CrossMark

High-dimensional simultaneous inference with the bootstrap

Ruben Dezeure[1] · **Peter Bühlmann**[1] ·
Cun-Hui Zhang[2]

Abstract We propose a residual and wild bootstrap methodology for individual and simultaneous inference in high-dimensional linear models with possibly non-Gaussian and heteroscedastic errors. We establish asymptotic consistency for simultaneous inference for parameters in groups G, where $p \gg n$, $s_0 = o(n^{1/2}/\{\log(p) \log(|G|)^{1/2}\})$ and $\log(|G|) = o(n^{1/7})$, with p the number of variables, n the sample size and s_0 the sparsity. The theory is complemented by many empirical results. Our proposed procedures are implemented in the R-package hdi (Meier et al. hdi: high-dimensional inference. R package version 0.1-6, 2016).

Keywords De-biased Lasso · De-sparsified Lasso · Gaussian approximation for maxima · High-dimensional linear model · Heteroscedastic errors · Multiple testing · Westfall–Young method

This invited paper is discussed in comments available at: doi:10.1007/s11749-017-0555-1; doi:10.1007/s11749-017-0556-0; doi:10.1007/s11749-017-0557-z; doi:10.1007/s11749-017-0558-y; doi:10.1007/s11749-017-0559-x.

Electronic supplementary material The online version of this article (doi:10.1007/s11749-017-0554-2) contains supplementary material, which is available to authorized users.

✉ Peter Bühlmann
buhlmann@stat.math.ethz.ch

[1] Seminar for Statistics, ETH Zürich, HG G 17, Rämistrasse 101, 8092 Zurich, Switzerland

[2] Department of Statistics and Biostatistics, Rutgers University, 569 Hill Center, Busch Campus, Piscataway, NJ 08854-8019, USA

⌂ Springer

Mathematics Subject Classification 62J07 · 62F40

1 Introduction

Recently, there has been growing interest for statistical inference, hypothesis tests and confidence regions in high-dimensional models. In fact, many applications nowadays involve high-dimensional models, and thus, accurate statistical inference methods and tools are very important. For general models and high-dimensional settings, sample splitting procedures (Wasserman and Roeder 2009; Meinshausen et al. 2009) and stability selection (Meinshausen and Bühlmann 2010; Shah and Samworth 2013) provide some statistical error control and significance. For the case of a linear model with homoscedastic and Gaussian errors, more recent and powerful techniques have been proposed (Bühlmann 2013; Zhang and Zhang 2014; van de Geer et al. 2014; Javanmard and Montanari 2014; Meinshausen 2015; Foygel Barber and Candès 2015) and some of these extend to generalized linear models. For a recent overview, see also Dezeure et al. (2015).

We focus in this paper on a linear model

$$Y = \mathbf{X}\beta^0 + \varepsilon,$$

where we use the notation Y for the $n \times 1$ response variable, $\mathbf{X}$ for the $n \times p$ design matrix, β^0 for the vector of unknown true regression coefficients, and ε for the errors; for more assumptions, see (1). One goal is to construct confidence intervals for individual coefficients β_j^0, for $j \in \{1, \ldots, p\}$, or corresponding statistical hypothesis tests of the form

$$H_{0,j} : \beta_j^0 = 0 \text{ versus the alternative } H_{A,j} : \beta_j^0 \neq 0 \ (j = 1, \ldots, p).$$

More generally, for groups $G \subseteq \{1, \ldots, p\}$ of variables, we consider

$$H_{0,G} : \beta_j^0 = 0 \text{ versus the alternative } H_{A,G} : \beta_j^0 \neq 0 \text{ for some } j \in G,$$

and of particular interest is also multiple testing adjustment when testing many individual or group hypotheses.

In this work we will argue that the bootstrap is very useful for individual and especially for simultaneous inference in high-dimensional linear models, that is for testing individual or group hypotheses $H_{0,j}$ or $H_{0,G}$, and for corresponding individual or simultaneous confidence regions. We thereby also demonstrate its usefulness to deal with potentially heteroscedastic or non-Gaussian errors. Instead of bootstrapping the Lasso estimator directly (see also the comment in Sect. 1.1), we propose to bootstrap the de-biased (Zhang and Zhang 2014) or de-sparsified Lasso which is a regular non-sparse estimator achieving asymptotic efficiency under certain assumptions (van de Geer et al. 2014). The idea of bootstrapping an estimator like the de-sparsified Lasso has been proposed in Belloni et al. (2015b), building on the fundamental work of Chernozhukov et al. (2013) for bootstrapping approximately linear estimators in high-dimensional problems: the relation of our work with existing results is described

in more detail at the end of Sect. 1.1. We discuss several advantages of bootstrapping the de-sparsified Lasso, including the issue of simultaneous inference for large groups of variables and statistically efficient multiple testing adjustment. These make our bootstrap approach a "state-of-the-art tool" for reliable inference in high-dimensional linear models with potentially heteroscedastic and strongly non-Gaussian errors. The resampling nature in general should further contribute additional stability and robustness to statistical results and conclusions, cf. Breiman (1996).

From a computational point of view, the bootstrap scheme is feasible and not substantially more expensive than the de-sparsified Lasso itself; especially when the number of variables is large, the extra cost of bootstrapping is not very severe. The bootstrap procedures which we propose and discuss are implemented and added to the R-package hdi (Meier et al. 2016). This supports their use for practical analysis of high-dimensional data.

1.1 Related work and our contribution

Besides the growing literature in assessing uncertainty in high-dimensional statistical inference mentioned at the beginning of the introductory section, the use of the bootstrap has been advocated in other works. In particular, the contributions by Belloni et al. (2015a, b) and Zhang and Cheng (2016) are closely related to ours: More details are given below. From a theoretical perspective, the results from Chernozhukov et al. (2013) are important for deriving results for simultaneous inference based on the bootstrap.

Bootstrapping the adaptive Lasso in high-dimensional linear models has been put forward and analyzed by Chatterjee and Lahiri (2011, 2013). A main difference to our proposal is that their approach is for a sparse Lasso-type estimator and they require a "beta-min" condition (saying that all nonzero regression coefficients are sufficiently large in absolute value) to ensure that the bootstrap captures the correct limiting distribution for the nonzero parameters. A related approach has been proposed by Liu and Yu (2013). Instead of the adaptive Lasso, a two-stage and sparse estimator based on Lasso selection (stage 1) and least squares or Ridge estimation using the selected variables (stage 2) is considered. A residual bootstrap (based on the two-stage procedure) is then employed to bootstrap this two-stage sparse estimator. We avoid a "beta-min" assumption because it is a main purpose of the inference method itself to find out which of the underlying regression coefficients are sufficiently large or not. Furthermore, from a practical perspective, bootstrapping a Lasso-type or other sparse (e.g., as in Liu and Yu 2013) estimator can be severely exposed to the super-efficiency phenomenon. It has been reported in numerical simulation studies, saying that inference for nonzero regression coefficients can be very poor (Dezeure et al. 2015). The bootstrap has also been used and studied in settings which are vaguely related to ours: Zhou (2014) presents an MCMC sampler for the distribution of an augmented Lasso estimator which allows for some inferential tasks, McKeague and Qian (2015) analyze the bootstrap for marginal correlation screening for high-dimensional linear models, and Shah and Bühlmann (2015) consider the use a bootstrap scheme for obtaining the exact distribution of scaled residuals in a high-dimensional linear model with Gaus-

sian errors, which in turn enables inference about the distribution for any estimator or function based on the scaled residuals.

We use the following abbreviations: work by Belloni et al. (2015b) *(BCK)*, Belloni et al. (2015a) *(BCCW) and* Zhang and Cheng (2016) *(ZC)*. These papers have analyzed the bootstrap for the linearized part of similar estimators like the de-sparsified Lasso. BCK (Belloni et al. 2015b) has been the first paper which considered this problem for a Gaussian multiplier bootstrap using results from Chernozhukov et al. (2013) for the Gaussian multiplier bootstrap for approximate means of random variables. The latter results have been improved in CCK (Chernozhukov et al. 2014), establishing faster rates and encompassing also the classical bootstrap based on the empirical distribution. The paper by BCCW (Belloni et al. 2015a) using a Gaussian multiplier bootstrap can be seen as an extension of BCK to many functional "Z-problems" (i.e., moment condition model), and software implementation is available from the R-package hdm (Chernozhukov et al. 2016). The work by ZC (Zhang and Cheng 2016) is rather similar to BCK. We provide here some bootstrap approaches for the more general case with heteroscedastic errors and exhibiting better finite sample performance, as explained next.

We discuss three different bootstrap methods: a residual bootstrap, a multiplier wild bootstrap and a special version of a paired bootstrap method, whereas BCK, BCCW and ZC consider a Gaussian multiplier wild bootstrap only. Our different procedures are motivated and carefully discussed from the viewpoint of dealing with heteroscedastic errors, while all other works consider homoscedastic errors only. We also allow for non-Gaussian multipliers in the wild bootstrap and develop the corresponding theory, motivated by potential advantages in performance (Mammen 1993) [although we could not find substantial empirical improvements in the setting of high-dimensional regression, in contrast to high-dimensional means (Deng and Zhang 2017)]. This is in contrast to BCK, BCCW and ZC who consider Gaussian multipliers only and directly rely on results from Chernozhukov et al. (2013) for the Gaussian multiplier bootstrap. (Extensions to the classical bootstrap might be rather straightforward using the results from Chernozhukov et al. 2014.)

As a methodological difference, we advocate to bootstrap the entire de-sparsified Lasso estimator, using the plug-in rule, whereas BCK, BCCW and ZC only bootstrap the linearized part of the estimator. As mentioned above, there is essentially no additional computational cost when bootstrapping the entire instead of only the linearized part of the estimator. In the presented theories, there is no need to bootstrap the nonlinear, asymptotically negligible, part of the estimator: Finite sample results though speak much in favor to bootstrap the entire estimator (as we propose here), see Sect. 5. Bootstrapping the entire procedure also makes unnecessary the version "RLDPE" (restricted low-dimensional projection estimator) of the de-sparsified estimator which was introduced by Zhang and Zhang (2014) to improve coverage of nominal confidence while paying a price for efficiency; see Sects. 5.1.1 and 5.1.2.

Regarding theory, our condition on the sparsity of the design is much weaker than in ZC. We require an ℓ_1-norm condition for the rows of the inverse covariance matrix, while they require a much more stringent ℓ_0-sparsity condition. The details are as follows: We require an ℓ_1-norm condition in the second part of (B2) which is implied

by the ℓ_0-sparsity condition $s_j = o(n/\log(p))$, where $s_j = \sum_{k \neq j} I\left(\left(\Sigma_X^{-1}\right)_{jk} \neq 0\right)$, due to $\|\gamma_j\|_1 \leq O(1)\sqrt{s_j}$ when $\lambda_{\min}(\Sigma) > c > 0$. In contrast, ZC require $s_j = o(\sqrt{n/\log(p)})$. For details of notation, see Sect. 3.3.

Our contribution here can be seen as a very general development of bootstrap methods for the de-biased or de-sparsified Lasso for confidence intervals and hypotheses testing in high-dimensional linear models with potentially heteroscedastic and non-Gaussian errors, with a particular emphasis on simultaneous inference and multiple testing adjustment. Our aim is to establish, by theory and empirical results, the practical usefulness and reliability of the bootstrap for high-dimensional inference.

2 High-dimensional linear model and the de-sparsified Lasso

We consider in this work a high-dimensional linear model

$$Y = \mathbf{X}\beta^0 + \varepsilon, \tag{1}$$

with $n \times 1$ response vector Y, $n \times p$ fixed design matrix $\mathbf{X}$, $p \times 1$ vector β^0 of the true underlying unknown regression parameters, and $n \times 1$ vector of error terms. The $n \times 1$ column vectors of $\mathbf{X}$ are denoted by X_j ($j = 1, \ldots, p$). The errors are assumed to be independent with mean $\mathbb{E}[\varepsilon_i] = 0$, but potentially heteroscedastic with variances $\mathbb{E}[\varepsilon_i^2] = \sigma_i^2$. We note that the case of fixed design arises when conditioning on the covariables. We focus on the high-dimensional regime where the dimension $p \gg n$ is much larger than sample size n. Then, the linearity itself is not a real restriction, as discussed in Sect. 6.1. The goal in this paper is inference for the unknown parameter vector β^0, in particular in terms of statistical hypothesis tests and confidence intervals.

We propose to do such inference based on non-sparse estimators. The non-sparsity of an estimator typically induces "regularity" and avoids the phenomenon of super-efficiency: We believe that this classical viewpoint (cf. Bickel et al. 1998) is actually important and leads to much better performance for constructing confidence intervals for nonzero parameters. Regularity typically enables asymptotic normality and efficiency, and it is also advantageous for consistency of the bootstrap due to fundamental results by Giné and Zinn (1989) and Giné and Zinn (1990).

2.1 The de-sparsified Lasso

The de-biased Lasso (Zhang and Zhang 2014), also called the de-sparsified Lasso (van de Geer et al. 2014), can be considered as a generalization of the ordinary least squares approach to the high-dimensional setting.

