The Annals of Statistics
2019, Vol. 47, No. 4, 2080–2116
https://doi.org/10.1214/18-AOS1741

PERTURBATION BOOTSTRAP IN ADAPTIVE LASSO

BY DEBRAJ DAS[1], KARL GREGORY[2] AND S. N. LAHIRI[1]

University of Wisconsin—Madison, University of South Carolina and North Carolina State University

The Adaptive Lasso (Alasso) was proposed by Zou [*J. Amer. Statist. Assoc.* **101** (2006) 1418–1429] as a modification of the Lasso for the purpose of simultaneous variable selection and estimation of the parameters in a linear regression model. Zou [*J. Amer. Statist. Assoc.* **101** (2006) 1418–1429] established that the Alasso estimator is variable-selection consistent as well as asymptotically Normal in the indices corresponding to the nonzero regression coefficients in certain fixed-dimensional settings. In an influential paper, Minnier, Tian and Cai [*J. Amer. Statist. Assoc.* **106** (2011) 1371–1382] proposed a perturbation bootstrap method and established its distributional consistency for the Alasso estimator in the fixed-dimensional setting. In this paper, however, we show that this (naive) perturbation bootstrap fails to achieve second-order correctness in approximating the distribution of the Alasso estimator. We propose a modification to the perturbation bootstrap objective function and show that a suitably Studentized version of our modified perturbation bootstrap Alasso estimator achieves second-order correctness even when the dimension of the model is allowed to grow to infinity with the sample size. As a consequence, inferences based on the modified perturbation bootstrap will be more accurate than the inferences based on the oracle Normal approximation. We give simulation studies demonstrating good finite-sample properties of our modified perturbation bootstrap method as well as an illustration of our method on a real data set.

1. Introduction.

Consider the multiple linear regression model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \dots, n, \tag{1.1}$$

where $y_1, \dots, y_n$ are responses, $\epsilon_1, \dots, \epsilon_n$ are independent and identically distributed (i.i.d.) random variables, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are known nonrandom design vectors and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ is the p-dimensional vector of regression parameters. When the dimension p is large, it is common to approach regression model (1.1) with the

Received November 2016; revised February 2018.

[1]Supported in part by NSF Grants DMS-1300068 and DMS 1613192.

[2]Supported by Deutsche Forschungsgemeinschaft through the Research Training Group RTG 1953, funded by the German Research Foundation (DFG) and by the state of Baden-Württemberg through bwHPC.

MSC2010 subject classifications. Primary 62J07; secondary 62G09, 62E20.

Key words and phrases. Alasso, naive perturbation bootstrap, modified perturbation bootstrap, second-order correctness, oracle.

assumption that the vector $\boldsymbol{\beta}$ is sparse, that is, that the set $\mathcal{A} = \{j : \beta_j \neq 0\}$ has cardinality $p_0 = |\mathcal{A}|$ much smaller than p, meaning that only a few of the covariates are "active." The Lasso estimator introduced by Tibshirani (1996) is well suited to the sparse setting because of its property that it sets some regression coefficients exactly equal to 0. One disadvantage of the Lasso, however, is that it produces nontrivial asymptotic bias for the nonzero regression parameters, primarily because it shrinks all estimators toward zero [cf. Knight and Fu (2000)].

Building on the Lasso, Zou (2006) proposed the Adaptive Lasso (hereafter referred to as Alasso) estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ in the regression problem (1.1) as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{t}} \left[\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i' \boldsymbol{t})^2 + \lambda_n \sum_{j=1}^{p} |\tilde{\beta}_{j,n}|^{-\gamma} |t_j| \right], \tag{1.2}$$

where $\tilde{\beta}_{j,n}$ is the jth component of a root-n-consistent estimator $\tilde{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$, such as the ordinary least squares (OLS) estimator when $p \leq n$ or the Lasso or Ridge regression estimator when $p > n$, $\lambda_n > 0$ is the penalty parameter, and $\gamma > 0$ is a constant governing the influence of the preliminary estimator $\tilde{\boldsymbol{\beta}}_n$ on the Alasso fit. Zou (2006) showed in the fixed-p setting that under some regularity conditions and with the right choice of λ_n, the Alasso estimator enjoys the so-called oracle property [cf. Fan and Li (2001)]; that is, it is variable-selection consistent and it estimates the nonzero regression parameters with the same precision as the OLS estimator which one would compute if the set of active covariates were known.

