed by our simulation study.

Now consider the problem and the approach in more detail. Let the data sample $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ consist of *independent* random observations and belong to the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We do not assume that the observations Y_i are identically distributed; moreover, no specific parametric structure of $\mathbb{P}$ is being required. In order to explain the idea of the approach we start here with a parametric case, however, assumption (1.1) below is not required for the results. Let $\mathbb{P}$ belong to some known regular parametric family $\{\mathbb{P}_{\boldsymbol{\theta}}\} \stackrel{\text{def}}{=} \{\mathbb{P}_{\boldsymbol{\theta}} \ll \mu_0, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$. In this case, the true parameter $\boldsymbol{\theta}^* \in \Theta$ is such that

$$\mathbb{P} \equiv \mathbb{P}_{\boldsymbol{\theta}^*} \in \{\mathbb{P}_{\boldsymbol{\theta}}\}, \tag{1.1}$$

and the initial problem of finding the properties of unknown distribution $\mathbb{P}$ is reduced to the equivalent problem for the finite-dimensional parameter $\boldsymbol{\theta}^*$. The parametric family $\{\mathbb{P}_{\boldsymbol{\theta}}\}$ induces the log-likelihood process $L(\boldsymbol{\theta})$ of the sample $\mathbf{Y}$,

$$L(\boldsymbol{\theta}) = L(\mathbf{Y}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \log\left(\frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mu_0}(\mathbf{Y})\right)$$

and the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}^*$,

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}). \tag{1.2}$$

The asymptotic Wilks phenomenon [Wilks (1938)] states that for the case of i.i.d. observations with the sample size tending to the infinity the likelihood ratio statistic converges in distribution to $\chi_p^2/2$, where p is the parameter dimension

$$2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\} \xrightarrow{w} \chi_p^2, \qquad n \to \infty.$$

Define the likelihood-based confidence set as

$$\mathcal{E}(\mathfrak{z}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}) \le \mathfrak{z}^2/2\}, \tag{1.3}$$

then the Wilks phenomenon implies

$$\mathbb{P}\{\boldsymbol{\theta}^* \in \mathcal{E}(\mathfrak{z}_{\alpha, \chi_p^2})\} \to \alpha, \qquad n \to \infty,$$

where $\mathfrak{z}^2_{\alpha, \chi_p^2}$ is the $(1 - \alpha)$-quantile for the χ_p^2 distribution. This result is very important and useful under the parametric assumption, that is, when (1.1) holds. In this case, the limit distribution of the likelihood ratio is independent of the model parameters or in other words it is *pivotal*. By this result, a sufficiently large sample size allows to construct the confidence sets for $\boldsymbol{\theta}^*$ with a given coverage probability. However, a possibly low speed of convergence of the likelihood ratio statistic makes the asymptotic Wilks result hardly applicable to the case of small or moderate samples. Moreover, the asymptotical pivotality breaks down if the parametric assumption (1.1) does not hold [see Huber (1967)] and, therefore, the whole approach may be misleading if the model is considerably misspecified. If the assumption (1.1) does not hold, then the "true" parameter is defined by the projection of the true measure $\mathbb{P}$ on the parametric family $\{\mathbb{P}_{\boldsymbol{\theta}}\}$:

$$\boldsymbol{\theta}^{*} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}L(\boldsymbol{\theta}). \tag{1.4}$$

The recent results by Spokoiny (2012, 2013) provide a nonasymptotic version of square-root Wilks phenomenon for the case of misspecified model. It holds with an exponentially high probability

$$\left|\sqrt{2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^{*})\}} - \|\boldsymbol{\xi}\|\right| \le \Delta_{\mathrm{W}} \simeq \frac{p}{\sqrt{n}}, \tag{1.5}$$

where $\boldsymbol{\xi} \stackrel{\text{def}}{=} D_0^{-1}\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^{*})$, $D_0^2 \stackrel{\text{def}}{=} -\nabla_{\boldsymbol{\theta}}^2 \mathbb{E}L(\boldsymbol{\theta}^{*})$. The bound is nonasymptotical, the approximation error term Δ_{W} has an explicit form (the precise statement is given in Theorem B.2, Section B.1 of the supplementary material [Spokoiny and Zhilova (2015)], and it depends on the parameter dimension p, sample size n and the probability of the random set on which the result holds.

Due to this bound, the original problem of finding a quantile of the LR test statistic $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^{*})$ is reduced to a similar question for the approximating quantity $\|\boldsymbol{\xi}\|$. The difficulty here is that in general $\|\boldsymbol{\xi}\|$ is nonpivotal, it depends on the unknown distribution $\mathbb{P}$ and the target parameter $\boldsymbol{\theta}^{*}$.

In the present work, we study the *multiplier bootstrap* (or *weighted bootstrap*) procedure for estimation of the quantiles of the likelihood ratio statistic. The idea of the procedure is to mimic a distribution of the likelihood ratio statistic by reweighing its summands with random multipliers independent of the data

$$L^{\circ}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \log\left(\frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mu_0}(Y_i)\right) u_i .$$

Here, the probability distribution is taken conditionally on the data $\mathbf{Y}$, which is denoted by the sign $^{\circ}$ (also $\mathbb{E}^{\circ}$ and Var° denote expectation and variance operators w.r.t. the probability measure conditional on $\mathbf{Y}$). The random weights $u_1, \ldots, u_n$ are i.i.d., independent of $\mathbf{Y}$ and it holds for them: $\mathbb{E}^{\circ}(u_i) = 1$, $\mathrm{Var}^{\circ}(u_i) = 1$, $\mathbb{E}^{\circ}\exp(u_i) < \infty$. Therefore, the multiplier bootstrap induces the probability space conditional on the data $\mathbf{Y}$. A simple but important observation is that $\mathbb{E}^{\circ}L^{\circ}(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta})$, and hence,

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}^{\circ}L^{\circ}(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \widetilde{\boldsymbol{\theta}}.$$

This means that the target parameter in the bootstrap world is precisely known and it coincides with the maximum likelihood estimator $\widetilde{\boldsymbol{\theta}}$ conditioned on $\mathbf{Y}$, therefore, the bootstrap likelihood ratio statistic $L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta}\in\Theta} L^{\circ}(\boldsymbol{\theta}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})$ is fully computable and leads to a simple computational procedure for the approximation of the distribution of $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$.

The goal of the present study is to show in a nonasymptotic way the validity of the described multiplier bootstrap procedure and to obtain an explicit bound on the error of coverage probability. In other words, we are interested in nonasymptotic approximation of the distribution of $\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}^{1/2}$ with the distribution of $\{L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})\}^{1/2}$. So far there exist very few theoretical nonasymptotic results about bootstrap validity. Classical asymptotic tools for showing the bootstrap consistency are based on weak convergence arguments which are not applicable in the finite sample set-up. Some different methods have to be applied. In particular, the approach of Liu (1988) based on Berry–Esseen theorem can be extended to a finite sample set-up with a univariate parameter. For a high dimensional parameter space, important contributions are done in the recent papers by Arlot, Blanchard and Roquain (2010) and Chernozhukov, Chetverikov and Kato (2013). The latter paper used a Gaussian approximation, Gaussian comparison and Gaussian anti-concentration technique in high dimension. Our approach is similar but we combine it with the square-root Wilks expansion and use Pinsker's inequality for Gaussian comparison and anti-concentration steps. The main steps of our theoretical study are illustrated by the following scheme:

$$
\begin{array}{lcccccl}
 & & \text{sq-Wilks theorem} & & \text{Gauss. approx.} & & \\
\mathbf{Y}\text{-world: } & \sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)} & \underset{p/\sqrt{n}}{\approx} & \|\boldsymbol{\xi}\| & \underset{(p^3/n)^{1/8}}{\overset{w}{\approx}} & \|\overline{\boldsymbol{\xi}}\| & \\
(1.6) & & & & & w \wr\!\wr\ \sqrt{p}\,\delta^2_{\mathrm{smb}} & \text{Gauss. compar.} \\
\text{Bootstrap world: } & \sqrt{2L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - 2L^\circ(\widetilde{\boldsymbol{\theta}})} & \underset{p/\sqrt{n}}{\approx} & \|\boldsymbol{\xi}^\circ\| & \underset{(p^3/n)^{1/8}}{\overset{w}{\approx}} & \|\overline{\boldsymbol{\xi}}^\circ\| &
\end{array}
$$

where

$$
\boldsymbol{\xi}^\circ \stackrel{\text{def}}{=} \boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} D_0^{-1}\nabla_{\boldsymbol{\theta}}\bigl[L^\circ(\boldsymbol{\theta}^*) - \mathbb{E}^\circ L^\circ(\boldsymbol{\theta}^*)\bigr].
$$

The vectors $\overline{\boldsymbol{\xi}}$ and $\overline{\boldsymbol{\xi}}^\circ$ are zero mean Gaussian and they mimic the covariance structure of the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^\circ$: $\overline{\boldsymbol{\xi}} \sim \mathcal{N}(0, \operatorname{Var}\boldsymbol{\xi})$, $\overline{\boldsymbol{\xi}}^\circ \sim \mathcal{N}(0, \operatorname{Var}^\circ \boldsymbol{\xi}^\circ)$.