In the low-dimensional $p < n$ setting with $\mathbf{X}$ having full rank p, denote by V_j the residual vector when doing an ordinary least squares regression of X_j versus $\mathbf{X}_{-j}$: Here $\mathbf{X}_{-j}$ is the $n \times (p-1)$ sub-matrix of $\mathbf{X}X$ without the jth column. Then, the ordinary least squares estimator for β^0 can be written as

$$\hat{\beta}_j^{\text{OLS}} = \frac{V_j^T Y}{V_j^T X_j}.$$

When $p > n$, the V_j's are zero vectors and we cannot use such a construction. Instead, we consider the residuals Z_j from a Lasso regression of X_j versus all other variables in $\mathbf{X}_{-j}$:

$$\hat{\gamma}_j = \hat{\gamma}_j(\lambda_X) = \mathrm{argmin}_{\gamma_j} \left(\|X_j - \mathbf{X}_{-j}\gamma_j\|_2^2/n + \lambda_X \|\gamma_j\|_1 \right),$$
$$Z_j = X_j - \mathbf{X}_{-j}\hat{\gamma}_j.$$

We then project on these regularized residuals while introducing a bias:

$$\hat{\beta}_j' = \frac{Z_j^T Y}{Z_j^T X_j} = \beta_j^0 + \sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \beta_k^0 + \frac{Z_j^T \varepsilon}{Z_j^T X_j}.$$

The introduced bias $\sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \beta_k^0$ can be estimated and corrected for by plugging in the Lasso from a regression of Y versus $\mathbf{X}$:

$$\hat{\beta} = \hat{\beta}(\lambda) = \mathrm{argmin}_{\beta} \left(\|Y - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1 \right).$$

This gives us the de-biased or de-sparsified Lasso:

$$\hat{b}_j = \hat{\beta}_j' - \sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \hat{\beta}_k = \beta_j^0 + \sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \left(\beta_k^0 - \hat{\beta}_k \right) + \frac{Z_j^T \varepsilon}{Z_j^T X_j}. \qquad (2)$$

The estimator $\hat{b}_j$ is not sparse anymore, and hence the name de-sparsified Lasso (van de Geer et al. 2014); we can also write it as

$$\hat{b}_j = \hat{\beta}_j + \frac{Z_j^T (Y - \mathbf{X}\hat{\beta})}{Z_j^T X_j},$$

which means that it equals the Lasso plus a one-step bias correction, and hence the alternative name de-biased Lasso (Zhang and Zhang 2014). In the sequel, we use the terminology de-sparsified Lasso.

When interested in all $j = 1, \ldots, p$, the procedure requires one to run the Lasso with tuning parameter λ for the regression of Y versus $\mathbf{X}$, and the nodewise Lasso (Meinshausen and Bühlmann 2006) which means the Lasso for every regression of X_j versus $\mathbf{X}_{-j}$ ($j = 1, \ldots, p$) with tuning parameter λ_X (the same for all j). The total computational requirement is thus to run $p + 1$ Lasso regressions which can be substantial if p is large. Luckily, parallel computation can be done very easily, as implemented in hdi (Meier et al. 2016; Dezeure et al. 2015).

It has been shown first by Zhang and Zhang (2014), for homoscedastic errors, that under some conditions,

$$\left(\hat{b}_j - \beta_j^0 \right) / \text{s.e.}_j \Rightarrow \mathcal{N}(0, 1) \ (j = 1, \ldots, p), \qquad (3)$$

with the approximate standard error given in Theorem 1 or Theorem 2 for the case of homoscedastic or heteroscedastic errors, respectively. The convergence is understood as both $p \geq n \to \infty$. For the homoscedastic case, the asymptotic variance reaches the semiparametric information bound (van de Geer et al. 2014).

Estimation of the standard error is discussed below in Sect. 2.2. With an approximate pivot at hand, we can construct confidence intervals and hypothesis tests: For homoscedastic errors, this has been pursued by various authors and Dezeure et al. (2015) present a review and description of how inference based on such pivots can be done with the R-package hdi (Meier et al. 2016).

In this work we will argue that bootstrapping the de-sparsified Lasso $\hat{b}$ will bring additional benefits over the asymptotic inference based on a Gaussian limiting distribution arising in (3).

2.2 Estimation of the standard error and robustness for heteroscedastic errors

Based on the developed theoretical results in Sect. 3.3, one can show that the asymptotic standard error of the de-sparsified estimator behaves like

$$\text{s.e.}_j = n^{-1/2} \frac{\sqrt{\text{Var}(n^{-1/2} \sum_{i=1}^{n} Z_{j;i} \varepsilon_i)}}{|Z_j^T X_j / n|}.$$

For the case of homoscedastic i.i.d. errors with $\text{Var}(\varepsilon_i) = \sigma_\varepsilon^2$, the inverse of the standard error is then asymptotically behaving like

$$\text{s.e.}_j = n^{-1/2} \frac{\sigma_\varepsilon \|Z_j\|_2 / \sqrt{n}}{|Z_j^T X_j / n|}.$$

This suggests to use as an estimate

$$\widehat{\text{s.e.}}_j = n^{-1/2} \frac{\hat{\sigma}_\varepsilon \|Z_j\|_2 / \sqrt{n}}{|Z_j^T X_j / n|},$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n - \hat{s}} \|Y - \mathbf{X}\hat{\beta}\|_2^2, \tag{4}$$

with $\hat{s}$ the number of nonzero coefficients in the estimate $\hat{\beta}$. This choice of $\hat{\sigma}_\varepsilon^2$ is based on the recommendation of Reid et al. (2016) and supported by our own empirical experience with different variance estimators. This standard error estimate is implemented in the R-package hdi (Meier et al. 2016).

For heteroscedastic but independent errors with $\text{Var}(\varepsilon_i) = \sigma_i^2$, the asymptotic standard error behaves as

$$\text{s.e.}_{\text{robust}, j} = n^{-1/2} \frac{\omega_j}{|Z_j^T X_j / n|},$$

$$\omega_j^2 = n^{-1} \sum_{i=1}^{n} Z_{j;i}^2 \sigma_i^2.$$

We then propose the robust estimator

$$\widehat{s.e.}_{\text{robust},j} = n^{-1/2} \frac{\hat{\omega}_j}{|Z_j^T X_j/n|},$$

$$\hat{\omega}_j^2 = \frac{1}{n-\hat{s}} \sum_{i=1}^{n} \left(\hat{\varepsilon}_i Z_{j;i} - n^{-1} \sum_{r=1}^{n} \hat{\varepsilon}_r Z_{j;r} \right)^2, \quad \hat{\varepsilon} = Y - X\hat{\beta}, \tag{5}$$

which has been used in Bühlmann and van de Geer (2015) for the different contexts of misspecified linear models with random design. We prove that under some conditions, $\widehat{s.e.}_j / \text{Var}(\hat{b}_j)^{1/2} = 1 + o_P(1)$ (Theorem 1 for the homoscedastic case) and $\widehat{s.e.}_{\text{robust},j} / \text{Var}(\hat{b}_j)^{1/2} = 1 + o_P(1)$ (Theorem 2 for the heteroscedastic case). In fact, the robust standard error estimator is consistent for both the homo- and heteroscedastic cases for the error terms: Therefore, it is robust against heteroscedasticity which explains its name. The phenomenon is closely related to the robust sandwich estimator for the standard error of the MLE in low-dimensional models (Eicker 1967; Huber 1967; White 1980; Freedman 1981).

We point out that the result

$$\left(\hat{b}_j - \beta_j^0 \right) / \widehat{s.e.}_{\text{robust},j} \Longrightarrow \mathcal{N}(0,1),$$

presented later in Theorem 2, is a new extension which covers the case with heteroscedastic errors. All that is conceptually needed is the robust standard error estimate $\widehat{s.e.}_{\text{robust},j}$.

3 Bootstrapping the de-sparsified Lasso

We consider first a residual bootstrap procedure. Two alternative bootstrap methods are discussed in Sects. 4.1 and 4.2. We use the Lasso for computing residuals $\hat{\varepsilon} = Y - X\hat{\beta}$ and centered residuals $\hat{\varepsilon}_{\text{cent},i} = \hat{\varepsilon}_i - \bar{\hat{\varepsilon}}$ $(i = 1, \ldots, n)$, where $\bar{\hat{\varepsilon}} = n^{-1} \sum \hat{\varepsilon}_i$. The bootstrapped errors are then constructed from the

Residual bootstrap:

$\varepsilon_1^*, \ldots, \varepsilon_n^*$ i.i.d. (re-)sampled from the centered residuals $\hat{\varepsilon}_{\text{cent},i}$ $(i = 1, \ldots, n)$.

We then construct the bootstrapped response variables as

$$Y^* = X\hat{\beta} + \varepsilon^*, \tag{6}$$

and the bootstrap sample is $\{(X_i, Y_i^*)\}_{i=1}^{n}$, reflecting the fact of fixed (non-random) design. Here and in the sequel X_i denotes the $p \times 1$ row vectors of X $(i = 1, \ldots, n)$.

3.1 Individual inference

We aim to estimate the distribution of the asymptotic pivot (see Theorems 1 and 2)

$$T_j = \frac{\hat{b}_j - \beta_j^0}{\widehat{s.e.}_{\text{robust},j}}, \tag{7}$$

where $\hat{b}_j$ is the de-sparsified estimator and $\widehat{s.e.}_{\text{robust},j}$ is the robust standard error in (5). We propose to always use this robust standard error in practice because it automatically provides protection (robustness) against heteroscedastic errors. At some places, we also discuss the use of the more usual standard error formula $\widehat{s.e.}_j$ from (4) for the case with homoscedastic errors: But this serves mainly for explaining some conceptual differences. For estimating the distribution in (7), we use the bootstrap distribution of

$$T_j^* = \frac{\hat{b}_j^* - \hat{\beta}_j}{\widehat{s.e.}_{\text{robust},j}^*}, \tag{8}$$

where $\hat{b}_j^*$ and $\widehat{s.e.}_{\text{robust},j}^*$ are computed by plugging in the bootstrap sample instead of the original data points. (Alternatively, when using the non-robust standard error, we would also use the bootstrap for the non-robust version.) Denote by $q_{j;v}^*$ the v-quantile of the bootstrap distribution of T_j^*. We then construct two-sided $100(1-\alpha)\%$ confidence intervals for the jth coefficient β_j^0 as

$$\text{CI}_j = \left[\hat{b}_j - q_{j;1-\alpha/2}^* \, \widehat{s.e.}_{\text{robust},j}, \hat{b}_j - q_{j;\alpha/2}^* \, \widehat{s.e.}_{\text{robust},j} \right]. \tag{9}$$

Corresponding p values for the null hypothesis $H_{0,j}$ versus the two-sided alternative $H_{A,j}$ can then be computed by duality. Bootstrapping pivots in classical low-dimensional settings are known to improve the level of accuracy of confidence intervals and hypothesis tests (Hall and Wilson 1991).

3.2 Simultaneous confidence regions, intervals and p values for groups

We can construct simultaneous confidence regions over a group of variables G. Rather than using the sup-norm, we build the region

$$C(1-\alpha) = \left\{ b \in \mathbb{R}^p; \max_{j \in G} \frac{\sqrt{n}(\hat{b}_j - b_j)}{\widehat{s.e.}_{\text{robust},j}} \right.$$
$$\left. \leq q_{\text{max};G}^*(1-\alpha/2) \text{ and } \min_{j \in G} \frac{\sqrt{n}(\hat{b}_j - b_j)}{\widehat{s.e.}_{\text{robust},j}} \geq q_{\text{min};G}^*(\alpha/2) \right\},$$

where $q_{\text{max};G}^*(v)$ is the v-quantile of the bootstrap distribution of $\max_{j \in G} T_j^*$ and $q_{\text{min};G}^*(v)$ the v-quantile of the bootstrap distribution of $\min_{j \in G} T_j^*$, respectively (with

T_j^* as in (8)). If the group G is large, a more informative view is to take the componentwise version of $C(1-\alpha)$: For each component $j \in G$ we consider the confidence interval for β_j^0 of the form

$$\text{CI}_{\text{simult},j} = \left[\hat{b}_j - \widehat{\text{s.e.}}_{\text{robust},j}q^*_{\text{max};G}(1-\alpha/2), \hat{b}_j - \widehat{\text{s.e.}}_{\text{robust},j}q^*_{\text{min};G}(\alpha/2)\right]. \quad (10)$$

We may also replace $q^*_{\text{max};G}(1-\alpha/2)$ and $q^*_{\text{min};G}(\alpha/2)$ by $\pm q^*_{\text{abs};G}(1-\alpha)$, where $q^*_{\text{abs};G}(\nu)$ is the ν-quantile of the bootstrap distribution of $\max_{j \in G}|T_j^*|$. In contrast to the confidence intervals in (9), the intervals in (10) are simultaneous and hence wider, providing approximate coverage in the form of

$$\mathbb{P}[\beta_j^0 \in \text{CI}_{\text{simult},j} \text{ for all } j \in G] \approx 1-\alpha.$$

Of particular interest is the case with $G = \{1, \ldots, p\}$. This construction often provides shorter intervals than using a Bonferroni correction, especially in the presence of positive dependence. See also the empirical results in Sect. 5.2 for the related problem of adjustment for multiple testing.