In an important recent work, Minnier, Tian and Cai (2011) introduced the perturbation bootstrap in the Alasso setup. To state their main results, let $\boldsymbol{\beta}_n^{*N} = (\beta_{1,n}^{*N}, \ldots, \beta_{p,n}^{*N})'$ be the naive perturbation bootstrap Alasso estimator prescribed by Minnier, Tian and Cai (2011) and define $\hat{\mathcal{A}}_n = \{j : \hat{\beta}_{j,n} \neq 0\}$ and $\mathcal{A}_n^{*N} = \{j : \beta_{j,n}^{*N} \neq 0\}$. These authors showed that under some regularity conditions and with p fixed as $n \to \infty$,

$$\mathbf{P}_*(\mathcal{A}_n^{*N} = \hat{\mathcal{A}}_n) \to 1 \quad \text{and} \quad \sqrt{n}(\boldsymbol{\beta}_n^{*N(1)} - \hat{\boldsymbol{\beta}}_n^{(1)})|\boldsymbol{\varepsilon} \asymp_d \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}^{(1)}),$$

where $\boldsymbol{\varepsilon}_n = (\epsilon_1, \ldots, \epsilon_n)$, $\boldsymbol{z}^{(1)}$ denotes the subvector of $\boldsymbol{z} \in \mathcal{R}^p$ corresponding to the coordinates in $\mathcal{A} = \{j : \beta_j \neq 0\}$, "$\asymp_d$" denotes asymptotic equivalence in distribution and $\mathbf{P}_*$ denotes bootstrap probability conditional on the data. Thus Minnier, Tian and Cai (2011) [hereafter referred to as MTC(11)] showed that, in the fixed-p setting and conditionally on the data, the naive perturbation bootstrap version of the Alasso estimator is variable-selection consistent in the sense that it recovers the support of the Alasso estimator with probability tending to one and that its distribution conditional on the data converges at the same time to that of the Alasso estimator for the nonzero regression parameters. But the accuracy of inference for nonzero regression parameters relies on the rate of convergence of the bootstrap distribution of $\sqrt{n}(\boldsymbol{\beta}_n^{*N(1)} - \hat{\boldsymbol{\beta}}_n^{(1)})|\boldsymbol{\varepsilon}$ to the distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}^{(1)})$ after proper Studentization. Furthermore, Chatterjee and Lahiri (2013) showed that the convergence of the Alasso estimators of the nonzero regression coefficients to their oracle Normal distribution is quite slow, owing to the bias induced by the penalty term in (1.2). Thus, it would be important for the accuracy of inference if second-order correctness can be achieved in approximating the distribution of the Alasso estimator by the perturbation bootstrap. Second-order correctness implies that the distributional approximation has a uniform error rate of $o_p(n^{-1/2})$. We show in this paper, however, that the distribution of the naive perturbation bootstrap version of the Alasso estimator, as defined by MTC(11), cannot be second-order correct even in fixed dimension. For more details, see Section 4.

We introduce a modified perturbation bootstrap for the Alasso estimator for which second-order correctness does hold, even when the number of regression parameters $p = p_n$ is allowed to increase with the sample size n. We also show in Proposition 2.1 that the modified perturbation bootstrap version of the Alasso estimator (defined in Section 2) can be computed by minimizing simple criterion functions. This makes our bootstrap procedure computationally simple and inexpensive.

In this paper, we consider some pivotal quantities based on Alasso estimators and establish that the modified perturbation bootstrap estimates the distribution of these pivotal quantities up to second order, that is, with an error that is of much smaller magnitude than what we would obtain by using the Normal approximation under the knowledge of the true active set of covariates. We will refer to the Normal approximation which uses knowledge of the true set of active covariates as the oracle Normal approximation. Our main results show that the modified perturbation bootstrap method enables, for example, the construction of confidence intervals for the nonzero regression coefficients with smaller coverage error than those based on the oracle Normal approximation.

More precisely, we consider pivots which are Studentizations of the quantities

$$\sqrt{n}\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \quad \text{and} \quad \sqrt{n}\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) + \breve{\boldsymbol{b}}_n,$$

where $\boldsymbol{D}_n$ is a $q \times p$ matrix (q fixed) producing q linear combinations of interest of $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}$ and where $\breve{\boldsymbol{b}}_n$ is a bias correction term which we will define in Section 5. We find that in the $p \leq n$ case, the modified perturbation bootstrap can estimate the distribution of the first pivot with an error of order $o_p(n^{-1/2})$ (see Theorem 5.1). This is much smaller than the error of the oracle Normal approximation, which was shown in Theorem 3.1 of Chatterjee and Lahiri (2013) to be of the order $O_p(n^{-1/2} + \|\boldsymbol{b}_n\| + c_n)$, where $\boldsymbol{b}_n$ is the bias targeted by $\breve{\boldsymbol{b}}_n$ and $c_n > 0$ is determined by the initial estimator $\tilde{\boldsymbol{\beta}}_n$ and the tuning parameters λ_n and γ; both $\|\boldsymbol{b}_n\|$ and c_n are typically greater in magnitude than $n^{-1/2}$, and hence determine the rate of the oracle Normal approximation. We also discover that the bias correction in the second pivot improves the error rate so that the modified perturbation bootstrap estimator achieves the rate $O_p(n^{-1})$ (see Theorem 5.2), which is a significant improvement over the best possible rate of oracle Normal approximation, namely