The error term shown below each arrow corresponds to the i.i.d. case considered in details in Section 4.4. The upper line of the scheme corresponds to the **Y**-world, the lower line—to the bootstrap world. In both lines, we apply two steps for approximating the corresponding likelihood ratio statistics. The first approximating step is the nonasymptotic square-root Wilks theorem: the bound (1.5) for the **Y**-case and a similar statement for the bootstrap world, which is obtained in Theorem B.4, Section B.2 in Spokoiny and Zhilova (2015). The corresponding error is of order $p/\sqrt{n}$ for the case of i.i.d. observations; in the bootstrap world the square-root Wilks expansion implies

$$
\Bigl|\sqrt{2L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - 2L^\circ(\widetilde{\boldsymbol{\theta}})} - \|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\|\Bigr| \le \mathtt{C}p/\sqrt{n}
$$

for $\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} D_0^{-1}\nabla_{\boldsymbol{\theta}}[L^{\circ}(\boldsymbol{\theta}) - \mathbb{E}^{\circ}L^{\circ}(\boldsymbol{\theta})]$. In our approximation diagram, we use $\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)$ instead of $\boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}})$ which is more convenient for the GAR step and is justified by Lemma B.7 in Spokoiny and Zhilova (2015) showing that $\|\boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\| \le \mathtt{C}p/\sqrt{n}$.

The next step is called *Gaussian approximation* (GAR) which means that the distribution of the Euclidean norm $\|\boldsymbol{\xi}\|$ of a centered random vector $\boldsymbol{\xi}$ is close to the distribution of the similar norm of a Gaussian vector $\|\overline{\boldsymbol{\xi}}\|$ with the same covariance matrix as $\boldsymbol{\xi}$. A similar statement holds for the vector $\boldsymbol{\xi}^{\circ}$. Thus, the initial problem of comparing the distributions of the likelihood ratio statistics is reduced to the comparison of the distributions of the Euclidean norms of two centered normal vectors $\overline{\boldsymbol{\xi}}$ and $\overline{\boldsymbol{\xi}}^{\circ}$ (Gaussian comparison). This last step links their distributions and encloses the approximating scheme. The Gaussian comparison step is done by computing the Kullback–Leibler divergence between two multivariate Gaussian distributions [i.e., by comparison of the covariance matrices of $\nabla_{\boldsymbol{\theta}}L(\boldsymbol{\theta}^*)$ and $\nabla_{\boldsymbol{\theta}}L^{\circ}(\boldsymbol{\theta}^*)$] and applying Pinsker's inequality [Lemma A.7 in Spokoiny and Zhilova (2015)]. At this point, we need to introduce the "small modelling bias" condition (SmB) from Section 4.2. It is formulated in terms of the following nonnegative-definite $p \times p$ symmetric matrices:

$$H_0^2 \stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)^{\top}], \tag{1.7}$$

$$B_0^2 \stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)]\mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)]^{\top} \tag{1.8}$$

for $\ell_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log(\frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mu_0}(Y_i))$, so that $\operatorname{Var}\{\nabla_{\boldsymbol{\theta}}L(\boldsymbol{\theta}^*)\} = H_0^2 - B_0^2$. If the parametric assumption (1.1) is true or if the data $\mathbf{Y}$ are i.i.d., then it holds $\mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)] \equiv 0$ and $B_0^2 = 0$. The (SmB) condition roughly means that the bias term B_0^2 is small relative to H_0^2. Below we show that the Kullback–Leibler distance between the distributions of two Gaussian vectors $\overline{\boldsymbol{\xi}}$ and $\overline{\boldsymbol{\xi}}^{\circ}$ is bounded by $p\|H_0^{-1}B_0^2H_0^{-1}\|^2/2$. The (SmB) condition precisely means that this quantity is small [in scheme (1.6) it is denoted by $\sqrt{p}\delta_{\mathrm{smb}}^2$]. In Section 4.3, the value $\|H_0^{-1}B_0^2H_0^{-1}\|$ is evaluated for some commonly used models: the case of i.i.d. observations, generalized linear model and linear quantile regression. Below we distinguish between two situations: when the condition (SmB) is fulfilled and the opposite case. Theorems 2.1 and 2.2 in Section 2 deal with the first case. It provides the cumulative error term for the coverage probability of the confidence set (1.3), taken at the $(1-\alpha)$-quantile computed with the multiplier bootstrap procedure. The proof of this result [see Section B.4 in Spokoiny and Zhilova (2015)] summarizes the steps of scheme (1.6). The biggest term in the full error is induced by Gaussian approximation and requires the ratio p^3/n to be small. In the case of a "large modelling bias," that is, when (SmB) does not hold, the multiplier bootstrap procedure continues to apply. It turns out that the bootstrap quantiles increase with the growing modelling bias; hence, the confidence set based on it remains valid, however, it may become conservative. This result is given in Theorem 2.5 of Section 2. The problems of Gaussian approximation and comparison for the Euclidean norm are considered in Sections A.2 and A.4 of the supplementary material [Spokoiny and Zhilova (2015)] in general terms independently of the statistical setting of the paper, and might be interesting by themselves. Section A.4 in Spokoiny and Zhilova (2015) presents also an anti-concentration inequality for the Euclidean norm of a Gaussian vector. This inequality shows how the deviation probability changes with a threshold. The general results on GAR are summarized in Theorem A.1 in the supplementary material [Spokoiny and Zhilova (2015)] and restated in Proposition B.12 in Spokoiny and Zhilova (2015) for the setting of scheme (1.6). These results are also nonasymptotic with explicit errors and apply under the condition that the ratio p^3/n to be small.

In Theorem 2.4, we consider the case of a scalar parameter $p = 1$ with an improved error term. Furthermore, in Section 2.2 we propose a modified version of a quantile function based on a smoothed probability distribution. In this case, the obtained error term is also better than in the general result.

Notation: $\|\cdot\|$ denotes Euclidean norm for vectors and spectral norm for matrices; $\mathtt{C}$ is a generic constant. The value $\mathtt{x} > 0$ describes our tolerance level: all the results will be valid on a random set of probability $(1 - C\mathrm{e}^{-\mathtt{x}})$ for an explicit constant C. Everywhere we give explicit error bounds and show how they depend on p and n for the case of the i.i.d. observations $Y_1, \dots, Y_n$ and $\mathtt{x} \le \mathtt{C}\log n$. More details on it are given in Section 4.4. In Section B.3 in the supplementary material [Spokoiny and Zhilova (2015)], we also consider generalized linear model and linear quantile regression, and show for them the dependence on p and n of all the values appearing in main results and their conditions.

The paper is organized as follows: the main results are stated in Section 2. Their proofs are given in Sections B.4, B.5 and B.6 of the supplementary material [Spokoiny and Zhilova (2015)]. Section 3 contains numerical results for misspecified linear and logistic regressions. In Section 4, we give all the required conditions, provide information about dependence of the involved terms on n and p and consider the (SmB) condition for some models. Section A in Spokoiny and Zhilova (2015) collects some useful statements on Gaussian approximation and Gaussian comparison.

2. Multiplier bootstrap procedure.

Let $\ell_i(\boldsymbol{\theta})$ denote the parametric log-density of the ith observation

$$\ell_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log\left(\frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mu_0}(Y_i)\right),$$

then $L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})$. Consider i.i.d. scalar random variables u_i independent of $\mathbf{Y}$ with $\mathbb{E}u_i = 1$, $\operatorname{Var} u_i = 1$, $\mathbb{E}\exp(u_i) < \infty$ for all $i = 1, \dots, n$. Multiply the summands of the likelihood function $L(\boldsymbol{\theta})$ with the new random variables

$$L^{\circ}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}) u_i,$$

then it holds $\mathbb{E}^{\circ} L^{\circ}(\boldsymbol{\theta}) = L(\boldsymbol{\theta})$, where $\mathbb{E}^{\circ}$ stands for the conditional expectation given $\mathbf{Y}$. Therefore, the quasi MLE for the $\mathbf{Y}$-world is a target parameter for the bootstrap world:

$$\underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \, \mathbb{E}^{\circ} L^{\circ}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \, L(\boldsymbol{\theta}) = \widetilde{\boldsymbol{\theta}}.$$

The corresponding quasi MLE under the conditional measure $\mathbb{P}^{\circ}$ is defined as

$$\widetilde{\boldsymbol{\theta}}^{\circ} \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \, L^{\circ}(\boldsymbol{\theta}).$$

The likelihood ratio statistic in the bootstrap world is equal to $L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})$ in which all the entries are known including the function $L^{\circ}(\boldsymbol{\theta})$ and the arguments $\widetilde{\boldsymbol{\theta}}^{\circ}, \widetilde{\boldsymbol{\theta}}$.