We might also be interested in p values for testing the null hypothesis

$$H_{0,G} : \beta_j^0 = 0 \text{ for all } j \in G,$$

against the alternative $H_{A,G} : \beta_j^0 \neq 0$ for some $j \in G$. We consider the max-type statistics $\max_{j \in G}|T_j|$ which should be powerful for detecting sparse alternatives. We can use the bootstrap under $H_{0,G}$, or alternatively under the complete null hypothesis, $H_{0,\text{complete}} : \beta_j^0 = 0 \ \forall j = 1, \ldots, p$, by exploiting (asymptotic) restricted subset pivotality. The details are given in Sect. 4.3. Resampling under $H_{0,\text{complete}}$ is computationally much more attractive when considering many groups since we can use the same bootstrap distribution to compute the p values for many groups. The p value is then given by

$$P_G = \mathbb{P}^{*0}\left[\max_{j \in G}|T_j^{*0}| > \max_{j \in G}|t_j|\right],$$

where the asterisk "*0" emphasizes that the bootstrap is constructed under the complete null hypothesis $H_{0,\text{complete}}$ and t_j is the observed realized value of the studentized statistics T_j.

In the presence of heteroscedasticity, the residual bootstrap is inconsistent for simultaneous inference, and the wild bootstrap or a paired bootstrap scheme described in Sects. 4.1 and 4.2 should be used instead.

3.3 Consistency of the residual bootstrap

For deriving the asymptotic consistency of the bootstrap, we make the following assumptions.

(A1) $\|\hat{\beta} - \beta^0\|_1 = o_P(1/\sqrt{\log(p)\log(1 + |G|)})$.

(A2) $\lambda_X \asymp \sqrt{\log(p)/n}$, $\|Z_j\|_2^2/n \geq L_Z$, $\|Z_j\|_{2+\delta}^{2+\delta} = o(\|Z_j\|_2^{2+\delta})$, $j \in G$.

(A3) $\varepsilon_1, \ldots, \varepsilon_n$ independent, $\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}\|\varepsilon\|_2^2/n = \sigma_\varepsilon^2$, $L \leq \mathbb{E}|\varepsilon_i|^2 = \sigma_i^2$, $\mathbb{E}|\varepsilon_i|^{2+\delta} \leq C$, for all i.

(A4) $\|\hat{\beta}^* - \hat{\beta}\|_1 = o_{P*}(1/\sqrt{\log(p)\log(1 + |G|)})$ in probability.

(A5) $\max_{ij} |X_{ij}| \leq C_X$.

(A6) $\max_{j \in G} \|Z_j\|_\infty \leq K$, $\delta = 2$, i.e., bounded fourth moment of ε, $\log(|G|) = o(n^{1/7})$.

Here $\sigma_\varepsilon, \delta, L, C, C_X, L_Z$ and K are positive constants uniformly bounded away from 0 and ∞, and $G \subseteq \{1, \ldots, p\}$ indicates a set of variables of interest, e.g., $G = \{j\}$ for inference of a single β_j. As our theoretical results require no more than the fourth moment of ε, we set $\delta \in (0, 2]$ for simplicity without loss of generality. The constant δ is the same in (A2), (A3) and (A6), e.g., $\delta = 2$ in (A3) when (A6) is imposed. Unless otherwise stated, (A2) is imposed with an arbitrarily small $\delta > 0$ when $|G| = O(1)$ and strengthened with (A6) when $|G| \to \infty$.

Justification of (A1), (A2), (A4) and (A6). Sufficient assumptions for (A1), (A2), (A4) and (A6) (and choosing $\lambda_X \asymp \sqrt{\log(p)/n}$) are as follows.

(B1) The rows of the design matrix are i.i.d. realizations from a distribution with covariance matrix Σ_X, and the smallest eigenvalue of Σ_X is larger than some $M > 0$. Furthermore, for some constants $C_1, C_2, 0 < C_1 \leq \tau_j^2 = 1/(\Sigma_X^{-1})_{jj} \leq C_2 < \infty$.

(B2) $s_0 = o(\sqrt{n}/\{\log(p)\sqrt{\log(|G|)}\})$, $\sum_{k \neq j} |(\Sigma_X^{-1})_{jk}| \leq o(\sqrt{n/\log p})$.

(B3) The smallest sparse eigenvalue of $\mathbf{X}^T\mathbf{X}/n$, with sparsity of the order s_0, is bounded from below by a positive constant.

Assumptions (B1, only the first requirement), (B2, only the first requirement) and (A5) imply that with high probability (w.r.t. i.i.d. sampling the rows of the design matrix), (B3) and the compatibility condition for the set S_0 hold (Bühlmann and van de Geer 2011, Cor.6.8). Alternatively, by Maurey's empirical method (Rudelson and Zhou 2013), (B3) and (A5) directly imply the compatibility condition for deterministic design. It is known (Bühlmann and van de Geer 2011, Th.6.1 and Ex14.3) that with the compatibility condition for S_0 and $\lambda \geq 2\|\mathbf{X}^T\varepsilon/n\|_\infty$ we have that $\|\hat{\beta} - \beta^0\|_1 = O_P(s_0\sqrt{\log(p)/n})$ and thus, (B2) implies (A1).

Let γ_j^0 be the population regression coefficients of X_j versus $\mathbf{X}_{-j}$ and $Z_j^0 = X_j - \sum_{k \neq j} X_k(\gamma_j^0)_k$. By Nemirovski's inequality, (B1) and (A5) imply $2 \max_{k \neq j} |X_k^T Z_j^0/n| \leq \lambda_X$ with large probability for a certain $\lambda_X = O_P(\sqrt{\log(p)/n})$. For such λ_X, the second part of (B2) implies

$$\|Z_j - Z_j^0\|_2^2/n + 2^{-1}\lambda_X\|\hat{\gamma}_j\|_1 \leq (3/2)\lambda_X\|\gamma_j^0\|_1 = o(1).$$

As $\|Z_j^0\|_\infty \leq C_X(1 + \|\gamma_j^0\|_1) = o(\sqrt{n/\log p})$ by (A5), the Bernstein inequality gives

$$\max_{j \leq p} \left| \tau_j^2 - \|Z_j^0\|_2^2/n \right| = o_P(1).$$

Thus, due to the second part of (B1) we have proved the requirement on $\|Z_j\|_2^2/n$ in (A2) and (A6). Moreover, as $\|Z_j - Z_j^0\|_{2+\delta} \le \|Z_j - Z_j^0\|_2 = o(n^{1/2})$,

$$\|Z_j\|_{2+\delta}^{2+\delta} \le 2^{1+\delta}\Big(\|Z_j^0\|_2^2\|Z_j^0\|_\infty^\delta + \|Z_j - Z_j^0\|_2^{2+\delta}\Big) \ll n^{1+\delta/2} \asymp \|Z_j\|_2^{2+\delta},$$

which proves the last statement in (A2). If the second requirement of (B2) is strengthened to $\max_{j \le p} \|\gamma_j^0\|_1 = C_\Sigma$, the ℓ_∞ bound in (A6) follows from $\|Z_j\|_\infty \le (1 + \|\hat{\gamma}_j\|_1)C_X \le (1 + 3C_\Sigma)C_X$.

Assumption (A4) holds when assuming (B1, only the first requirement), (B2, only the first requirement) and (A5) (and these assumptions imply the compatibility condition as mentioned earlier), ensuring that $\hat{s}_0 = \|\hat{\beta}\|_0 = O_P(s_0) = o_P(\sqrt{n}/\log(p))$. The latter holds under a sparse eigenvalue condition on the design (Zhang and Huang 2008) or when using, e.g., the adaptive or thresholded Lasso in the construction of the bootstrap samples (van de Geer et al. 2011) and (Bühlmann and van de Geer 2011, Ch.7.8-7.9).

3.3.1 Homoscedastic errors

The bootstrap is used to estimate the distribution of the studentized statistic

$$\left(\hat{b}_j - \beta_j^0\right)/\widehat{\text{s.e.}}_j,$$

$$1/\widehat{\text{s.e.}}_j = \sqrt{n}\frac{|Z_j^T X_j/n|}{\hat{\sigma}_\varepsilon\|Z_j\|_2/\sqrt{n}},$$

where $\widehat{\text{s.e.}}_j$ is the approximate standard error for $\hat{b}_j$ when the Lasso is nearly fully de-biased, with the estimated standard deviation of the error.

Theorem 1 *Assume (A1)–(A5) with common $\mathbb{E}\,\varepsilon_i^2 = \sigma_\varepsilon^2$ throughout the theorem. Let $\mathbb{P}^*$ represent the residual bootstrap. Then,*

$$T_j = \left(\hat{b}_j - \beta_j^0\right)/\widehat{\text{s.e.}}_j \xRightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

$$T_j^* = \left(\hat{b}_j^* - \hat{\beta}_j\right)/\widehat{\text{s.e.}}_j^* \xRightarrow{\mathcal{D}^*} \mathcal{N}(0,1) \text{ in probability},$$

for each $j \in G$; here, "$\mathcal{D}$" and "$\mathcal{D}^*$" denote convergence in distribution with respect to the original and to the bootstrap measure, respectively. If $|G| = O(1)$, then

$$\sup_{(t_j, j \in G)}\left|\mathbb{P}^*\left[T_j^* \le t_j, j \in G\right] - \mathbb{P}\left[T_j \le t_j, j \in G\right]\right| = o_P(1).$$

If (A6) holds, then

$$\sup_{c \in \mathbb{R}}\left|\mathbb{P}^*\left[\max_{j \in G} h\left(T_j^*\right) \le c\right] - \mathbb{P}\left[\max_{j \in G} h\left(T_j\right) \le c\right]\right| = o_P(1)$$

for $h(t) = t$, $h(t) = -t$ and $h(t) = |t|$.

A proof is given in the Electronic Supplementary Material. We note that Theorem 1 only requires a weak form of homoscedasticity in the sense of equal variance, instead of the stronger assumption of equal distribution, and that under this weak homoscedasticity, the original and the bootstrap distributions have asymptotically the same (estimated) standard errors

$$\widehat{\text{s.e.}}_{\cdot j} \sim \sqrt{\text{Asymp.Var}(\hat{b}_j)} \sim \sqrt{\text{Asymp.Var}^*(\hat{b}_j^*)} \sim \widehat{\text{s.e.}}_{\cdot j}^*,$$

where we omit that the statements are with high probability (in P^* and/or in P). See also after the proof of Theorem 1 in the Electronic Supplementary Material.

3.3.2 Heteroscedastic errors

Consider the inverse of the robust standard error formula:

$$1/\widehat{\text{s.e.}}_{\text{robust},j} = \sqrt{n} \frac{|Z_j^T X_j / n|}{\hat{\omega}_j},$$

$$\hat{\omega}_j^2 = n^{-1} \sum_{i=1}^{n} \left(\hat{\varepsilon}_i Z_{j;i} - n^{-1} \sum_{r=1}^{n} \hat{\varepsilon}_r Z_{j;r} \right)^2.$$

For deriving the consistency of the bootstrap in the presence of heteroscedastic errors, we remove the homoscedasticity assumption on the variance, $\mathbb{E}\varepsilon_i^2 = \sigma_\varepsilon^2$, imposed in Theorem 1.

Theorem 2 *Assume (A1)–(A5). Let $\mathbb{P}^*$ represent the residual bootstrap. Then, for each $j \in G$,*

$$\left(\hat{b}_j - \beta_j^0 \right) / \widehat{\text{s.e.}}_{\text{robust},j} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

$$\left(\hat{b}_j^* - \hat{\beta}_j \right) / \widehat{\text{s.e.}}_{\text{robust},j}^* \xrightarrow{\mathcal{D}^*} \mathcal{N}(0, 1) \text{ in probability.}$$

Here, "$\mathcal{D}$" and "$\mathcal{D}^*$" denote convergence in distribution with respect to the original and to the bootstrap measure, respectively.

A proof is given in the Electronic Supplementary Material. Different than for the homoscedastic case, the original and the bootstrap distributions have asymptotically different (estimated) standard errors

$$\widehat{\text{s.e.}}_{\text{robust},j} \sim \sqrt{\text{Asymp.Var}(\hat{b}_j)} \not\sim \sqrt{\text{Asymp.Var}^*(\hat{b}_j^*)} \sim \widehat{\text{s.e.}}_{\text{robust},j}^*,$$

where we omit that the statements are with high probability (in P^* and/or in P). Similarly, the residual bootstrap does not provide consistent estimation of the correlation

between different $\hat{b}_j$ to justify simultaneous inference as considered in Theorem 1. The reason is that the bootstrap constructs i.i.d. errors and does not mimic the heteroscedastic structure in the original sample. See also the sentences after the proof of Theorem 2 in the Electronic Supplementary Material. Simultaneous inference with heteroscedastic errors is treated in the following section.

4 Simultaneous inference with the bootstrap

We discuss here the advantages of the bootstrap for simultaneous inference and multiple testing adjustment in the presence of heteroscedasticity. Of particular interest here is the problem of simultaneous inference over a group $G \subseteq \{1, \ldots, p\}$ of components of the regression parameter β, including the case where $G = \{1, \ldots, p\}$ is very large and includes all components. More precisely, we want to estimate the distribution of

$$\max_{j \in G} h(T_j), \ T_j = \left(\hat{b}_j - \beta_j^0\right)/\widehat{\text{s.e.}}_{\text{robust},\, j}, \tag{11}$$

by using the bootstrap for $h(t) = t$, $h(t) = -t$ and $h(t) = |t|$.