$O(n^{-1/2})$. In the $p > n$ case, we find that the modified perturbation bootstrap estimates the distributions of Studentized versions of both the bias-corrected and unbias-corrected pivots with the rate $o_p(n^{-1/2})$ (see Theorems 5.3, 5.4 and 5.5), establishing the second-order correctness of our modified perturbation bootstrap in the high-dimensional setting. We have explored the cases when the dimension p is increasing polynomially with n and when p is increasing exponentially with n. Our adding to the pivot a bias correction term may bring to mind the desparsified Lasso introduced independently by Zhang and Zhang (2014) and van de Geer et al. (2014); these authors construct a nonsparse estimator of $\boldsymbol{\beta}$ by adding a Lasso-based bias correction to a biased, nonsparse estimator of $\boldsymbol{\beta}$ which is linear in the response values $y_1, \ldots, y_n$. They consider first-order properties of this nonsparse estimator, establishing asymptotic normality under sparsity conditions. In contrast, we consider the sparse Alasso estimator of $\boldsymbol{\beta}$ and correct the bias of a pivot based on the form of the Alasso estimator. We establish second-order results of our proposed perturbation bootstrap method in both before and after bias correction. The main motivation behind the bias correction is to achieve the error rate $O_p(n^{-1})$.

We show that the naive perturbation bootstrap of MTC(11) is not second-order correct (see Theorem 4.1) by investigating the Karush–Kuhn–Tucker (KKT) condition [cf. Boyd and Vandenberghe (2004)] corresponding to their minimization problem. It is shown that second-order correctness is not attainable by the naive version of the perturbation bootstrap, primarily due to lack of proper centering of the naive bootstrapped Alasso criterion function. We derive the form of the centering constant by analyzing the corresponding approximation errors using the theory of Edgeworth expansion. To accommodate the centering correction, we modify the perturbation bootstrap criterion function for the Alasso; see Section 2 for details. In addition, we also find out that it is beneficial, from both theoretical and computational perspectives, to modify the perturbation bootstrap version of the initial estimators in a similar way. To prove second-order correctness of the modified perturbation bootstrap Alasso, the key steps are to find an Edgeworth expansion of the bootstrap pivotal quantities based on the modified criterion function and to compare it with the Edgeworth expansion of the sample pivots. We want to mention that the dimension p of the regression parameter vector can grow polynomially in the sample size n at a rate depending on the number of finite polynomial moments of the error distribution. Extension to the case in which p grows exponentially with n is possible under the assumption of finiteness of moment generating function of the regression errors. In this regime, we have explored separately two important special cases, namely when the errors are sub-Gaussian and subexponential.

We conclude this section with a brief literature review. The perturbation bootstrap was introduced by Jin, Ying and Wei (2001) as a resampling procedure where the objective function has a U-process structure. Work on the perturbation bootstrap in the linear regression setup is limited. Some work has been carried out by Chatterjee and Bose (2005), MTC(11), Zhou, Song and Thompson (2012) and Das and Lahiri (2019). As a variable selection procedure, Tibshirani (1996) introduced the Lasso. Zou (2006) proposed the Alasso as an improvement over the Lasso. For the Alasso and related popular penalized estimation and variable selection procedures, the residual bootstrap has been investigated by Knight and Fu (2000), Hall et al. (2009), Chatterjee and Lahiri (2010, 2011, 2013), Wang and Song (2011), MTC(11), van de Geer et al. (2014) and Camponovo (2015), among others.

The rest of the paper is organized as follows. The modified perturbation bootstrap for the Alasso is introduced and discussed in Section 2. Assumptions and explanations of those are presented in Section 3. Negative results on the naive perturbation bootstrap approximation proposed by MTC(11) are discussed in Section 4. Main results concerning the estimation properties of the Studentized modified perturbation bootstrap pivotal quantities as well as intuitions and explanations behind the modification of the modified perturbation bootstrap are given in Section 5. Section 6 presents simulation results exploring the finite-sample performance of the modified perturbation bootstrap in comparison with other methods for constructing confidence intervals based on Alasso estimators. Additional simulation results are relegated to the Supplementary Material [Das, Gregory and Lahiri (2019)]. Section 7 gives an illustration on real data. An outline of the proofs are presented in Section 8. Details are provided in the Supplementary Material [Das, Gregory and Lahiri (2019)]. Section 9 states concluding remarks.