Let $1 - \alpha \in (0, 1)$ be an unknown desirable confidence level of the set $\mathcal{E}(\mathfrak{z})$:

$$\mathbb{P}\bigl(\boldsymbol{\theta}^* \in \mathcal{E}(\mathfrak{z})\bigr) \geq 1 - \alpha. \tag{2.1}$$

Here, the parameter $\mathfrak{z} \geq 0$ determines the size of the confidence set. Define $\mathfrak{z}_{\alpha}$ as the minimal possible value of $\mathfrak{z}$ such that (2.1) is fulfilled:

$$\mathfrak{z}_{\alpha} \stackrel{\text{def}}{=} \inf\bigl\{\mathfrak{z} \geq 0 \colon \mathbb{P}\bigl(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\bigr) \leq \alpha\bigr\}. \tag{2.2}$$

For evaluating this value, we apply the multiplier bootstrap procedure which replaces the unknown data distribution with the artificial bootstrap distribution given the observed sample. The target value $\mathfrak{z}_{\alpha}$ is approximated by the value $\mathfrak{z}_{\alpha}^{\circ}$ defined as the upper α-quantile of $\{2L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - 2L^{\circ}(\widetilde{\boldsymbol{\theta}})\}^{1/2}$:

$$\mathfrak{z}_{\alpha}^{\circ} \stackrel{\text{def}}{=} \inf\bigl\{\mathfrak{z} \geq 0 \colon \mathbb{P}^{\circ}\bigl(L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\bigr) \leq \alpha\bigr\}. \tag{2.3}$$

Note that the bootstrap probability $\mathbb{P}^{\circ}$ and log-likelihood excess $L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})$ depends on the data $\mathbf{Y}$ and thus, $\mathfrak{z}_{\alpha}^{\circ}$ is random as well. Theoretical results of the next section justify the proposed approach.

2.1. *Main results.* Now we state the main results for the general set-up. The approximating error terms and the conditions are specified in Section B.3 of the supplementary material [Spokoiny and Zhilova (2015)] for popular examples including i.i.d. observations, generalized regression model and linear quantile regression. Our first result claims that the random quantity $\mathbb{P}^{\circ}(L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2)$ is close in probability to the value $\mathbb{P}(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2)$ for a wide range of $\mathfrak{z}$-values.

THEOREM 2.1. *Let the conditions of Section 4 be fulfilled, then it holds for* $\mathfrak{z} \ge \max\{2, \sqrt{p}\} + \mathtt{C}(p + \mathtt{x})/\sqrt{n}$ *with probability* $\ge 1 - 12\mathrm{e}^{-\mathtt{x}}$:

$$\left|\mathbb{P}\big(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\big) - \mathbb{P}^{\circ}\big(L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\big)\right| \le \Delta_{\mathrm{full}}.$$

The error term $\Delta_{\mathrm{full}} \le \mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$ *in the case of i.i.d. model*; *see Section* 4.4. *Explicit definition of the error term* Δ_{full} *is given in Section* B.4 *of the supplementary material* [*Spokoiny and Zhilova* (2015)]; *see* (B.41) *and* (B.42) *therein*.

The term Δ_{full} can be viewed as a sum of the error terms corresponding to each step in the scheme (1.6). The largest error term equal to $\mathtt{C}\{(p + \mathtt{x})^3/n\}^{1/8}$ is induced by GAR. This error rate is not always optimal for GAR, for example, in the case of $p = 1$ or for the i.i.d. observations [see Remark A.2 in Spokoiny and Zhilova (2015)]. In Theorems 2.4 and 2.6, the rate is $\mathtt{C}\{(p + \mathtt{x})^3/n\}^{1/2}$.

The next result can be viewed as "bootstrap validity."

THEOREM 2.2 (Validity of the bootstrap under a small modelling bias). *Assume the conditions of Theorem* 2.1. *Then for* $\alpha \le 1 - 8\mathrm{e}^{-\mathtt{x}}$, *it holds*

$$\left|\mathbb{P}\big(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > (\mathfrak{z}^{\circ}_{\alpha})^2/2\big) - \alpha\right| \le \Delta_{\mathfrak{z},\mathrm{full}}.$$

The error term $\Delta_{\mathfrak{z},\mathrm{full}} \le \mathtt{C}\{(p + \mathtt{x})^3/n\}^{1/8}$ *in the case of the i.i.d. model*; *see Section* 4.4. *For a precise description*, *see* (B.46) *and* (B.47) *of the supplementary material* [*Spokoiny and Zhilova* (2015)].

In view of definition (1.3) of the likelihood-based confidence set, Theorem 2.1 implies the following:

COROLLARY 2.3 (Coverage probability error). *Under the conditions of Theorem* 2.2, *it holds that*

$$\left|\mathbb{P}\{\boldsymbol{\theta}^* \in \mathcal{E}(\mathfrak{z}^{\circ}_{\alpha})\} - (1 - \alpha)\right| \le \Delta_{\mathfrak{z},\mathrm{full}}.$$

REMARK 2.1 (Critical dimension). The error term Δ_{full} depends on the ratio p^3/n. The bootstrap validity can be only stated if this ratio is small. The obtained error bound seems to be mainly of theoretical interest, because the condition "$(p^3/n)^{1/8}$ is small" may require a huge sample. However, it provides some qualitative information about the bootstrap behavior as the parameter dimension grows. Our numerical results show that the accuracy of bootstrap approximation is very reasonable in a variety of examples with $p \ll n$.

In the following theorem, we consider the case of the scalar parameter $p = 1$. The obtained error rate is $1/\sqrt{n}$, which is sharper than $1/n^{1/8}$. Instead of the GAR for the Euclidean norm from Section A in Spokoiny and Zhilova (2015), we use here the Berry–Esseen theorem [see also Remark A.2 in Spokoiny and Zhilova (2015)].

THEOREM 2.4 (The case of $p = 1$, using the Berry–Esseen theorem). *Let the conditions of Section 4 be fulfilled.*

1. *For* $\mathfrak{z} \geq 1 + \mathtt{C}(1 + \mathtt{x})/\sqrt{n}$, *it holds with probability* $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$
$$\left|\mathbb{P}\big(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\big) - \mathbb{P}^{\circ}\big(L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\big)\right| \leq \Delta_{\mathrm{B.E.,full}}.$$
2. *For* $\alpha \leq 1 - 8\mathrm{e}^{-\mathtt{x}}$
$$\left|\mathbb{P}\big(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > (\mathfrak{z}_{\alpha}^{\circ})^2/2\big) - \alpha\right| \leq \Delta_{\mathrm{B.E.}\ \mathfrak{z},\mathrm{full}}.$$

The error terms $\Delta_{\mathrm{B.E.,full}}, \Delta_{\mathrm{B.E.}\ \mathfrak{z},\mathrm{full}} \leq \mathtt{C}(1 + \mathtt{x})/\sqrt{n}$ *in the case* 4.4. *Explicit definitions of* $\Delta_{\mathrm{B.E.,full}}$ *is given in* (B.48) *and* (B.49) *in Section* B.4 *of the supplementary material* [*Spokoiny and Zhilova* (2015)].

REMARK 2.2 (Bootstrap validity and weak convergence). The standard way of proving the bootstrap validity is based on weak convergence arguments; see, for example, Mammen (1992), van der Vaart and Wellner (1996), Janssen and Pauls (2003), Chatterjee and Bose (2005). If the statistic $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ weakly converges to a χ^2-type distribution, one can state an asymptotic version of the results of Theorems 2.1, 2.4. Our way is based on a kind of nonasymptotic Gaussian approximation and Gaussian comparison for random vectors and allows to get explicit error terms.

REMARK 2.3 (Use of Edgeworth expansion). The classical results on confidence sets for the mean of population states the accuracy of order $1/n$ based on the second-order Edgeworth expansion; see Hall (1992). Unfortunately, if the considered parametric model can be misspecified, even the leading term is affected by the modelling bias, and the use of Edgeworth expansion cannot help in improving the bootstrap accuracy.

REMARK 2.4 (Choice of the weights). In our construction, similarly to Chatterjee and Bose (2005), we apply a general distribution of the bootstrap weights u_i under some moment conditions. One particularly can use Gaussian multipliers as suggested by Chernozhukov, Chetverikov and Kato (2013). This leads to the exact Gaussian distribution of the vectors $\boldsymbol{\xi}^{\circ}$ and is helpful to avoid one step of Gaussian approximation for these vectors.