We propose below bootstrap schemes which are consistent and work well for either homoscedastic or heteroscedastic errors.

4.1 The multiplier wild bootstrap

We introduce a multiplier wild bootstrap (Wu 1986; Liu and Singh 1992; Mammen 1993). Consider the centered residuals $\hat{\varepsilon}_{\text{cent}} = \hat{\varepsilon} - \bar{\hat{\varepsilon}}$, where $\hat{\varepsilon} = Y - \mathbf{X}\hat{\beta}$, and construct the multiplier bootstrapped residuals as

$\varepsilon_i^{*W} = W_i\, \hat{\varepsilon}_{\text{cent},i} \ (i = 1, \ldots, n),$
$W_1, \ldots, W_n$ i.i.d. independent of the data with $\mathbb{E}W_i = 0$, $\mathbb{E}W_i^2 = 1$ and $\mathbb{E}W_i^4 < \infty$.
$$\tag{12}$$

We then proceed as with the standard residual bootstrap for constructing $Y^* = \mathbf{X}\hat{\beta} + \varepsilon^{*W}$, and the bootstrap sample is then $\{(X_i, Y_i^*)\}_{i=1}^n$ as input to compute the bootstrapped estimator $T_j^* = (\hat{b}_j^* - \hat{b}_j)/\widehat{\text{s.e.}}_{\text{robust},\, j}^*$, i.e., using the plug-in rule of the bootstrap sample to the estimator.

This wild bootstrap scheme is asymptotically consistent for simultaneous inference with heteroscedastic (as well as homoscedastic) errors, see Sect. 4.4.

4.2 The *xyz*-paired bootstrap

We modify here the paired bootstrap for regression (Efron 1979; Liu and Singh 1992) to deal with the case of heteroscedastic errors (Freedman 1981). As recomputation of Z_j with bootstrap data would be expensive, we propose to append z-variables to the *xy*-matrix as additional regressors and bootstrap the entire rows of the *xyz*-matrix.

However, to create an unbiased regression model for the bootstrap, the variables have to be correctly centered to ensure $\mathbb{E}^*[(X_j^*)^T \varepsilon^*] = \mathbb{E}^*[(Z_j^*)^T \varepsilon^*] = 0$. We note that this is not a problem in the low-dimensional case because the residual vector in the least squares estimation is automatically orthogonal to all design vectors. The wild bootstrap does not have a centering problem either because the newly generated multiplier variables W_i all have zero mean. For the paired bootstrap, we propose to i.i.d. sample rows of the $n \times (2p+1)$ matrix $(\hat{\mathbf{X}}, \hat{Y}, \hat{\mathbf{Z}})$, and hence the name xyz-paired bootstrap,

$$\hat{X}_j = X_j - \frac{X_j^T \hat{\varepsilon}_{\text{cent}}}{\|\hat{\varepsilon}_{\text{cent}}\|_2^2} \hat{\varepsilon}_{\text{cent}}, \ \hat{Y} = \hat{\mathbf{X}}\hat{\beta} + \hat{\varepsilon}_{\text{cent}}, \ \hat{Z}_j = Z_j - \frac{Z_j^T \hat{\varepsilon}_{\text{cent}}}{\|\hat{\varepsilon}_{\text{cent}}\|_2^2} \hat{\varepsilon}_{\text{cent}},$$

where $\hat{\varepsilon}_{\text{cent}}$ is as in the residual bootstrap. Indeed, for the resulting $(\mathbf{X}^*, Y^*, \mathbf{Z}^*)$,

$$\mathbb{E}^*[\varepsilon^*] = \mathbb{E}^*[(X_j^*)^T \varepsilon^*] = \mathbb{E}^*[(Z_j^*)^T \varepsilon^*] = 0 \ \text{ with } \varepsilon^* = Y^* - \mathbf{X}^* \hat{\beta} = (\hat{\varepsilon}_{\text{cent}})^*.$$

The bootstrapped estimators $\hat{b}_j^*$, $\hat{\omega}_j^*$ and $\widehat{\text{s.e.}}_{\text{robust},j}^*$ are then defined by the plug-in rule as in the wild bootstrap, with $T_j^* = (\hat{b}_j^* - \hat{\beta}_j)/\widehat{\text{s.e.}}_{\text{robust},j}^*$.

The *xyz*-paired bootstrap is shown to be consistent for simultaneous inference with heteroscedastic errors, see Sect. 4.4. However, limited empirical results (not shown in the paper) suggested that it may not be competitive in comparison with the Gaussian multiplier wild bootstrap from Sect. 4.1.

4.3 The Westfall–Young procedure for multiple testing adjustment

The Westfall–Young procedure (Westfall and Young 1993) is a very attractive powerful approach for multiple testing adjustment based on resampling. It uses the bootstrap to approximate joint distributions of *p* values and test statistics, therefore taking their dependencies into account. This in turn leads to efficiency gains: The procedure has been proven for certain settings to be (nearly) optimal for controlling the familywise error rate (Meinshausen et al. 2011).

A standard assumption for the Westfall–Young procedure is the so-called subset pivotality for the statistics $T_j = \hat{b}_j/\widehat{\text{s.e.}}_{\text{robust},j}$ (or using the version for the homoscedastic case with $\widehat{\text{s.e.}}_j$). Note that in this subsection, T_j is without the centering at $\hat{\beta}_j$.

(subs-piv) Subset pivotality holds if, for every possible subset G, the marginal distribution for $\{T_j \ j \in G\}$ remains the same under the restriction $H_{0,G} : \beta_j = 0$ for all $j \in G$ and $H_{0,\text{complete}} : \beta_j = 0$ for all $j = 1, \ldots, p$.

When focusing specifically on a max-type statistic, we can weaken subset pivotality to a restricted form.

(restricted subs-piv) Restricted subset pivotality holds if, for every possible subset G, the distribution of $\max_{j \in G} |T_j|$ remains the same under the restriction $H_{0,G} : \beta_j = 0$ for all $j \in G$ and $H_{0,\text{complete}} : \beta_j = 0$ for all $j = 1, \ldots, p$.

Subset pivotality can be justified in an asymptotic sense. For groups G with finite cardinality, Theorems 1 and 3 presented below imply that asymptotic subset pivotality

holds. For large groups G (with $|G|$ as large as $p \gg n$) and assuming Gaussian errors, the restricted form of subset pivotality holds, see, e.g., Zhang and Zhang (2014). For large groups and non-Gaussian errors, restricted subset pivotality can be established under the conditions in Theorems 1 and 3. (The proof of these theorems implies the restricted subset pivotality, by using arguments from Chernozhukov et al. (2013).)

Assuming restricted subset pivotality (in an asymptotic sense) we immediately obtain that for any group $G \subseteq \{1, \ldots, p\}$:

$$\mathbb{P}\left[\max_{j \in G} |T_j| \leq c | H_{0,G}\right] = \mathbb{P}\left[\max_{j \in G} |T_j| \leq c | H_{0,\text{complete}}\right] \quad (c \in \mathbb{R}), \qquad (13)$$

or its asymptotic version with approximate equality. This suggests to approximate the distribution of $\max_{j \in G} |T_j|$ under the complete null hypothesis $H_{0,\text{complete}}$ by using a bootstrap scheme under the complete null hypothesis $H_{0,\text{complete}}$. We use

$$Y^{*0} = \varepsilon^{*W} \text{ for the multiplier wild bootstrap in (12),} \qquad (14)$$

that is, the construction as before but replacing $\hat{\beta}$ by the zero vector. For the heteroscedastic residual bootstrap, this means that we perform i.i.d. resampling of the rows of $(\hat{\varepsilon}_{\text{cent}}, X, Z_j)$. We notationally emphasize the bootstrap under $H_{0,\text{complete}}$ by the asterisk "*0". The bootstrap approximation is then as follows:

$$\mathbb{P}^{*0}\left[\max_{j \in G} |T_j^{*0}| \leq c\right] \approx \mathbb{P}\left[\max_{j \in G} |T_j| \leq c | H_{0,\text{complete}}\right] \quad (c \in \mathbb{R}),$$

and when invoking (13) we obtain that $\mathbb{P}^{*0}[\max_{j \in G} |T_j^{*0}| \leq c] \approx \mathbb{P}[\max_{j \in G} |T_j| \leq c | H_{0,G}]$. A rigorous justification for this approximation and the parallel approximation by the xyz-paired bootstrap is given in Theorem 3 below.

We then easily obtain multiplicity-adjusted p values which approximately control the familywise error rate for testing all the hypotheses $H_{0,j} : \beta_j^0 = 0$ for all $j = 1, \ldots, p$:

$$P_{j,\text{corr}} = \mathbb{P}^{*0}\left[\max_{k \in \{1,\ldots,p\}} |T_k^{*0}| > |t_j|\right],$$

where $T_k = \hat{b}_k / \widehat{s.e.}_{\text{robust},k}$ (or using the non-robust version $\widehat{s.e.}_k$), T_k^{*0} its bootstrapped version under $H_{0,\text{complete}}$ using (14) and t_j is the observed, realized value of the test statistic T_j. Because the bootstrap is constructed under the complete $H_{0,\text{complete}}$ we can compute the bootstrap distribution of $\max_{k \in \{1,\ldots,p\}} |T_k^{*0}|$ once and then use it to calibrate the p values for all components $j = 1, \ldots, p$: Obviously, this is computationally very efficient.

As described in Westfall and Young (1993), this method improves upon Bonferroni-style and Sidak adjustments, mainly because the bootstrap is taking dependence among the test statistics into account and hence is not overly conservative like the Bonferroni-type or Sidak correction. Furthermore, the Westfall–Young method does not rely on the assumption that the p values are uniformly distributed under $H_{0;j}$, for all j. Finally, a Bonferroni-type correction goes far into the tails of the distributions of the individual

test statistics, in particular if p is large: One typically would need some importance sampling for a computationally efficient bootstrap approximation of a single test statistic in the tails. We found that the Westfall–Young method is much less exposed to this issue (because the maximum statistics are directly bootstrapped without doing additional corrections in the tail).

4.4 Consistency of the multiplier wild and *xyz*-paired bootstrap

We discuss under which assumptions the multiplier wild and *xyz*-paired bootstrap schemes achieve consistency for estimating the distribution of $T_j = (\hat{b}_j - \beta_j^0)/\widehat{s.e.}_{\text{robust},j}$, $\max_{j \in G}(\pm T_j)$, and $\max_{j \in G} |T_j|$, where $G \subseteq \{1, \ldots, p\}$. The centered and standardized bootstrapped estimator is $T_j^* = (\hat{b}_j^* - \hat{\beta}_j)/\widehat{s.e.}^*_{\text{robust},j}$.

Theorem 3 *Assume (A1)–(A5) (and thus allowing for heteroscedastic errors). Let* $\mathbb{P}^*$ *represent the multiplier wild bootstrap. Then,*

$$T_j = \left(\hat{b}_j - \beta_j^0\right)/\widehat{s.e.}_{\text{robust},j} \implies \mathcal{N}(0, 1),$$

$$T_j^* = \left(\hat{b}_j^* - \hat{\beta}_j\right)/\widehat{s.e.}^*_{\text{robust},j} \stackrel{\mathcal{D}^*}{\implies} \mathcal{N}(0, 1) \text{ in probability,}$$

for each $j \in G$. *If* $|G| = O(1)$; *then,*

$$\sup_{(t_j, j \in G)} \left| \mathbb{P}^* \left[T_j^* \leq t_j, j \in G \right] - \mathbb{P} \left[T_j \leq t_j, j \in G \right] \right| = o_P(1).$$

If (A6) holds, then

$$\sup_{c \in \mathbb{R}} \left| \mathbb{P}^* \left[\max_{j \in G} h \left(T_j^* \right) \leq c \right] - \mathbb{P} \left[\max_{j \in G} h \left(T_j \right) \leq c \right] \right| = o_P(1)$$

for $h(t) = t$, $h(t) = -t$ *and* $h(t) = |t|$.

Moreover, all the above statements also hold when $\mathbb{P}^*$ *represents the xyz-paired bootstrap, provided that* $\delta = 2$ *in (A2) and (A3),* $\max_{j \in G}(\|Z_j\|_2/|Z_j^T X_j|) = o_P(1/\sqrt{\log(2|G|)})$, *and* $\sqrt{\log(p)/n} = o_P(1/\sqrt{\log(p) \log(1 + |G|)})$.

A proof is given in the Electronic Supplementary Material. We note that the assumption (A4) is meant to be with respect to the multiplier wild or the paired *xyz*-bootstrap, respectively: It is ensured by the same conditions as outlined in Sects. 3.3 and 3.3.2. Furthermore, condition (A6) can be relaxed to $\log(|G|) = o(n^{1/5})$ for Mammen's wild bootstrap and the xyz-paired bootstrap (Deng and Zhang 2017).

For the *xyz*-paired bootstrap, the additional condition $\log(p)/n = o_P(1/\sqrt{\log(p) \log(1 + |G|)})$ is a consequence of (A1) and the ℓ_1 minimax rate of the Lasso (Ye and Zhang 2010), and upper bounds of the form $\max_j \|Z_j\|_2/|Z_j^T X_j| = O_P(n^{-1/2})$, implying the requirement in Theorem 3 and the uniform $n^{-1/2}$ rate for the standard error of $\hat{b}_j$, can be found in Zhang and Zhang (2014) and van de Geer et al. (2014).

4.4.1 Conceptual differences between the multiplier wild and residual bootstrap

We briefly discuss some conceptual differences between the multiplier bootstrap and residual bootstrap while (mostly) not distinguishing whether the inference is simultaneous or for individual parameters. (The residual bootstrap also works for simultaneous inference as discussed in Theorem 1.)

The multiplier wild bootstrap leads to the correct standard error of the estimator for both cases of either homo- or heteroscedastic errors, i.e.,

$$\widehat{s.e.}^*_{\text{robust},j} \sim \sqrt{\text{Asym.Var}^*(\hat{b}^*_j)} \sim \sqrt{\text{Asym.Var}(\hat{b}_j)} \sim \widehat{s.e.}_{\text{robust},j}.$$

The asymptotic equivalence $\text{Asym.Var}^*(\hat{b}^*_j) \sim \text{Asym.Var}(\hat{b}_j)$ does not hold for the residual bootstrap in the case of heteroscedastic errors. However, this property is not needed when constructing the inference based on the pivots as in (9), and the absence of the asymptotic equivalence between studentized $\hat{b}^*_j$ and $\hat{b}_j$ is theoretically supported by Theorem 2. Nevertheless, the fact that the residual bootstrap does not capture the correct asymptotic variance in the non-standardized case, which has been a major reason to introduce the wild bootstrap (Mammen 1993), might remain a disadvantage for the residual bootstrap.

When the multiplier variables W_i are i.i.d. $\mathcal{N}(0, 1)$, the wild bootstrap as in (12) induces an *exact* Gaussian distribution (given the data) for the linear part $Z_j^T \varepsilon^* / Z_j^T X_j$, the leading term of $\hat{b}^*_j$. This is considered in Belloni et al. (2015b), Belloni et al. (2015a) and Zhang and Cheng (2016). For the finite sample case with non-Gaussian errors, the distribution of the original quantity $Z_j^T \varepsilon / Z_j^T X_j$ is non-Gaussian: By construction, the Gaussian multiplier bootstrap cannot capture such a non-Gaussianity. The residual bootstrap is better tailored to potentially pick up such non-Gaussianity and hence might have an advantage over the Gaussian multiplier wild bootstrap. Still, if heteroscedasticity is a concern, one should use non-Gaussian multipliers as advocated in Mammen (1993) and justified in Theorem 3.

Our limited empirical results suggest that the residual bootstrap and Gaussian multiplier wild bootstrap lead to very similar empirical results in terms of type I (actual level of significance for tests, and actual confidence coverage) and type II errors (power of tests, and size of confidence regions) for (i) the case of homoscedastic errors and for individual and simultaneous inference and (ii) the case of heteroscedastic errors and individual inference when using the robust standard error formula for the residual bootstrap. For the case of heteroscedastic errors and simultaneous inference, the wild bootstrap seems to be the preferred method. Some supporting empirical results are given in the Electronic Supplementary Material.

5 Empirical results

We compare the bootstrapped estimator to the original de-sparsified Lasso in terms of single testing confidence intervals and multiple testing corrected p values. We also consider the restricted low-dimensional projection estimator (RLDPE) which
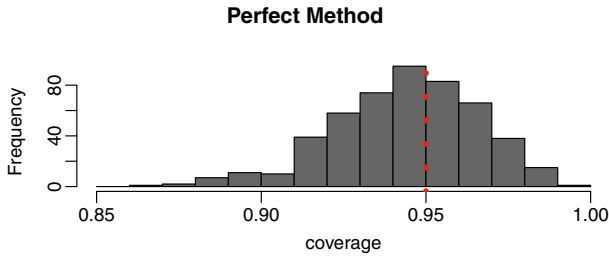
Perfect Method

Fig. 1 Histogram of the coverage probabilities of two-sided 95% confidence intervals for 500 parameters. It illustrates how the results look like for a perfectly correct method for creating confidence intervals and one uses only 100 realizations to compute the probabilities (color figure online)

has been introduced by Zhang and Zhang (2014) as a version of the de-biased (or de-sparsified) Lasso to enhance the reliability of coverage while paying a price for efficiency, and we also compare with the ZC approach from Zhang and Cheng (2016) which applies the bootstrap only to the linear part of the de-sparsified estimator without bootstrapping the estimated bias correction term. We always consider the residual bootstrap, unless explicitly specified that the wild bootstrap (with Gaussian multipliers) is used. Moreover, when considering scenarios with homoscedastic errors, we always studentize with the non-robust estimator $\widehat{s.e.}_j$ and for heteroscedastic errors, we always studentize with the robust estimator $\widehat{s.e.}_{\text{robust},j}$ (unless specified differently).

Of particular interest is the accuracy of the bootstrap when dealing with non-Gaussian and even heteroscedastic errors. For multiple testing, one would like to find out how much there is to gain when using the Westfall–Young procedure over a method that does not exploit dependencies between the outcomes of the tests, such as Bonferroni–Holm. To this end, it is interesting to look at a variety of dependency structures for the design matrix and to look at real data as well.

For confidence intervals, we visualize the overall average coverage probability as well as the occurrence of too high or too low coverage probabilities. We work with histograms of the coverage probabilities for all coefficients in the model, as in example in Fig. 1. These probabilities are always computed based on 100 realizations of the corresponding linear model. For those cases where coverage is too low, we visualize the confidence intervals themselves to illustrate the poor coverage. An example of the plot we will work with is shown in Fig. 2.

For multiple testing, we look at the power and the familywise error rate:

$$\text{Power} = \sum_{j \in S_0} \mathbb{P}[H_{0,j} \text{ is rejected}]/s_0,$$
$$\text{FWER} = \mathbb{P}[\exists j \in S_0^c : H_{0,j} \text{ is rejected}],$$

where the probabilities are computed based on 100 realizations of the linear model.

We use boxplots to visualize the power and error rates, similar to Fig. 3, where each data point is the result of the probability calculation described above. In order to generate interesting and representative data points, we look at different choices for
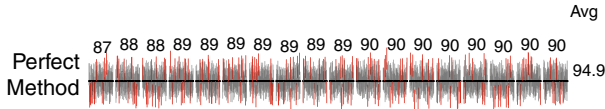
Fig. 2 Plot of two-sided 95% confidence intervals. It illustrates how the results would look like for a correct method for creating confidence intervals when one only computes 100 confidence intervals. Eighteen coefficients are chosen and are drawn in 18 columns from left to right with a black horizontal bar indicating the coefficient size. If any coefficients differ from zero then they are plotted first from the left (in order of increasing magnitude). This particular example does not exhibit any of those nonzero coefficients. The other coefficients are chosen to be those with the lowest coverage such that we can investigate potential causes for this poor coverage. The 100 computed confidence intervals are drawn from left to right in the column for the corresponding coefficient. The line segment is colored black in case it contains the truth, red otherwise. The number of confidence intervals that cover the truth for a particular coefficient is written above the confidence intervals in the respective column. The overall average coverage probability over all coefficients is displayed in the right-most column (color figure online)
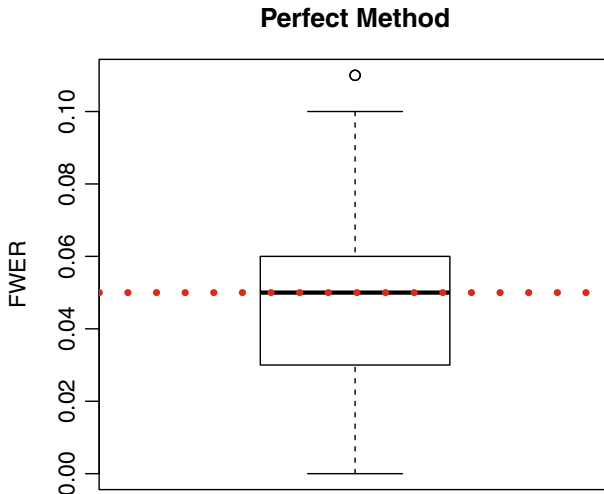
Fig. 3 Plot of multiple testing performance in terms of familywise error rate (FWER) control. It illustrates how the results look like for a correct method for multiple testing correction, if one computes the error rates over 100 realizations of the model. The target is controlling the FWER at level 0.05. This target is highlighted by a horizontal, dotted red line. We sample independent and identically distributed p values $p_j \sim U(0, 1)$, for $j = 1, \ldots, 500$, and compute the familywise error rate over 100 realizations when using the rejection threshold $\alpha = 0.05/500 = 0.0001$. The boxplot based on 300 data points is the result of repeating this experiment 300 times (color figure online)

the signal and different seeds for the data generation. As a rule, results for different design types are put in separate plots.

5.1 Varying the distribution of the errors

We first consider the performance of the bootstrap when varying the distribution of the errors for simulated data.

The design matrix will be generated $\sim \mathcal{N}_p(0, \Sigma)$ with a covariance matrix Σ of two possible types (although mainly of Toeplitz type):

$$\text{Toeplitz:} \quad \Sigma_{j,k} = 0.9^{|j-k|}.$$
$$\text{Independence:} \quad \Sigma = I_p.$$

In case the model contains signal, the coefficient vector will have $s_0 = 3$ coefficients that differ from zero. This is a rather small number of active variables. However, the theory requires (at least) $s_0 = o(\sqrt{n}/\log(p))$ (Assumption (B2)) leading to the value 1.61 within the $o(.)$-term for $n = 100$ and $p = 500$ which we typically consider in our simulations. Larger values of s_0 have been considered empirically in van de Geer et al. (2014): For such cases, the coverage of confidence intervals or the type I error control in testing was found to be rather unreliable. The coefficients are picked in 6 different ways:

$$\text{Randomly generated}: \ \text{U}(0, 2), \text{U}(0, 4), \text{U}(-2, 2),$$
$$\text{A fixed value}: \ 1, 2 \text{ or } 10.$$

5.1.1 Homoscedastic Gaussian errors

Data are generated from a linear model with Toeplitz design matrix and homoscedastic Gaussian errors of variance $\sigma^2 = 1$, $\varepsilon \sim \mathcal{N}_n(0, I_n)$. The sample size is chosen to be $n = 100$, and the number of parameters $p = 500$.

For confidence intervals, we focus on one generated design matrix $\mathbf{X}$ and one generated coefficient vector of type $U(-2, 2)$. The histograms for the coverage probabilities are shown in Fig. 4. The coverage probabilities are more correct for the bootstrapped estimator. The original estimator has a bias for quite a few coefficients resulting in low coverage, as shown in Fig. 5. In addition, it tends to have too high coverage for many coefficients. The conservative RLDPE has much wider confidence intervals, which addresses the problem of low coverage, but results in too high overall coverage.

For multiple testing, we generate 50 Toeplitz design matrices $\mathbf{X}$ which are combined with 50 coefficient vectors for each coefficient type $U(0, 2)$, $U(0, 4)$, $U(-2, 2)$, fixed 1, fixed 2 and fixed 10. For each of these 300 linear models, the coefficient vector undergoes a different random permutation. A value for the familywise error rate and power is then computed by generating 100 realizations of the linear models, as described in the introduction of Sect. 5. The boxplots of the power and familywise error rate are shown in Fig. 6. The bootstrap is the least conservative option. In addition, one can conclude that it still has proper error control by comparing the results to perfect error control in Fig. 3. One would expect to see a difference in power, but there does not seem to be a visible difference between the bootstrap approach and the original estimator for our dataset. The RLDPE, on the other hand, does turn out to be more conservative.
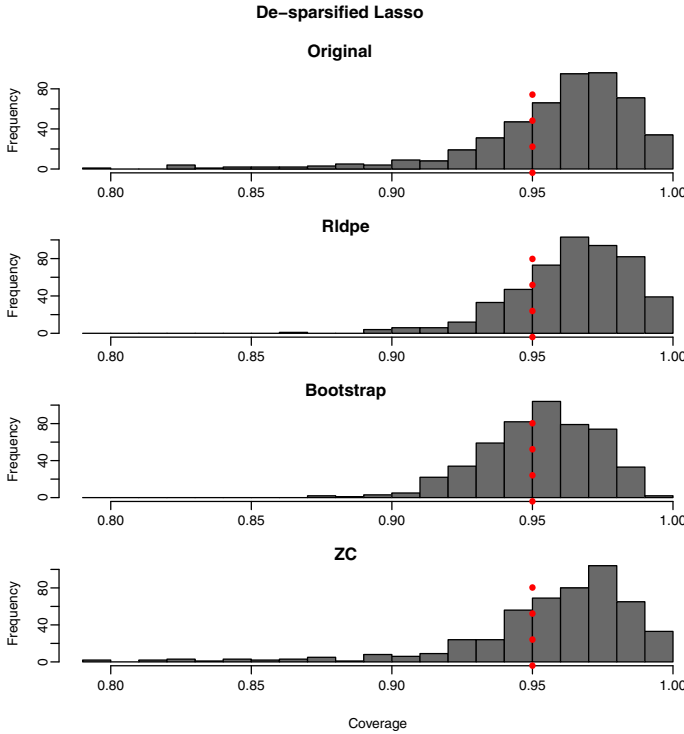
Fig. 4 Histograms of the coverage probabilities of two-sided 95% confidence intervals for all 500 parameters in a linear model ($n = 100$, $p = 500$), computed from 100 independent replications. Perfect performance would look like Fig. 1. The fixed design matrix is of Toeplitz type, the single coefficient vector of type $U(-2, 2)$ and *homoscedastic Gaussian errors*. The original estimator has more over-coverage and under-coverage than the bootstrapped estimator. The RLDPE has little under-coverage, like the bootstrapped estimator, but it has too high coverage probabilities overall. The ZC approach to bootstrapping, which only bootstraps the linearized part of the estimator, does not show any improvements over the original de-sparsified Lasso (color figure online)

5.1.2 Homoscedastic non-Gaussian errors

Data are generated from a linear model with Toeplitz design matrix and homoscedastic centered Chi-squared errors $\varepsilon_1, \ldots, \varepsilon_n$, of variance $\sigma^2 = 1$,

$$\zeta_1, \ldots, \zeta_n \text{ i.i.d.} \sim \chi_1^2, \quad \varepsilon_i = \frac{\zeta_i - 1}{\sqrt{2}}, \quad i = 1, \ldots, n.$$

The sample size is chosen to be $n = 100$, and the number of parameters $p = 500$.