REMARK 2.5 (Skipping the Gaussian approximation step). The biggest error term $\mathtt{C}\{(p + \mathtt{x})^3/n\}^{1/8}$ in Theorem 2.1 is induced by the Gaussian approximation step. In some particular cases, the Gaussian approximation step can be avoided leading to better error bounds. For example, if the marginal score vectors $\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)$ are normally distributed, and the random bootstrap weights are normal as well, $u_i \sim \mathcal{N}(1, 1)$, then the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^{\circ}$ are automatically normal, and the GAR step can be skipped. If the marginal score vectors $\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)$ are i.i.d. and symmetrically distributed [s.t. $\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \sim -\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)$], and the centered bootstrap weights follow the Rademacher distribution [$u_i \sim 2\,\mathrm{Bernoulli}(0.5)$], then the recent results by Arlot, Blanchard and Roquain (2010) can be applied to show that the conditional distribution of $\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\|$ given the data is close to the distribution of $\|\boldsymbol{\xi}\|$. However, such methods require some special structural conditions on the underlying measure $\mathbb{P}$ like symmetricity or Gaussianity of the errors and may fail if these conditions are violated. It remains a challenging question how a nice performance of a general bootstrap procedure even for small or moderate samples can be explained.

Now we discuss the impact of modelling bias, which comes from a possible misspecification of the parametric model. As explained by the approximating diagram (1.6), the distance between the distributions of the likelihood ratio statistics can be characterized via the distance between two multivariate normal distributions. To state the result, let us recall the definition of the full Fisher information matrix $D_0^2 \stackrel{\mathrm{def}}{=} -\nabla_{\boldsymbol{\theta}}^2 \mathbb{E} L(\boldsymbol{\theta}^*)$. For the matrices H_0^2 and B_0^2, given in (1.7) and (1.8), it holds $H_0^2 > B_0^2 \geq 0$. If the parametric assumption (1.1) is true or in the case of an i.i.d. sample $\mathbf{Y}$, $B_0^2 = 0$. Under the condition (SmB) $\|H_0^{-1} B_0^2 H_0^{-1}\|$ enters linearly in the error term Δ_{full} in Theorem 2.1.

The first statement in Theorem 2.5 below says that the effective coverage probability of the confidence set based on the multiplier bootstrap is *larger* than the nominal coverage probability up to the error term $\Delta_{\mathrm{b,full}} \leq \mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$. The inequalities in the second part of Theorem 2.5 prove the *conservativeness of the bootstrap quantiles*: the quantity $\sqrt{\mathrm{tr}\{D_0^{-1} H_0^2 D_0^{-1}\}} - \sqrt{\mathrm{tr}\{D_0^{-1}(H_0^2 - B_0^2) D_0^{-1}\}} \geq 0$ increases with the growing modelling bias.

THEOREM 2.5 (Performance of the bootstrap for a large modelling bias). *Under the conditions of Section 4 except for* (SmB), *it holds for* $\mathfrak{z} \geq \max\{2, \sqrt{p}\} + \mathtt{C}(p+\mathtt{x})/\sqrt{n}$ *with probability* $\geq 1 - 14\mathrm{e}^{-\mathtt{x}}$:

1.
$$\mathbb{P}\big(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\big) \leq \mathbb{P}^{\circ}\big(L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\big) + \Delta_{\mathrm{b,full}}.$$

2.
$$\begin{aligned}\mathfrak{z}_{\alpha}^{\circ} &\geq \mathfrak{z}_{(\alpha+\Delta_{\mathrm{b,full}})} \\ &\quad + \sqrt{\mathrm{tr}\{D_0^{-1} H_0^2 D_0^{-1}\}} - \sqrt{\mathrm{tr}\{D_0^{-1}(H_0^2 - B_0^2) D_0^{-1}\}} - \Delta_{\mathtt{qf},1}, \\ \mathfrak{z}_{\alpha}^{\circ} &\leq \mathfrak{z}_{(\alpha-\Delta_{\mathrm{b,full}})} \\ &\quad + \sqrt{\mathrm{tr}\{D_0^{-1} H_0^2 D_0^{-1}\}} - \sqrt{\mathrm{tr}\{D_0^{-1}(H_0^2 - B_0^2) D_0^{-1}\}} + \Delta_{\mathtt{qf},2}.\end{aligned}$$

The term $\Delta_{\mathrm{b,full}} \leq \mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$ *is given in* (B.51) *in Section* B.5 *of the supplementary material* [*Spokoiny and Zhilova* (2015)]. *The positive values* $\Delta_{\mathtt{qf},1}$, $\Delta_{\mathtt{qf},2}$

are given in (B.55), (B.54) *in Section* B.5 *in* Spokoiny and Zhilova (2015); *they are bounded from above with* $(\mathfrak{a}^2 + \mathfrak{a}_B^2)(\sqrt{8\mathtt{x}p} + 6\mathtt{x})$ *for the constants* $\mathfrak{a}^2 > 0$, $\mathfrak{a}_B^2 \geq 0$ *from conditions* $(\mathcal{I})$, $(\mathcal{I}_B)$.

REMARK 2.6. There exists some literature on robust (and heteroscedasticity robust) bootstrap procedures; see, for example, Mammen (1993), Aerts and Claeskens (2001), Kline and Santos (2012). However, to our knowledge there are no robust bootstrap procedures for the likelihood ratio statistic, most of the results compare the distribution of the estimator obtained from estimating equations, or Wald/score test statistics with their bootstrap counterparts in the i.i.d. setup. In our context, this would correspond to the noise misspecification in the log-likelihood function and it is addressed automatically by the multiplier bootstrap. Our notion of modelling bias includes the situation when the target value $\boldsymbol{\theta}^*$ from (1.4) only defines a projection (the best parametric fit) of the data distribution. In particularly, the quantities $\mathbb{E}\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)$ for different i do not necessarily vanish yielding a significant modelling bias. Similar notion of misspecification is used in the literature on Generalized Method of Moments; see, for example, Hall (2005). Chapter 5 therein considers the hypothesis testing problem with two kinds of misspecification: local and nonlocal, which would correspond to our small and large modelling bias cases.

An interesting message of Theorem 2.5 is that the multiplier bootstrap procedure ensures a prescribed coverage level for this target value $\boldsymbol{\theta}^*$ even without small modelling bias restriction; however, in this case, the method is somehow conservative because the modelling bias is transferred into the additional variance in the bootstrap world. The numerical experiments in Section 3 agree with this result.

2.2. *Smoothed version of a quantile function.* This section explains how to improve the accuracy of bootstrap approximation using a smoothed quantile function. The $(1-\alpha)$-quantile of $\sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)}$ is defined as

$$
\begin{aligned}
\mathfrak{z}_\alpha &\stackrel{\text{def}}{=} \inf\{\mathfrak{z} \geq 0 : \mathbb{P}(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2) \leq \alpha\} \\
&= \inf\{\mathfrak{z} \geq 0 : \mathbb{E}\mathbb{1}\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\} \leq \alpha\}.
\end{aligned}
$$

Introduce for $x \geq 0$ and $z, \Delta > 0$ the following function:

$$
g_\Delta(x, z) \stackrel{\text{def}}{=} g\left(\frac{1}{2\Delta z}(x^2 - z^2)\right), \tag{2.4}
$$

where $g(x)$ is a three times differentiable nonnegative function, and grows monotonously from 0 to 1, $g(x) = 0$ for $x \leq 0$ and $g(x) = 1$ for $x \geq 1$, therefore,

$$
\mathbb{1}\{x > 1\} \leq g(x) \leq \mathbb{1}\{x > 0\} \leq g(x+1).
$$

An example of such function is given in (A.8) in Spokoiny and Zhilova (2015). It holds

$$\mathbb{1}\{x - z > \Delta\} \le g_{\Delta}(x, z) \le \mathbb{1}(x - z > 0) \le g_{\Delta}(x, z + \Delta).$$

This approximation is used in the proofs of Theorems 2.1, 2.2 and 2.5 in the part of Gaussian approximation of Euclidean norm of a sum of independent vectors [see Section A.2 in Spokoiny and Zhilova (2015)] yielding the error rate $(p^3/n)^{1/8}$ in the final bound [Theorems 2.1, 2.2 and A.1 in Spokoiny and Zhilova (2015)]. The next result shows that the use of a smoothed quantile function helps to improve the accuracy of bootstrap approximation: it becomes $(p^3/n)^{1/2}$ instead of $(p^3/n)^{1/8}$. The reason is that we do not need to account for the error induced by a smooth approximation of the indicator function.