For confidence intervals, we focus on one generated design matrix **X** and one generated coefficient vector of type $U(-2, 2)$. The histograms for the coverage probabilities are shown in Fig. 7.

The performance for the confidence intervals looks similar to that for Gaussian errors; only the under-coverage of the original estimator is even more pronounced.
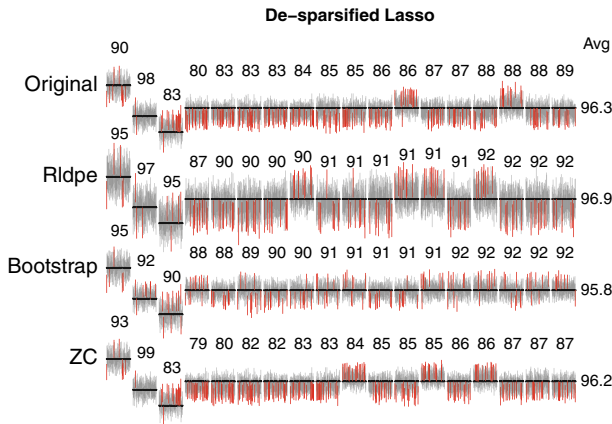
Fig. 5 Two-sided 95% confidence intervals for the de-sparsified Lasso estimator. From left to right 18 coefficients are shown with a black horizontal bar of a certain height illustrating the value of the coefficient. Only the first three coefficients differ from zero. The other 15 coefficients presented are those with the lowest confidence interval coverage for that particular method (in increasing order from left to right). One hundred response vectors were generated for a linear model with *homoscedastic Gaussian errors*, fixed design of type Toeplitz, a single coefficient vector of type $U(-2, 2)$, sample size $n = 100$ and dimension $p = 500$. Each of these realizations was fitted to produce a confidence interval for each coefficient in the model. The 100 confidence intervals are drawn as vertical lines and ordered from left to right in the column corresponding to that particular coefficient. The line segments are colored black if they cover the true coefficient and colored red otherwise. The number above each coefficient corresponds to the number of confidence intervals, out of 100, which end up covering the truth. The average coverage probability over all coefficients is provided in a column to the right of all coefficients. The original estimator has some bias for a few coefficients, which results in a lower-than-desired coverage for those coefficients. The RLDPE has wider confidence intervals exhibiting over-coverage. The ZC approach to bootstrapping, which only bootstraps the linearized part of the estimator, does not show any improvements over the original de-sparsified Lasso (color figure online)

The coverage for the bootstrapped estimator looks as good as in the Gaussian case. As shown in Fig. 8, the cause for the poor coverage of the non-bootstrapped estimator is again bias. Using the robust standard error estimation does not impact the results, as shown in the Electronic Supplementary Material.

For multiple testing, the same setups were looked at as in Sect. 5.1.1, but now with the different errors. As shown in Fig. 9, the poor single testing confidence interval coverage does not translate into poor multiple testing error control. The original method with Bonferroni–Holm is on the conservative side, while the bootstrap is slightly closer to the correct level.

5.1.3 Heteroscedastic non-Gaussian errors

Data are generated from a linear model with heteroscedastic non-Gaussian errors. The example is taken from Mammen (1993) with sample size $n = 50$, but where we increased the number of parameters to $p = 250$ from the original $p = 5$. The model has no signal $\beta_1^0 = \beta_2^0 = \cdots = \beta_p^0 = 0$ and introduces heteroscedasticity while still maintaining the correctness of the linear model.
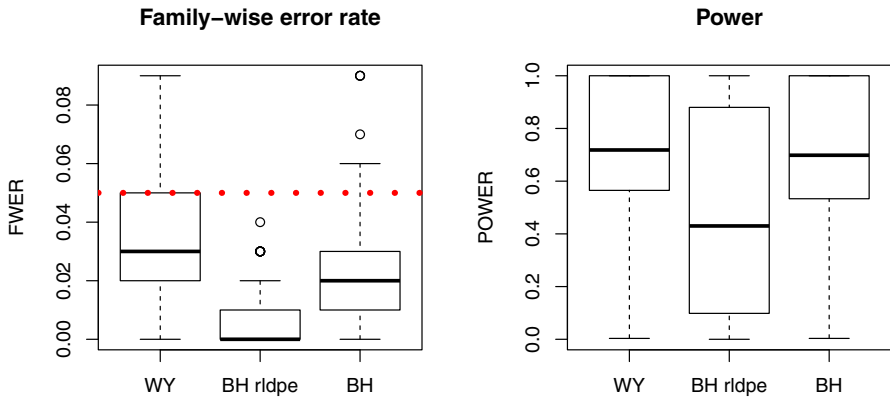
Fig. 6 Boxplot of the familywise error rate and the power for multiple testing for the de-sparsified Lasso. The target is controlling the FWER at level 0.05, highlighted by a horizontal, dotted red line. Two different approaches for multiple testing correction are compared: Westfall–Young (WY) and Bonferroni–Holm (BH). For Bonferroni–Holm, we make the distinction between the original method and the RLDPE approach. Three hundred linear models are investigated in total, where 50 Toeplitz design matrices are combined with 50 coefficient vectors for each of the 6 types: $U(0, 2)$, $U(0, 4)$, $U(-2, 2)$, fixed 1, fixed 2, fixed 10. The variables belonging to the active set are chosen randomly. The errors in the linear model were chosen to be *homoscedastic Gaussian*. Each of the models has a data point for the error rate and the power in the boxplot. The error rate and power probabilities were calculated by averaging over 100 realizations (color figure online)
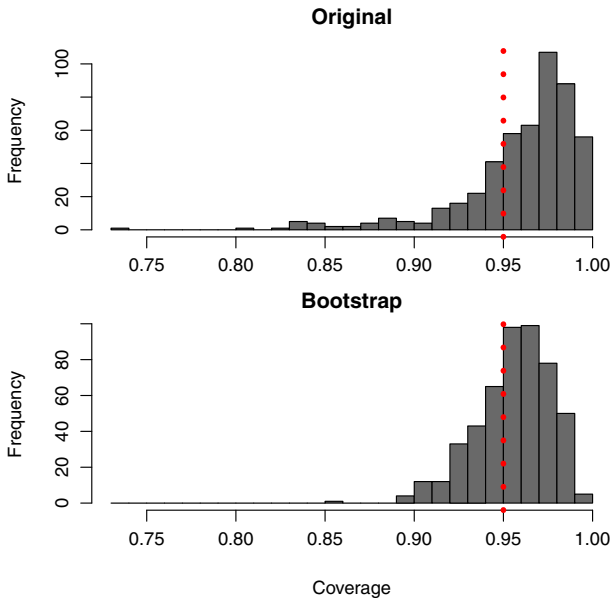

Fig. 7 The same plot as Fig. 4 but for *homoscedastic Chi-squared errors*. The bootstrapped estimator has better coverage properties (color figure online)
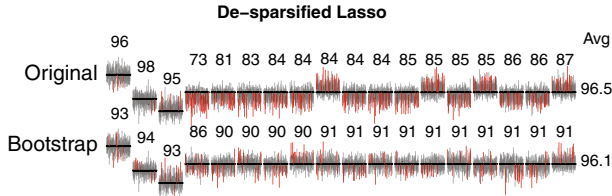
Fig. 8 The same plot as Fig. 5 but for *homoscedastic Chi-squared errors*. The original estimator has quite some bias for a few coefficients, which results in a lower-than-desired coverage for those coefficients (color figure online)
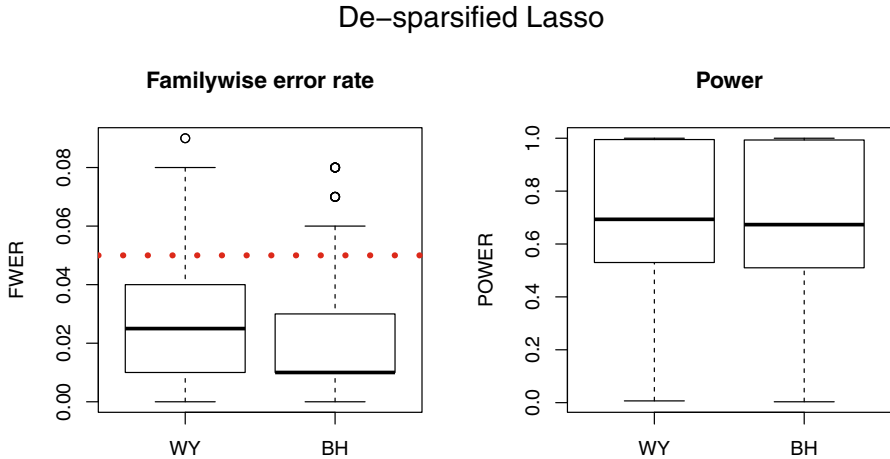

Fig. 9 The same plot as Fig. 6, but for *homoscedastic Chi-squared errors* (color figure online)

Each row of the design matrix $\mathbf{X}$ is generated independently $\sim \mathcal{N}_p(0, I_p)$ and then given a different variance. Each row is multiplied with the value $Z_i/2$, where the $\{Z_1, \ldots, Z_n\}$ are chosen i.i.d. $\sim U(1, 3)$.

The errors ε_i are chosen to be a mixture of normal distributions

$$\varepsilon_1, \ldots, \varepsilon_n \text{ i.i.d, } \varepsilon_i = l_i \zeta_i + (l_i - 1)\eta_i,$$
$$l_i \sim \text{Bernoulli}(0.5), \ \zeta_i \sim \mathcal{N}(1/2, (1.2)^2), \ \eta_i \sim \mathcal{N}(-1/2, (0.7)^2),$$

with l_i, ζ_i, η_i independent of each other.

The responses are generated by introducing heteroscedasticity in the errors

$$Y_i = Q_i \varepsilon_i + \varepsilon_i \quad \forall i = 1, \ldots, n$$

with $Q_i = X_{i,1}^2 + X_{i,2}^2 + X_{i,3}^2 + X_{i,4}^2 + X_{i,5}^2 - \mathbb{E}[Z_i^2]$.

For confidence intervals, we focus on one generated design matrix $\mathbf{X}$. The histograms for the coverage probabilities are shown in Fig. 10, and the plot of the actual confidence intervals is shown in Fig. 11. What is immediately clear from Fig. 10 is that it makes a big difference if one uses the robust version of the standard error estimation
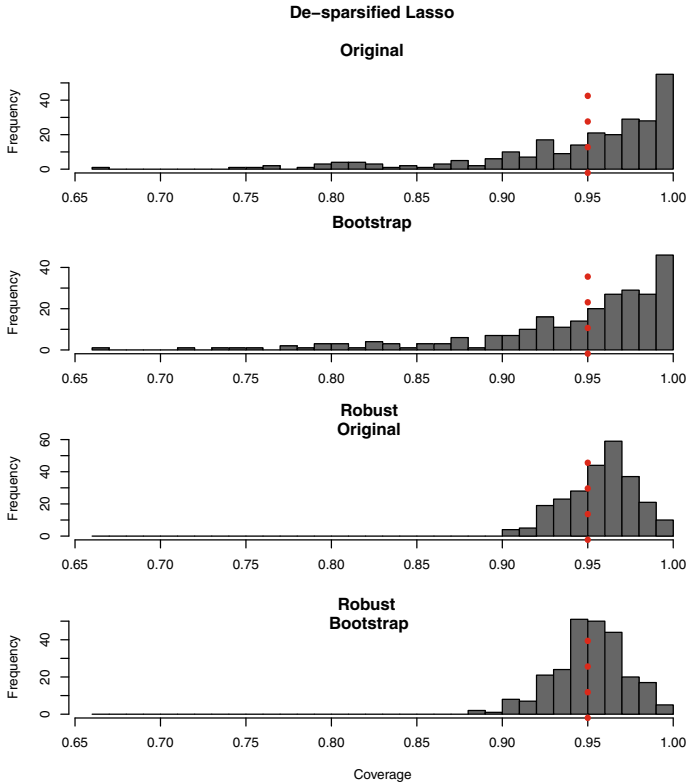
Fig. 10 The same plot as Fig. 4, but for *heteroscedastic non-Gaussian errors* and without signal. The robust standard error estimation clearly outperforms the non-robust version. There seems to be hardly any difference between the bootstrap and the original estimator after choosing the standard error estimation (color figure online)

or not. The coverage is very poor for the non-robust methods, while for the robust methods the performance looks like perfect coverage (Fig. 1).

There isn't any benefit for the bootstrap over the original estimator for this dataset. The robust original estimator does not show any bias in Fig. 11 and has great coverage already. The overall average coverage is slightly more correct for the bootstrap with a value of 95.1 versus 95.9.

In contrast to the single testing confidence intervals, all methods (robust and non-robust) perform adequately for multiple testing as shown in Fig. 9. Due to the lack of signal in the dataset, we can only investigate error rates. Fifty different design matrices were generated to produce the 50 data points in the boxplots. The bootstrap is less conservative and has actual error rate closer to the true level (Fig. 12).

5.1.4 Discussion

Bootstrapping the de-sparsified Lasso turns out to improve the coverage of confidence intervals without increasing the confidence interval lengths (that is, without losing
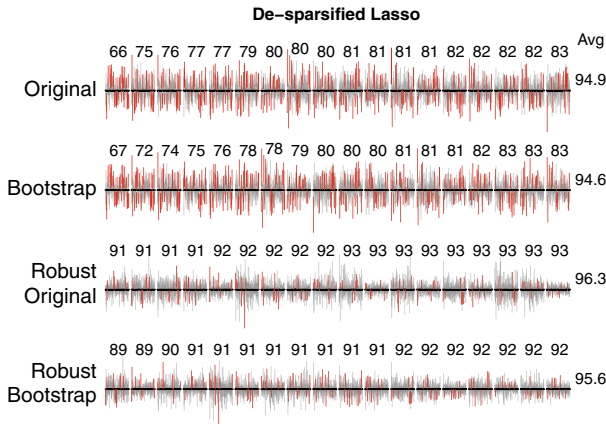
Fig. 11 The same plot as Fig. 5, but for *heteroscedastic non-Gaussian errors* and without signal. The non-robust estimators have low coverage for many coefficients. Unlike the other setups, there does not seem to be a bias in the original estimator for this dataset (color figure online)
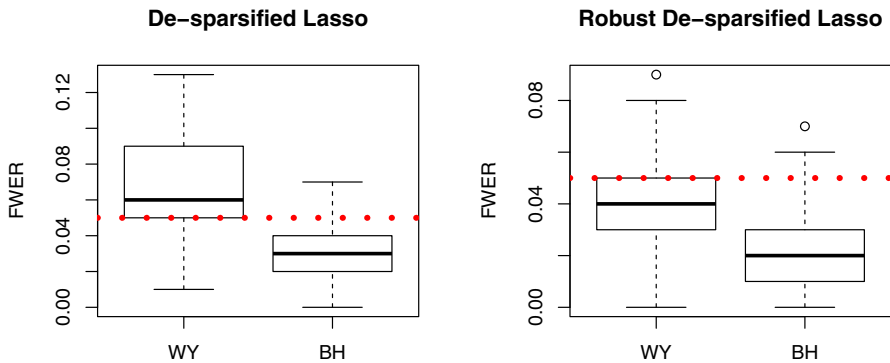

Fig. 12 The same plot as Fig. 6, but for *heteroscedastic non-Gaussian errors* and without signal. We only report the error rate because all null hypotheses are true for the generated dataset. The plot on the left is for the non-robust methods, and the one on the right for the robust ones (color figure online)

efficiency). The use of the conservative RLDPE (Zhang and Zhang 2014) is not necessary: The bootstrap achieves reliable coverage, while for the original de-sparsified Lasso, the RLDPE seems worthwhile to achieve reasonable coverage while paying a price in terms of efficiency. Furthermore, bootstrapping only the linearized part of the de-sparsified estimator as proposed by Zhang and Cheng (2016), and implicit in the work by Belloni et al. (2015a, b), is clearly sub-ideal in comparison with bootstrapping the entire estimator and using the plug-in principle as advocated here.

For multiple testing, the bootstrapped estimator had familywise error rates that were closer to the target level, while Bonferroni–Holm adjustment is too conservative. This finding was not reflected in any noticeable power improvements, but some gains are found, see Sect. 5.2.

The robust standard error turned out to be critical when dealing with heteroscedastic errors. Therefore, we recommend the bootstrapped estimator with robust standard error estimation as the method to be used: If the errors were homoscedastic, we pay a price of efficiency; see also the sentence at the end of Sect. 5.2.1.

As can be seen in the Electronic Supplementary Material, the Gaussian multiplier bootstrap also performs well. The performance is very similar to the residual bootstrap, and as one would expect, it handles heteroscedastic errors as good as the robust standard error bootstrap approach.

5.2 A closer look at multiple testing

The examples from Sect. 5.1 showed little to no power difference between the bootstrap and the original estimator. One straightforward explanation for this is that the signal in the simulated datasets did not fall into the (possibly small) differences in rejection regions.

As another more signal-independent way to investigate multiple testing performance, we compare the computed rejection regions. Unfortunately, the actual values of the rejection thresholds are often quite unintuitive to compare. Instead, it can be more informative to invert the Bonferroni–Holm adjustment rule to compute some *equivalent number of tests*, which is essentially equivalent to the number of tests under independence. The Westfall–Young procedure computes a rejection threshold t_{rej} for the absolute value of the test statistic, and we can then compute the equivalent number of tests (with the Bonferroni adjustment) p_{equiv} as

$$p_{\text{equiv}} = \frac{\alpha}{2(1 - \Phi(t_{rej}))}, \tag{15}$$

for controlling the familywise error rate at level α and with $\Phi(.)$ the cumulative distribution function for $\mathcal{N}(0, 1)$. Improvements in rejection threshold are then reflected in p_{equiv} being a lot smaller than the actual number of hypotheses tested p, while still properly controlling the error rates.

Looking at the rejection thresholds presented in Fig. 13, we can see that the bootstrap does improve substantially over the original estimator with a Bonferroni correction. Multiple testing with the bootstrap is often equivalent to testing about 300 (independent) tests with Bonferroni correction in comparison with the original 500.

5.2.1 Real measurements design

We take design matrices from real data (8 datasets measuring gene expressions, see table below) and simulate a linear model with known signal and homoscedastic Gaussian errors. We look at all 6 signal options described in Sect. 5.1.

For every signal type, we only look at 5 different seeds for generating the coefficients and for the permutations of the coefficient vector (in contrast to the typical 50 as used in Sect. 5.1.1). As usual, the familywise error rates are computed based on 100 realizations of each model.
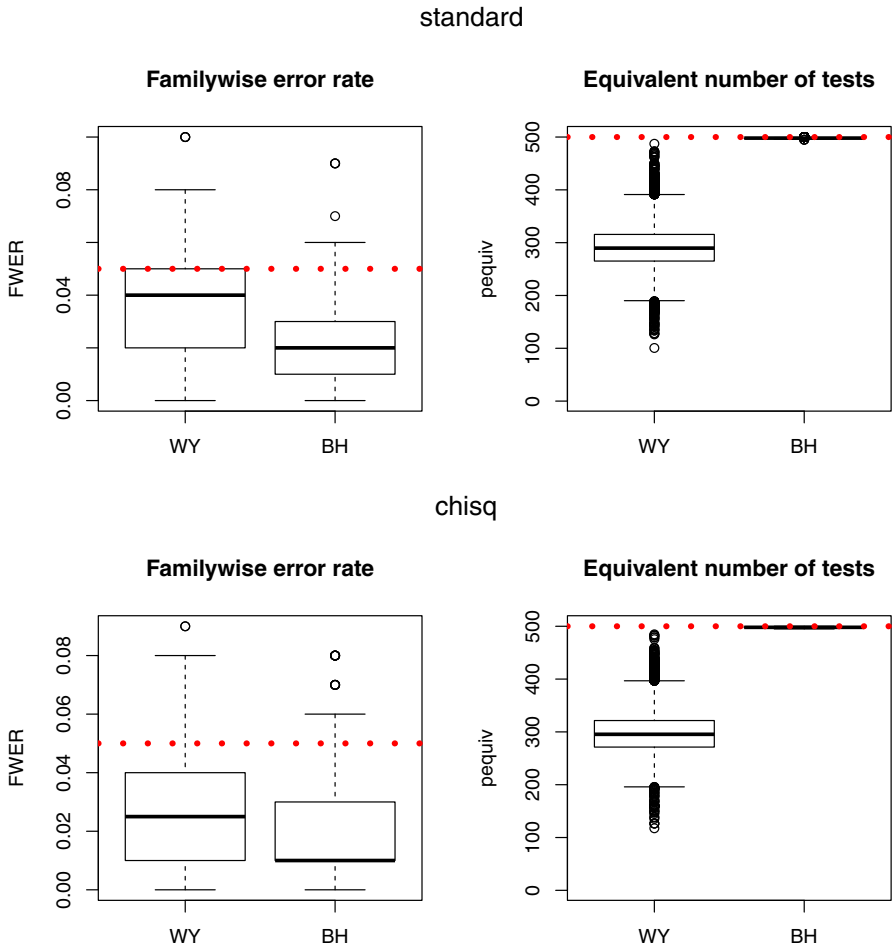
Fig. 13 The same plots as Fig. 6 for the *homoscedastic Gaussian errors* (top) and Fig. 9 for the *homoscedastic non-Gaussian errors* (bottom), but displaying the number of equivalent tests p_{equiv} instead of the power. The actual number of hypotheses tested is highlighted by a horizontal, dotted red line (color figure online)

Boxplots of the familywise error rate and p_{equiv} for the `lymphoma` dataset are shown in Fig. 14. The median values of the equivalent number of tests and the FWER for all the different designs are as follows:

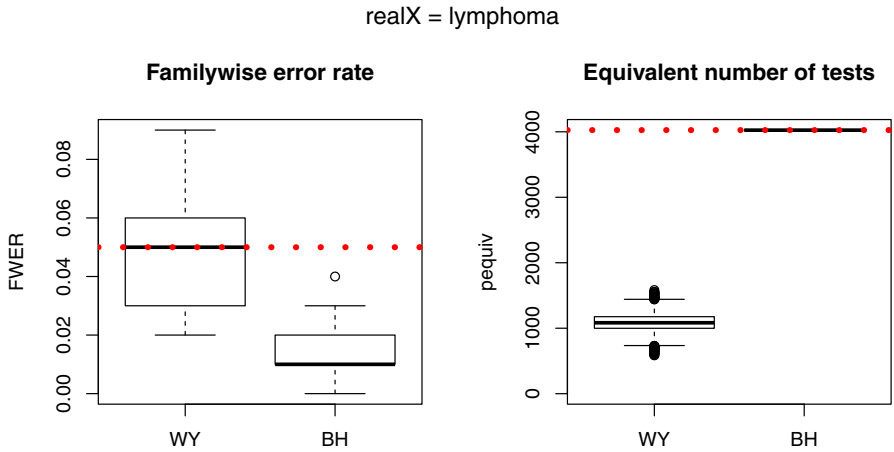|	dsmN71	Brain	Breast	Lymphoma	Leukemia	Colon	Prostate	nci
Median p_{equiv} WY	1264	886	1162	1083	1230	655	2466	1289
Median p_{equiv} BH	4088	5596	7129	4025	3570	2000	6032	5243
Dimension p	4088	5597	7129	4026	3571	2000	6033	5244
Median FWER WY	0.02	0.06	0.05	0.05	0.05	0.03	0.06	0.03
Median FWER BH	0.00	0.02	0.01	0.01	0.03	0.00	0.04	0.01

realX = lymphoma

Fig. 14 The same plot as in Fig. 13, but with design matrix coming from real measurements (lymphoma in this case) with simulated signal and *homoscedastic Gaussian errors* (color figure online)

The bootstrap achieves substantial reductions in the median equivalent number of tests for all datasets investigated here.

We note that when studentizing the test statistics with the robust standard error, the power gain with the bootstrap (Westfall–Young method) is often rather marginal. This is illustrated in the Electronic Supplementary Material.

5.2.2 Real data example

We revisit a dataset about riboflavin production by bacillus subtilis (Bühlmann et al. 2014), already studied in Bühlmann (2013), van de Geer et al. (2014) and Dezeure et al. (2015). The dataset has dimensions $n = 71$, $p = 4088$, and the original de-sparsified Lasso does not manage to reject any null hypothesis $H_{0,j}$ at the 5% significance level after multiple testing correction with Bonferroni–Holm.

Despite the power gain that is possible with this design matrix (see dsmN71 in Table in Sect. 5.2.1), the bootstrapped estimator does not reject any hypotheses either with the Westfall–Young procedure.

We investigate what signal strength one would be able to detect in this real dataset by adding artificial signal to the original responses. This is done by adding a linear component $X_j c$ of increasing signal strength c for a single variable j,

$$Y^{'} = Y + X_j c.$$

One can keep track of the *p* value for this particular coefficient and repeat the experiment for all possible columns of the design $j = 1, \ldots, p$. Boxplots of this experiment are shown in Fig. 15.