THEOREM 2.6 [Validity of the bootstrap in the smoothed case under (SmB) condition]. *Let the conditions of Section 4 be fulfilled. It holds for* $\mathfrak{z} \ge \max\{2, \sqrt{p}\} + \mathtt{C}(p + \mathtt{x})/\sqrt{n}$ *and* $\Delta \in (0, 0.22]$ *with probability* $\ge 1 - 12\mathrm{e}^{-\mathtt{x}}$:

$$\left|\mathbb{E} g_{\Delta}\bigl(\sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)}, \mathfrak{z}\bigr) - \mathbb{E}^{\circ} g_{\Delta}\bigl(\sqrt{2L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - 2L^{\circ}(\widetilde{\boldsymbol{\theta}})}, \mathfrak{z}\bigr)\right| \le \Delta_{\mathrm{sm}},$$

where $\Delta_{\mathrm{sm}} \le \mathtt{C}\{(p + \mathtt{x})^3/n\}^{1/2}\Delta^{-3}$ *in the case* 4.4. *An explicit definition of* Δ_{sm} *is given in* (B.59), (B.60) *in Section* B.6 *of the supplementary material* [*Spokoiny and Zhilova* (2015)].

The modified bootstrap quantile function reads as

$$\mathfrak{z}^{\circ}_{\Delta,\alpha} \stackrel{\mathrm{def}}{=} \min\bigl\{\mathfrak{z} \ge 0 : \mathbb{E}^{\circ} g_{\Delta}\bigl(\sqrt{2L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - 2L^{\circ}(\widetilde{\boldsymbol{\theta}})}, \mathfrak{z}\bigr) \le \alpha\bigr\}.$$

3. Numerical results. This section illustrates the performance of the multiplier bootstrap for some artificial examples. We especially aim to address the issues of noise misspecification and of increasing modelling bias. It should be mentioned that the obtained results are nicely consistent with the theoretical statements.

In all the experiments, we took 10^4 data samples for estimation of the empirical c.d.f. of $\sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)}$, and 10^4 $\{u_1, \dots, u_n\}$ samples for each of the 10^4 data samples for the estimation of the quantiles of $\sqrt{2L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - 2L^{\circ}(\widetilde{\boldsymbol{\theta}})}$.

3.1. *Computational error.* Here, we check numerically how well the multiplier procedure works in the case of the correct model. Here, the modelling bias term $\|H_0^{-1} B_0^2 H_0^{-1}\|$ from the (SmB) condition equals to zero by its definition. Let the data come from the following model: $Y_i = \Psi_i^{\top}\boldsymbol{\theta}_0 + \varepsilon_i$, for $i = 1, \dots, n$, where $\varepsilon_i \sim \mathcal{N}(0, 1)$, $\Psi_i \stackrel{\mathrm{def}}{=} (1, X_i, X_i^2, \dots, X_i^{p-1})^{\top}$, the design points $X_1, \dots, X_n$ are equidistant on [0, 1], and the parameter vector $\boldsymbol{\theta}_0 = (1, \dots, 1)^{\top} \in \mathbb{R}^p$. The true likelihood function is $L(\boldsymbol{\theta}) = -\sum_{i=1}^{n}(Y_i - \Psi_i^{\top}\boldsymbol{\theta})^2/2$. In this experiment, we

TABLE 1
Coverage probabilities for the correct model

			Confidence levels					
n	p	$\mathcal{L}(u_i)$	**0.99**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**
50	1	2 Bernoulli(0.5)	0.986	0.942	0.892	0.838	0.792	0.745
		$\mathcal{N}(1, 1)$	0.988	0.945	0.895	0.847	0.803	0.751
		exp(1)	0.988	0.942	0.885	0.833	0.784	0.729
50	3	2 Bernoulli(0.5)	0.984	0.938	0.885	0.838	0.788	0.736
		$\mathcal{N}(1, 1)$	0.994	0.949	0.897	0.844	0.789	0.736
		exp(1)	0.984	0.917	0.835	0.776	0.707	0.650
50	10	2 Bernoulli(0.5)	0.975	0.923	0.866	0.813	0.764	0.715
		$\mathcal{N}(1, 1)$	0.996	0.950	0.877	0.780	0.721	0.644
		exp(1)	0.952	0.827	0.710	0.617	0.541	0.473

consider three cases: the scalar parameter $p = 1$, and the multivariate parameter $p = 3, 10$.

Table 1 shows the effective coverage probabilities of the quantiles estimated using the multiplier bootstrap. The second line contains the range of the nominal confidence levels: $0.99, \dots, 0.75$. The first left column shows the sample size n and the second column—the parameter's dimension p. The third left column describes the distribution of the bootstrap weights: 2 Bernoulli(0.5), $\mathcal{N}(1, 1)$ or exp(1). Below its second line, the table contains the frequencies of the event: "the real likelihood ratio $\le$ the quantile of the bootstrap likelihood ratio."

3.2. *Linear regression with misspecified heteroscedastic errors.* Here, we show on a linear regression model that the quality of the confidence sets obtained by the multiplier bootstrap procedure is not significantly deteriorated by misspecified heteroscedastic errors. Let the data be defined as $Y_i = \Psi_i^\top \boldsymbol{\theta}_0 + \sigma_i \varepsilon_i$, $i = 1, \dots, n$. The i.i.d. random variables $\varepsilon_i \sim \mathrm{Laplace}(0, 2^{-1/2})$ are s.t. $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = 1$. The coefficients σ_i are deterministic: $\sigma_i \stackrel{\mathrm{def}}{=} 0.5\{4 - i (\mathrm{mod}\ 4)\}$. The regressors Ψ_i are the same as in the experiment 3.1. The quasi-likelihood function is also the same as in the previous section: $L(\boldsymbol{\theta}) = -\sum_{i=1}^n (Y_i - \Psi_i^\top \boldsymbol{\theta})^2/2$, and it is misspecified, since it corresponds to $\sigma_i \varepsilon_i \sim \mathcal{N}(0, 1)$. The target point $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, therefore, the modelling bias term $\|H_0^{-1} B_0^2 H_0^{-1}\|$ from the (SmB) condition equals to zero.

Here, we also consider three different parameter's dimensions: $p = 1, 3, 10$ with $\boldsymbol{\theta}_0 = (1, \dots, 1)^\top \in \mathbb{R}^p$. Table 2 describes the second experiment's results similarly to the Table 1.

One can see from the Tables 1 and 2 that the bootstrap procedure does a good job even for small or moderate samples like 50 or 100 if the parameter dimension is not too large. The results are stable w.r.t. the noise misspecification.

TABLE 2
Coverage probabilities for case of misspecified heteroscedastic noise

			Confidence levels					
n	p	$\mathcal{L}(u_i)$	**0.99**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**
50	1	2 Bernoulli(0.5)	0.988	0.947	0.896	0.849	0.799	0.752
		$\mathcal{N}(1, 1)$	0.990	0.949	0.893	0.844	0.794	0.746
		exp(1)	0.989	0.941	0.881	0.825	0.770	0.714
50	3	2 Bernoulli(0.5)	0.984	0.937	0.885	0.834	0.788	0.739
		$\mathcal{N}(1, 1)$	0.996	0.955	0.897	0.839	0.780	0.722
		exp(1)	0.988	0.924	0.846	0.765	0.701	0.634
50	10	2 Bernoulli(0.5)	0.976	0.927	0.870	0.815	0.765	0.715
		$\mathcal{N}(1, 1)$	0.998	0.959	0.891	0.810	0.731	0.655
		exp(1)	0.967	0.850	0.726	0.630	0.552	0.479
100	10	2 Bernoulli(0.5)	0.985	0.935	0.885	0.833	0.781	0.733
		$\mathcal{N}(1, 1)$	0.998	0.970	0.917	0.857	0.786	0.723
		exp(1)	0.989	0.921	0.826	0.741	0.663	0.591

The Rademacher and Gaussian weights demonstrate nearly the same nice performance while the procedure with exponential weights tends to underestimate the real quantiles. This effect becomes especially prominent when the parameter dimension grows to 10.

3.3. *Biased constant regression with misspecified errors*. In the third experiment, we consider biased regression with misspecified i.i.d. errors:

$$Y_i = \beta \sin(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim \text{Laplace}\big(0, 2^{-1/2}\big), \qquad \text{i.i.d.},$$
$$X_i \text{ are equidistant in } [0, 2\pi].$$

Taking the likelihood function $L(\boldsymbol{\theta}) = -\sum_{i=1}^{n}(Y_i - \boldsymbol{\theta})^2/2$ yields $\boldsymbol{\theta}^* = 0$. Therefore, the larger is the deterministic amplitude $\beta > 0$, the bigger is bias of the mean constant regression. The (SmB) condition reads as

$$\begin{aligned}\|H_0^{-1} B_0^2 H_0^{-1}\| &= 1 - \frac{\sum_{i=1}^{n} \operatorname{Var} Y_i}{\beta^2 \sum_{i=1}^{n} \sin^2(X_i) + \sum_{i=1}^{n} \operatorname{Var} Y_i} \\ &= 1 - \frac{1}{\beta^2 (n-1)/2n + 1} \\ &\le 1/\sqrt{n}.\end{aligned}$$

Consider the sample size $n = 50$, and two cases: $\beta = 0.25$ with fulfilled (SmB) condition and $\beta = 1.25$ when (SmB) does not hold. Table 3 shows that for the large bias quantiles yielded by the multiplier bootstrap are conservative. This conservative property of the multiplier bootstrap quantiles is also illustrated with the

TABLE 3
Coverage probabilities for the noise-misspecified biased regression

			Confidence levels					
n	$\mathcal{L}(u_i)$	β	**0.99**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**
50	$\mathcal{N}(1, 1)$	0.25	0.98	0.94	0.89	0.84	0.79	0.74
		1.25	1.0	0.99	0.97	0.94	0.91	0.87

graphs in Figure 1. They show the empirical distribution functions of the likelihood ratio statistics $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ and $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$ for $\beta = 0.25$ and $\beta = 1.25$. On the right graph for $\beta = 1.25$ the empirical distribution functions for the bootstrap case are smaller than the one for the $\mathbf{Y}$ case. It means that for the large bias the bootstrap quantiles are bigger than the $\mathbf{Y}$ quantiles, which increases the diameter of the confidence set based on the bootstrap quantiles. This confidence set remains valid, since it still contains the true parameter with a given confidence level.

Figure 2 shows the growth of the difference between the quantiles of $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$ and $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ with increasing β for the range of the confidence levels: $0.75, 0.8, \dots, 0.99$.

3.4. *Logistic regression with bias.* In this example, we consider logistic regression. Let the data come from the following distribution:

$$Y_i \sim \text{Bernoulli}(\beta X_i), \qquad X_i \text{ are equidistant in } [0, 2], \beta \in (0, 1/2].$$

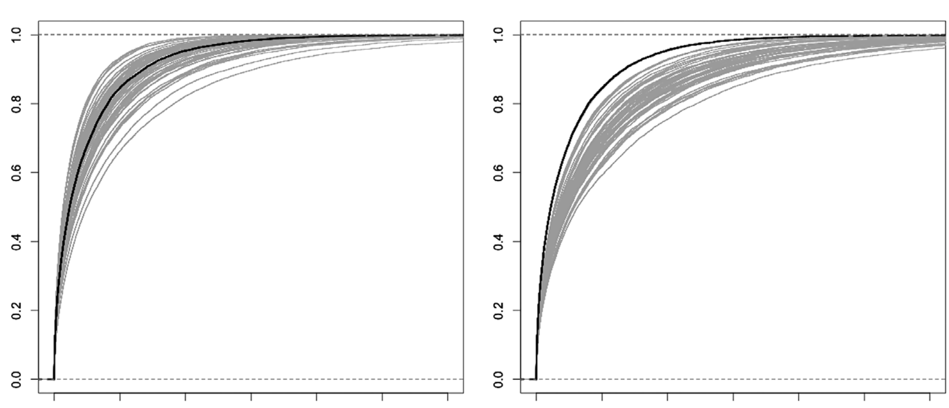


FIG. 1. *Empirical distribution functions of the likelihood ratios.* —— *Empirical distribution function of* $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ *estimated with* 10^4 $\mathbf{Y}$ *samples.* —— 50 *empirical distribution functions of* $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$ *estimated with* $10^4\{u_i\} \sim \exp(1)$ *samples.*

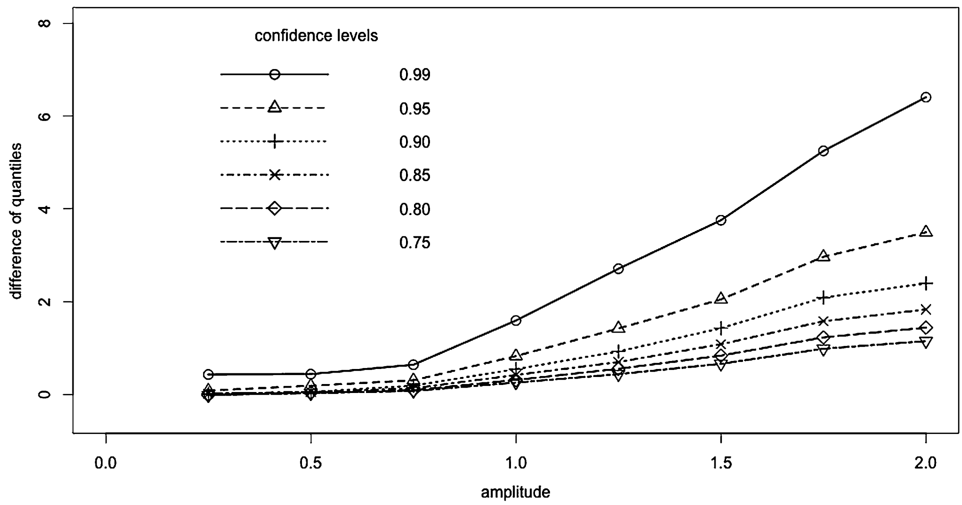


FIG. 2. *The difference ("Bootstrap quantile"−"$\mathbf{Y}$-quantile") growing with modelling bias.*

Consider the likelihood function corresponding to the i.i.d. observations

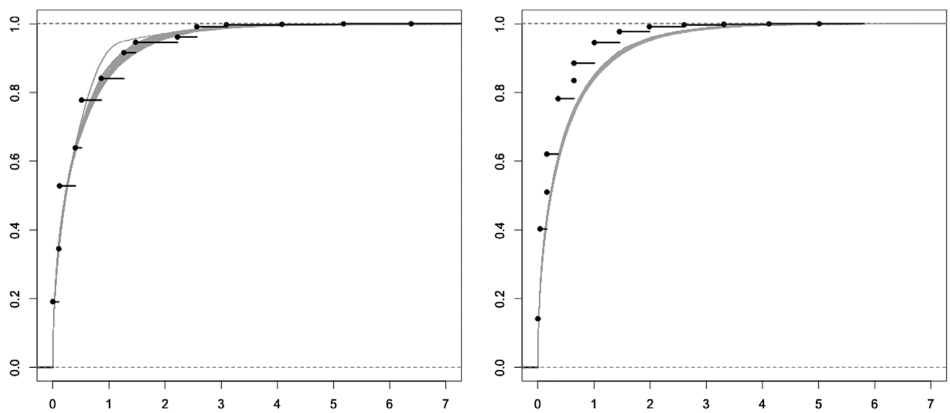
$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{Y_i \boldsymbol{\theta} - \log(1 + \mathrm{e}^{\boldsymbol{\theta}})\}.$$

By definition (1.4) $\boldsymbol{\theta}^* = \log\{\beta/(1-\beta)\}$, bigger values of β induce larger modelling bias. Indeed, the (SmB) condition reads as

$$\begin{aligned}\|H_0^{-1} B_0^2 H_0^{-1}\| &= \frac{\beta^2 \sum_{i=1}^n (X_i - 1)^2}{n\beta^2 + \beta(1-2\beta)\sum_{i=1}^n X_i} \\ &= \frac{\beta}{1-\beta} \cdot \frac{n+1}{3(n-1)} \\ &\le 1/\sqrt{n}.\end{aligned}$$

The graphs on Figure 3 demonstrate the conservativeness of bootstrap quantiles. Here, we consider two cases: $\beta = 0.1$ and $\beta = 0.5$. Similarly to the Example 3.3 in the case of the bigger β on the right graph of Figure 3, the empirical distribution functions of $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$ are smaller than the one for $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$.

4. Conditions. Here, we state the conditions required for the main results. The conditions in Section 4.1 come from the general finite sample theory by Spokoiny (2012). They are required for the results of Sections B.1 and B.2 in the supplementary material [Spokoiny and Zhilova (2015)]. The conditions in Section 4.2 are necessary to prove the results on multiplier bootstrap from Section 2. In Section B.3 in Spokoiny and Zhilova (2015), we consider these conditions in detail for several examples: i.i.d. observations, generalized linear model and linear quantile regression.



$Y_i \sim \text{Bernoulli}(0.1X_i), n = 50$ $Y_i \sim \text{Bernoulli}(0.5X_i), n = 50$

FIG. 3. *Empirical distribution functions of the likelihood ratios for logistic regression.* —— *Empirical distribution function of* $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ *estimated with* 10^4 $\mathbf{Y}$ *samples.* —— 50 *empirical distribution functions of* $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$ *estimated with* $10^4 \{u_i\} \sim \exp(1)$ *samples.*

4.1. *Basic conditions*. Introduce the stochastic part of the likelihood process: $\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta})$, and its marginal summand: $\zeta_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell_i(\boldsymbol{\theta}) - \mathbb{E}\ell_i(\boldsymbol{\theta})$.