The bootstrap results in smaller *p* values for the same signal values. It rejects the relevant null hypothesis almost all the time for signal values above $c > 2.5$. Note that
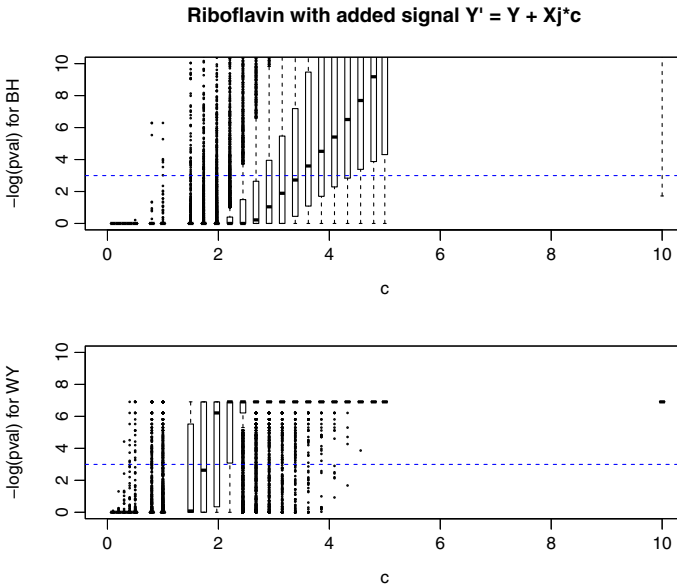
Riboflavin with added signal Y' = Y + Xj*c

Fig. 15 Multiple testing corrected p values for increasing artificial signal added to the real dataset about riboflavin (dsmN71). Signal is added only one variable at a time $Y' = Y + X_j c$, and only the p value for that coefficient p_j is stored for each such experiment and value of c. The values $-\log(p_j)$ for all different experiments ($j = 1, \ldots, p = 4088$) are plotted in boxplots grouped by value of c. A horizontal, dashed blue line indicates the rejection threshold 0.05. The bootstrap approach to multiple testing clearly has higher power as it picks up on the signal quicker. The bootstrap has a lower bound on the minimal achievable p value due to the finite number of bootstrap samples $B = 1000$, namely $-\log(1/1000) \approx 6.9$

we do not have access to replicates and therefore, we cannot determine the actual error rate.

5.2.3 Discussion

The bootstrap with the Westfall–Young (WY) multiple testing adjustment leads to reliable familywise error control while providing a rejection threshold which is far more powerful than using the Bonferroni adjustment (in the case of homoscedastic errors), especially in the presence of dependence among the components for testing (while for heteroscedastic errors and when using the robust standard error for studentization, the efficiency gain of WY often seems less substantial). Since the efficient WY adjustment is not adding additional computational costs to the one from bootstrapping, such simultaneous inference and WY multiple testing adjustment is highly recommended.

6 Further considerations

We discuss here additional points before providing some conclusions.

6.1 Model misspecification

So far, the entire discussion has been for a linear model as in (1) with a sparse regression vector β^0. The linearity is not really a restriction: Suppose that the true model would involve a nonlinear regression function

$$Y_i = f^0(X_i) + \varepsilon_i \ (i = 1, \ldots n).$$

The $n \times 1$ vector of the true function at the observed data values $(f^0(X_1), \ldots, f^0(X_n))^T$ can be represented as

$$(f^0(X_1), \ldots, f^0(X_n))^T = \mathbf{X}\beta^0 \tag{16}$$

for many possible solutions β^0: This is always true as long as rank$(\mathbf{X}) = n$, which typically holds in $p \geq n$ settings. The issue is whether there are solutions β^0 in (16) which are sparse: A general result to address this point is not available, see Bühlmann and van de Geer (2015).

It is argued in Bühlmann and van de Geer (2015) that weak sparsity in terms of $\|\beta^0\|_r$ for $0 < r < 1$ suffices to guarantee that the de-sparsified Lasso has an asymptotic Gaussian distribution centered at β^0, as described in Theorem 1 or 2. Thus, assuming that there is a weakly sparse solution β^0 in (16) is relaxing the requirement for ℓ_0 sparsity. The presented theory for the bootstrap could be adapted to cover the case for weakly sparse β^0.

The interpretation of a confidence interval for β^0, based on the Gaussian limiting distribution of the de-sparsified Lasso or using its bootstrapped version as described in this paper, is that it covers all ℓ_r $(0 < r < 1)$ weakly sparse solutions β^0 which are solutions of (16). For this, we implicitly assume that there is at least one such ℓ_r weakly sparse solution.

6.2 Random design

The distinction between fixed and random design becomes crucial for misspecified models. If the true model with random design is linear, then by conditioning on the covariables, the corresponding fixed design model is again linear. And if the inference is correct conditional on $\mathbf{X}$, it is also correct unconditional for random design. If the true random design model is nonlinear, one can look at the best projected random design linear model: But then, when conditioning, the obtained projected fixed design linear model has a bias (or nonzero conditional mean for the error). In other words, conditioning on the covariables is not appropriate when the model is wrong, and one should rather do unconditional inference in a random design (best approximating) linear model, see Bühlmann and van de Geer (2015).

Thus, there are certainly situations where one would like to do unconditional inference in a random design linear model, see also Freedman (1981) who proposes the "paired bootstrap" in a low-dimensional setting. The bootstrap which we discussed in this paper is for fixed design only: for random design, one should resample the

covariables as well. Because of the latter the computational task becomes much more demanding: For the de-sparsified Lasso, and when p is large, most of the computation is spent on computing all the residual vectors $Z_1, \ldots, Z_p$, which requires running the Lasso p times. For bootstrapping with fixed design, this computation has to be done only once (since $Z_1, \ldots, Z_p$ are deterministic values of the fixed design $\mathbf{X}$); with random design, it seems unavoidable to do it $B \approx 100 - 1'000$ times which would result in a major additional computational cost.

6.3 Conclusions

We propose residual, wild and paired bootstrap methodologies for individual and simultaneous inference in high-dimensional linear models with possibly non-Gaussian and heteroscedastic errors. The bootstrap is used to approximate the distribution of the de-sparsified Lasso, a regular non-sparse estimator which is not exposed to the unpleasant super-efficiency phenomenon.

We establish asymptotic consistency for possibly simultaneous inference for parameters in a group $G \subseteq \{1, \ldots, p\}$ of variables, where $p \gg n$ but $s_0 = o(n^{1/2}/\{\log(p) \log(|G|)\})$ and $\log(|G|) = o(n^{1/7})$ with s_0 denoting the sparsity. The presented general theory is complemented by many empirical results, demonstrating the advantages of our approach over other proposals. Especially for simultaneous inference and multiple testing adjustment, the bootstrap is very powerful.

For homoscedastic errors, the residual bootstrap and wild bootstrap perform similarly. For heteroscedastic errors, the wild bootstrap is more natural and can be used for simultaneous inference (whereas the residual bootstrap fails to be consistent). Thus, for protecting against heteroscedastic errors, the wild bootstrap seems to be the preferred method. Our proposed procedures are implemented in the R-package `hdi` (Meier et al. 2016).

References

Belloni A, Chernozhukov V, Chetverikov D, Wei Y (2015a) Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation. Preprint arXiv:1512.07619

Belloni A, Chernozhukov V, Kato K (2015b) Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. Biometrika 102(1):77–94

Bickel P, Klaassen C, Ritov Y, Wellner J (1998) Efficient and adaptive estimation for semiparametric models. Springer, Berlin

Breiman L (1996) Heuristics of instability and stabilization in model selection. Ann Stat 24:2350–2383

Bühlmann P (2013) Statistical significance in high-dimensional linear models. Bernoulli 19:1212–1242

Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, Berlin

Bühlmann P, van de Geer S (2015) High-dimensional inference in misspecified linear models. Electron J Stat 9:1449–1473

Bühlmann P, Kalisch M, Meier L (2014) High-dimensional statistics with a view towards applications in biology. Annu Rev Stat Appl 1:255–278

Chatterjee A, Lahiri S (2011) Bootstrapping Lasso estimators. J Am Stat Assoc 106:608–625

Chatterjee A, Lahiri S (2013) Rates of convergence of the adaptive LASSO estimators to the oracle distri-
    bution and higher order refinements by the bootstrap. Ann Stat 41:1232–1259

Chernozhukov V, Chetverikov D, Kato K (2013) Gaussian approximations and multiplier bootstrap for
    maxima of sums of high-dimensional random vectors. Ann Stat 41:2786–2819

Chernozhukov V, Chetverikov D, Kato K (2014) Central limit theorems and bootstrap in high dimensions.
    The Annals of Probabiliy, To appear, Preprint arXiv:1412.3661

Chernozhukov V, Hansen C, Spindler M (2016) hdm: high-dimensional metrics. Preprint arXiv:1608.00354

Deng H, Zhang C-H (2017) Beyond Gaussian approximation: bootstrap in large scale simultaneous infer-
    ence. unpublished work in progress

Dezeure R, Bühlmann P, Meier L, Meinshausen N (2015) High-dimensional inference: confidence intervals,
    p-values and R-software hdi. Stat Sci 30:533–558

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26

Eicker F (1967) Limit theorems for regressions with unequal and dependent errors. In: Proceedings of the
    fifth Berkeley symposium on mathematical statistics and probability, vol 1, pp 59–82

Foygel Barber R, Candès EJ (2015) Controlling the false discovery rate via knockoffs. Ann Stat 43:2055–
    2085

Freedman DA (1981) Bootstrapping regression models. Ann Stat 9:1218–1228

Giné E, Zinn J (1989) Necessary conditions for the bootstrap of the mean. Ann Stat 17:684–691

Giné E, Zinn J (1990) Bootstrapping general empirical measures. Ann Probab 18:851–869

Hall P, Wilson SR (1991) Two guidelines for bootstrap hypothesis testing. Biometrics 47:757–762

Huber PJ (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceed-
    ings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1, pp 221–233

Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regres-
    sion. J Mach Learn Res 15:2869–2909

Liu RY, Singh K (1992) Efficiency and robustness in resampling. Ann Stat 20:370–384

Liu H, Yu B (2013) Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear
    regression. Electron J Stat 7:3124–3169

Mammen E (1993) Bootstrap and wild bootstrap for high dimensional linear models. Ann Stat 21:255–285

McKeague IW, Qian M (2015) An adaptive resampling test for detecting the presence of significant predic-
    tors. J Am Stat Assoc 110:1422–1433

Meier L, Dezeure R, Meinshausen N, Mächler M, Bühlmann P (2016) hdi: high-dimensional inference. R
    package version 0.1-6

Meinshausen N (2015) Group bound: confidence intervals for groups of variables in sparse high dimensional
    regression without assumptions on the design. J R Stat Soc B 77:923–945

Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the Lasso. Ann
    Stat 34:1436–1462

Meinshausen N, Bühlmann P (2010) Stability selection (with discussion). J R Stat Soc B 72:417–473

Meinshausen N, Meier L, Bühlmann P (2009) P-values for high-dimensional regression. J Am Stat Assoc
    104:1671–1681

Meinshausen N, Maathuis MH, Bühlmann P (2011) Asymptotic optimality of the Westfall-Young permu-
    tation procedure for multiple testing under dependence. Ann Stat 39:3369–3391

Reid S, Tibshirani R, Friedman J (2016) A study of error variance estimation in Lasso regression. Stat
    Sinica 26:35–67

Rudelson M, Zhou S (2013) Reconstruction from anisotropic random measurements. IEEE Trans Inf Theory
    59:3434–3447

Shah R, Samworth R (2013) Variable selection with error control: another look at stability selection. J R
    Stat Soc B 75:55–80

Shah R, Bühlmann P (2015) Goodness of fit tests for high-dimensional linear models. J R Stat Soc B. doi:10.
    1111/rssb.12234

van de Geer S, Bühlmann P, Zhou S (2011) The adaptive and the thresholded Lasso for potentially mis-
    specified models (and a lower bound for the Lasso). Electron J Stat 5:688–749

van de Geer S, Bühlmann P, Ritov Y, Dezeure R (2014) On asymptotically optimal confidence regions and
    tests for high-dimensional models. Ann Stat 42:1166–1202

Wasserman L, Roeder K (2009) High dimensional variable selection. Ann Stat 37:2178–2201

Westfall P, Young S (1993) Resampling-based multiple testing: examples and methods for P-value adjust-
    ment. Wiley, Hoboken

White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48:817–838

Wu C-FJ (1986) Jackknife, bootstrap and other resampling methods in regression analysis. Ann Stat 14:1261–1295

Ye F, Zhang C-H (2010) Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. J Mach Learn Res 11:3481–3502

Zhang C-H, Huang J (2008) The sparsity and bias of the Lasso selection in high-dimensional linear regression. Ann Stat 36:1567–1594

Zhang C-H, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc B 76:217–242

Zhang X, Cheng G (2016) Simultaneous inference for high-dimensional linear models. J Am Stat Assoc. doi:10.1080/01621459.2016.1166114

Zhou Q (2014) Monte Carlo simulation for Lasso-type problems by estimator augmentation. J Am Stat Assoc 109:1495–1516