(ED_0) There exist a positive-definite symmetric matrix V_0^2 and constants $\mathtt{g} > 0$, $\nu_0 \ge 1$ such that $\operatorname{Var}\{\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\theta}^*)\} \le V_0^2$ and

$$\sup_{\boldsymbol{\gamma} \in \mathbb{R}^p} \log \mathbb{E} \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}^*)}{\|V_0 \boldsymbol{\gamma}\|}\right\} \le \nu_0^2 \lambda^2/2, \qquad |\lambda| \le \mathtt{g}.$$

(ED_2) There exist a constant $\omega > 0$ and for each $\mathtt{r} > 0$ a constant $\mathtt{g}_2(\mathtt{r})$ such that it holds for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ and for $j = 1, 2$

$$\sup_{\boldsymbol{\gamma}_j \in \mathbb{R}^p \|\boldsymbol{\gamma}_j\| \le 1} \log \mathbb{E} \exp\left\{\frac{\lambda}{\omega} \boldsymbol{\gamma}_1^\top D_0^{-1} \nabla_{\boldsymbol{\theta}}^2 \zeta(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2\right\} \le \nu_0^2 \lambda^2/2, \qquad |\lambda| \le \mathtt{g}_2(\mathtt{r}).$$

$(\mathcal{L}_0)$ For each $\mathtt{r} \in [0, \mathtt{r}_0]$ [$\mathtt{r}_0$ comes from condition (B.1) of Theorem B.1 in Spokoiny and Zhilova (2015)] there exists a constant $\delta(\mathtt{r}) \in [0, 1/2]$ s.t. for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ it holds

$$\|D_0^{-1} D^2(\boldsymbol{\theta}) D_0^{-1} - \mathbf{I}_p\| \le \delta(\mathtt{r}),$$

where $D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla_{\boldsymbol{\theta}}^2 \mathbb{E}L(\boldsymbol{\theta})$, $\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}\}$.

$(\mathcal{I})$ There exists a constant $\mathfrak{a} > 0$ s.t. $\mathfrak{a}^2 D_0^2 \ge V_0^2$.

$(\mathcal{L}\mathtt{r})$ For each $\mathtt{r} > \mathtt{r}_0$ there exists a value $\mathtt{b}(\mathtt{r}) > 0$ s.t. $\mathtt{r}\mathtt{b}(\mathtt{r}) \to +\infty$ for $\mathtt{r} \to +\infty$ and $\forall \boldsymbol{\theta} : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}$ it holds

$$-2\{\mathbb{E}L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta}^*)\} \ge \mathtt{r}^2 \mathtt{b}(\mathtt{r}).$$

4.2. *Conditions required for the bootstrap validity.*

(SmB) There exists a constant $\delta^2_{\text{smb}} \in [0, 1/8]$ such that it holds for the matrices H_0^2, B_0^2 defined in (1.7) and (1.8).

$$\|H_0^{-1} B_0^2 H_0^{-1}\| \le \delta^2_{\text{smb}} \le \mathtt{C} p n^{-1/2}.$$

(ED_{2m}) For each $\mathtt{r} > 0$, $i = 1, \dots, n$, $j = 1, 2$ and for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ it holds for the values $\omega \ge 0$ and $\mathtt{g}_2(\mathtt{r})$ from the condition (ED_2)

$$\sup_{\substack{\boldsymbol{\gamma}_j \in \mathbb{R}^p \\ \|\boldsymbol{\gamma}_j\| \le 1}} \log \mathbb{E} \exp\left\{ \frac{\lambda}{\omega} \boldsymbol{\gamma}_1^\top D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 \right\} \le \frac{\nu_0^2 \lambda^2}{2n}, \qquad |\lambda| \le \mathtt{g}_2(\mathtt{r}).$$

$(\mathcal{L}_{0m})$ For each $\mathtt{r} > 0$, $i = 1, \dots, n$ and for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$, there exists a constant $\mathtt{C}_m(\mathtt{r}) \ge 0$ such that

$$\|D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \mathbb{E} \ell_i(\boldsymbol{\theta}) D_0^{-1}\| \le \mathtt{C}_m(\mathtt{r}) n^{-1}.$$

$(\mathcal{I}_B)$ There exists a constant $\mathfrak{a}_B^2 \ge 0$ s.t. $\mathfrak{a}_B^2 D_0^2 \ge B_0^2$.

(SD_1) There exists a constant $0 \le \delta_v^2 \le \mathtt{C} p/n$. such that it holds for all $i = 1, \dots, n$ with exponentially high probability

$$\|H_0^{-1} \{\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top - \mathbb{E}[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top]\} H_0^{-1}\| \le \delta_v^2.$$

(Eb) The bootstrap weights u_i are i.i.d., independent of the data $\mathbf{Y}$, and

$$\mathbb{E} u_i = 1, \qquad \operatorname{Var} u_i = 1,$$
$$\log \mathbb{E} \exp\{\lambda(u_i - 1)\} \le \nu_0^2 \lambda^2 / 2, \qquad |\lambda| \le \mathtt{g}.$$

4.3. *Small modelling bias condition for some models.* Here, we specify the condition (SmB) for some particular models. If the observations $Y_1, \dots, Y_n$ are i.i.d., then $\nabla_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}^*) = n \nabla_{\boldsymbol{\theta}} \mathbb{E} \ell_i(\boldsymbol{\theta}^*) = 0$, and $B_0^2 = 0$. The next example is the generalized linear model: the parametric probability distribution family $\{\mathbb{P}_\upsilon\}$ is an exponential family with a canonical parameterization. The log-density for this family can be expressed as

$$\ell(\upsilon) = y\upsilon - h(\upsilon)$$

for a convex function $h(\cdot)$. Table 4 provides some examples of $\{\mathbb{P}_\upsilon\}$ and $h(\cdot)$. Taking $\{\mathbb{P}_\upsilon\}$ as a parametric family and $\Psi_i^\top \boldsymbol{\theta}$ as linear predictors for some deterministic regressors $\Psi_i \in \mathbb{R}^p$ yields the following quasi log-likelihood function:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \{Y_i \Psi_i^\top \boldsymbol{\theta} - h(\Psi_i^\top \boldsymbol{\theta})\}.$$

TABLE 4
Examples of the GLM

$\mathbb{P}_{\upsilon}$	$h(\upsilon)$	$h'(\upsilon)$ (natural parameter)
$\mathcal{N}(\upsilon, 1)$	$\upsilon^2/2$	υ
$\mathrm{Exp}(-\upsilon)$	$-\log(-\upsilon)$	$-1/\upsilon$
$\mathrm{Pois}(\mathrm{e}^{\upsilon})$	e^{υ}	e^{υ}
$\mathrm{Binom}(1, \frac{\mathrm{e}^{\upsilon}}{\mathrm{e}^{\upsilon}+1})$	$\log(\mathrm{e}^{\upsilon}+1)$	$\frac{\mathrm{e}^{\upsilon}}{\mathrm{e}^{\upsilon}+1}$

It holds

$$\|H_0^{-1} B_0^2 H_0^{-1}\| \le 1 - \min_{1\le i\le n} \frac{\operatorname{Var} Y_i}{\operatorname{Var} Y_i + \{\mathbb{E} Y_i - h'(\Psi_i^\top \boldsymbol{\theta}^*)\}^2} \in [0, 1).$$

It is important that $\mathbb{E}_{\boldsymbol{\theta}^*} Y_i = h'(\Psi_i^\top \boldsymbol{\theta}^*)$, that is, in the case of the correct parametric model $\mathbb{P} \in \{\mathbb{P}_\upsilon\}$ the modelling bias is indeed equal to zero.

Now let us consider the linear quantile regression. Let the observations $Y_1, \dots, Y_n$ be scalar, and the design points $X_1, \dots, X_n$ be deterministic. Let $\tau \in (0, 1)$ denote a fixed known quantile level. The object of estimation is a quantile function $q_\tau(x)$ s.t.

$$\mathbb{P}\bigl(Y_i < q_\tau(X_i)\bigr) = \tau \qquad \forall i = 1, \dots, n.$$

Using the quantile regression approach by Koenker and Bassett (1978), this problem can be treated with the quasi maximum likelihood method and the following log-likelihood function:

$$L(\boldsymbol{\theta}) = -\sum_{i=1}^n \rho_\tau\bigl(Y_i - \Psi_i^\top \boldsymbol{\theta}\bigr), \qquad (4.1)$$
$$\rho_\tau(x) \stackrel{\text{def}}{=} x\bigl(\tau - \mathbb{1}\{x < 0\}\bigr),$$

where $\Psi_i \in \mathbb{R}^p$ are known regressors. This log-likelihood function corresponds to asymmetric Laplace distribution with the density $\tau(1-\tau)\mathrm{e}^{-\rho_\tau(x-a)}$. It holds

$$\|H_0^{-1} B_0^2 H_0^{-1}\| \le 1 - \min_{1\le i\le n} \frac{\operatorname{Var}(\tau - \mathbb{1}\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\})}{\operatorname{Var}(\tau - \mathbb{1}\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\}) + (\tau - \mathbb{P}\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\})^2}.$$

If $\mathbb{P}\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\} \equiv \tau$, then the right-hand side of the last inequality is equal to zero.

4.4. *Dependence of the involved terms on the sample size and parameter dimension.* Here, we consider the case of the i.i.d. observations $Y_1, \dots, Y_n$ and $\mathtt{x} = \mathtt{C}\log n$ in order to specify the dependence of the nonasymptotic bounds on n and p. In Section B.3 of the supplementary material [Spokoiny and Zhilova (2015)], we also consider generalized linear model and quantile regression. Example 5.1 in Spokoiny (2012) demonstrates that in this situation $\mathtt{g} = \mathtt{C}\sqrt{n}$ and $\omega = \mathtt{C}/\sqrt{n}$. This yields $\mathfrak{z}(\mathtt{x}) = \mathtt{C}\sqrt{p+\mathtt{x}}$ for some constant $\mathtt{C} \ge 1.85$, for the function $\mathfrak{z}(\mathtt{x})$ given in (B.4) in Section B.1 of the supplementary material [Spokoiny and Zhilova (2015)]. Similarly, it can be checked that $\mathtt{g}_2(\mathtt{r})$ from condition (ED_2) is proportional to $\sqrt{n}$: due to independence of the observations

$$\begin{aligned}
&\log \mathbb{E}\exp\biggl\{\frac{\lambda}{\omega}\boldsymbol{\gamma}_1^\top D_0^{-1}\nabla^2_{\boldsymbol{\theta}}\zeta(\boldsymbol{\theta})D_0^{-1}\boldsymbol{\gamma}_2\biggr\} \\
&\quad = \sum_{i=1}^n \log \mathbb{E}\exp\biggl\{\frac{\lambda}{\sqrt{n}}\frac{1}{\omega\sqrt{n}}\boldsymbol{\gamma}_1^\top d_0^{-1}\nabla^2_{\boldsymbol{\theta}}\zeta_i(\boldsymbol{\theta})d_0^{-1}\boldsymbol{\gamma}_2\biggr\} \\
&\quad \le n\frac{\lambda^2}{n}\mathtt{C} \qquad \text{for } |\lambda| \le \overline{\mathtt{g}}_2(\mathtt{r})\sqrt{n},
\end{aligned}$$

where $\zeta_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell_i(\boldsymbol{\theta}) - \mathbb{E}\ell_i(\boldsymbol{\theta})$, $d_0^2 \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}}\mathbb{E}\ell_i(\boldsymbol{\theta}^*)$ and $D_0^2 = n d_0^2$ in the i.i.d. case. Function $\overline{\mathtt{g}}_2(\mathtt{r})$ denotes the marginal analog of $\mathtt{g}_2(\mathtt{r})$.

Let us show that for the value $\delta(\mathtt{r})$ from condition $(\mathcal{L}_0)$ it holds $\delta(\mathtt{r}) = \mathtt{C}\mathtt{r}/\sqrt{n}$. Suppose for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ and $\boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\| = 1$ $\|D_0^{-1}\boldsymbol{\gamma}^\top \nabla^3_{\boldsymbol{\theta}}\mathbb{E}L(\boldsymbol{\theta})D_0^{-1}\| \le \mathtt{C}$, then it holds for some $\overline{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})$

$$\begin{aligned}
\|D_0^{-1}D^2(\boldsymbol{\theta})D_0^{-1} - \mathbf{I}_p\| &= \|D_0^{-1}(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top \nabla^3_{\boldsymbol{\theta}}\mathbb{E}L(\overline{\boldsymbol{\theta}})D_0^{-1}\| \\
&= \|D_0^{-1}(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top D_0 D_0^{-1}\nabla^3_{\boldsymbol{\theta}}\mathbb{E}L(\overline{\boldsymbol{\theta}})D_0^{-1}\| \\
&\le \mathtt{r}\|D_0^{-1}\|\|D_0^{-1}\boldsymbol{\gamma}^\top \nabla^3_{\boldsymbol{\theta}}\mathbb{E}L(\overline{\boldsymbol{\theta}})D_0^{-1}\| \le \mathtt{C}\mathtt{r}/\sqrt{n}.
\end{aligned}$$

Similarly, $\mathtt{C}_m(\mathtt{r}) \le \mathtt{C}\mathtt{r}/\sqrt{n} + \mathtt{C}$ 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 is compact and n is sufficiently large, then the value $\mathtt{b}(\mathtt{r})$ from condition $(\mathcal{L}\mathtt{r})$ can be taken as $\mathtt{C}\{1 - \mathtt{r}/\sqrt{n}\} \approx \mathtt{C}$. Indeed, for $\boldsymbol{\theta} : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}$

$$\begin{aligned}
-2\{\mathbb{E}L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta}^*)\} &\ge \mathtt{r}^2\{1 - \mathtt{r}\|D_0^{-1}\|\|D_0^{-1}\boldsymbol{\gamma}^\top \nabla^3_{\boldsymbol{\theta}}\mathbb{E}L(\overline{\boldsymbol{\theta}})D_0^{-1}\|\} \\
&\ge \mathtt{r}^2(1 - \mathtt{C}\mathtt{r}/\sqrt{n}).
\end{aligned}$$

Due to the obtained orders, conditions (B.1) and (B.19) of Theorems B.1 and B.6 (in the supplementary material [Spokoiny and Zhilova (2015)]) on concentration of the MLEs $\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}^{\circ}$ require $\mathtt{r}_0 \ge \mathtt{C}\sqrt{p+\mathtt{x}}$.

Acknowledgments. The authors are very grateful to the anonymous referees for their careful reading of the manuscript and many helpful remarks and suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Bootstrap confidence sets under model misspecification" (DOI: 10.1214/15-AOS1355SUPP; .pdf). The supplementary material contains a proof of the square-root Wilks approximation for the bootstrap world, proofs of the main results from Section 2, and results on Gaussian approximation for ℓ_2-norm of a sum of independent vectors.

REFERENCES

AERTS, M. and CLAESKENS, G. (2001). Bootstrap tests for misspecified models, with application to clustered binary data. *Comput. Statist. Data Anal.* **36** 383–401. MR1843893

ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010). Some nonasymptotic results on resampling in high dimension. I. Confidence regions. *Ann. Statist.* **38** 51–82. MR2589316

BARBE, P. and BERTAIL, P. (1995). *The Weighted Bootstrap. Lecture Notes in Statistics* **98**. Springer, New York. MR2195545

BÜCHER, A. and DETTE, H. (2013). Multiplier bootstrap of tail copulas with applications. *Bernoulli* **19** 1655–1687. MR3129029

CHATTERJEE, S. and BOSE, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.* **33** 414–436. MR2157808

CHEN, X. and POUZO, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *J. Econometrics* **152** 46–60. MR2562763

CHEN, X. and POUZO, D. (2015). Sieve Wald and QLR inferences on semi/nonparametric conditional moment models. *Econometrica* **83** 1013–1079. MR3357484

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York. MR1145237

HALL, A. R. (2005). *Generalized Method of Moments*. Oxford Univ. Press, Oxford. MR2135106

HOROWITZ, J. L. (2001). The bootstrap. *Handbook of Econometrics* **5** 3159–3228.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I: Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

JANSSEN, P. (1994). Weighted bootstrapping of U-statistics. *J. Statist. Plann. Inference* **38** 31–41. MR1256846

JANSSEN, A. and PAULS, T. (2003). How do bootstrap and permutation tests work? *Ann. Statist.* **31** 768–806. MR1994730

KLINE, P. and SANTOS, A. (2012). Higher order properties of the wild bootstrap under misspecification. *J. Econometrics* **171** 54–70. MR2970336

KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644

LAVERGNE, P. and PATILEA, V. (2013). Smooth minimum distance estimation and testing with conditional estimating equations: Uniform in bandwidth theory. *J. Econometrics* **177** 47–59. MR3103915

LIU, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.* **16** 1696–1708. MR0964947

MA, S. and KOSOROK, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *J. Multivariate Anal.* **96** 190–217. MR2202406

MAMMEN, E. (1992). *When Does Bootstrap Work? Lecture Notes in Statistics* **77**. Springer, New York.

MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** 255–285. MR1212176

NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **56** 3–48. MR1257793

SPOKOINY, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.* **40** 2877–2909. MR3097963

SPOKOINY, V. (2013). Bernstein–von Mises theorem for growing parameter dimension. Preprint. Available at arXiv:1302.3430.

SPOKOINY, V. and ZHILOVA, M. (2015). Supplement to "Bootstrap confidence sets under model misspecification." DOI:10.1214/15-AOS1355SUPP.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* **9** 60–62.

WU, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1261–1350. MR0868303

WEIERSTRASS-INSTITUTE
MOHRENSTR. 39
10117 BERLIN
GERMANY
E-MAIL: spokoiny@wias-berlin.de
zhilova@wias-berlin.de