2. The modified perturbation bootstrap for Alasso. Let $G_1^*, \dots, G_n^*$ be n independent copies of a nondegenerate random variable $G^* \in [0, \infty)$ having expectation μ_{G^*}. These quantities will serve as perturbation quantities in the construction of the perturbation bootstrap Alasso estimator. We define our bootstrap version of the Alasso estimator as the minimizer of a carefully constructed penalized objective function which involves the Alasso predicted values $\hat{y}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_n$, $i = 1, \dots, n$ as well as the observed values $y_i, \dots, y_n$. These sets of values appear in the objective function in two perturbed least-squares criteria. Similar modification is also needed in defining the bootstrap versions of the Alasso initial estimators; see (2.2). The motivation behind this construction is detailed in Section 4. We point out in Section 5 why the naive perturbation bootstrap formulation of MTC(11) fails to achieve second-order correctness.

We formally define the modified perturbation bootstrap version $\hat{\boldsymbol{\beta}}_n^*$ of the Alasso estimator $\hat{\boldsymbol{\beta}}_n$ as

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\boldsymbol{*}} = \underset{\boldsymbol{t}^*}{\arg\min} \Bigg[& \sum_{i=1}^n (y_i - \boldsymbol{x}_i' \boldsymbol{t}^*)^2 (G_i^* - \mu_{G^*}) \\
& + \sum_{i=1}^n (\hat{y}_i - \boldsymbol{x}_i' \boldsymbol{t}^*)^2 (2\mu_{G^*} - G_i^*) + \mu_{G^*} \lambda_n \sum_{j=1}^p |\tilde{\beta}_{j,n}^*|^{-\gamma} |t_j^*| \Bigg],
\end{aligned}
\tag{2.1}
$$

where $\tilde{\beta}^*_{j,n}$ is the jth component of $\tilde{\boldsymbol{\beta}}^*_n$, the modified perturbation bootstrap version of the Alasso initial estimator $\tilde{\boldsymbol{\beta}}_n$. We construct $\tilde{\boldsymbol{\beta}}^*_n$ as

$$\begin{aligned}\tilde{\boldsymbol{\beta}}^*_n = \arg\min_{\boldsymbol{t}^*}\Bigg[&\sum_{i=1}^n (y_i - \boldsymbol{x}_i'\boldsymbol{t}^*)^2 (G_i^* - \mu_{G^*}) \\ &+ \sum_{i=1}^n (\hat{y}_i - \boldsymbol{x}_i'\boldsymbol{t}^*)^2 (2\mu_{G^*} - G_i^*) + \mu_{G^*}\tilde{\lambda}_n \sum_{j=1}^p |t_j^*|^l\Bigg],\end{aligned} \tag{2.2}$$

where $\tilde{\lambda}_n = 0$ when $\tilde{\boldsymbol{\beta}}_n$ is taken as the OLS, which we use when $p \le n$, and $l = 1$ or 2 according as the initial estimator $\tilde{\boldsymbol{\beta}}_n$ is taken as the Lasso or Ridge regression estimator when $p > n$. Note that $\tilde{\lambda}_n$ may be different from λ_n.

We point out that the modified perturbation bootstrap estimators can be computed using existing algorithms. Define $\boldsymbol{L}_1(\boldsymbol{t}) = \sum_{i=1}^n (y_i - \boldsymbol{x}_i'\boldsymbol{t})^2 (G_i^* - \mu_{G^*}) + \sum_{i=1}^n (\hat{y}_i - \boldsymbol{x}_i'\boldsymbol{t})^2 (2\mu_{G^*} - G_i^*) + \mu_{G^*}\tilde{\lambda}_n \sum_{j=1}^p c_j |t_j|^l$ for some nonnegative constants c_j, $j = 1, \dots, p$. Now set $z_i = \hat{y}_i + \hat{\epsilon}_i \mu_{G^*}^{-1}(G_i^* - \mu_{G^*})$, where $\hat{\epsilon}_i = y_i - \hat{y}_i$ for $i = 1, \dots, n$ and let $\boldsymbol{L}_2(\boldsymbol{t}) = \sum_{i=1}^n (z_i - \boldsymbol{x}_i'\boldsymbol{t})^2 + \tilde{\lambda}_n \sum_{j=1}^p c_j |t_j|^l$. Then we have the following proposition.

PROPOSITION 2.1. $\arg\min_{\boldsymbol{t}} \boldsymbol{L}_1(\boldsymbol{t}) = \arg\min_{\boldsymbol{t}} \boldsymbol{L}_2(\boldsymbol{t})$.

This proposition allows us to compute $\tilde{\boldsymbol{\beta}}^*_n$ as well as $\hat{\boldsymbol{\beta}}^*_n$ by minimizing standard objective functions on some pseudo-values. Note that the modified perturbation bootstrap versions of the Alasso estimator as well as of the Alasso initial estimator can be obtained simply by properly perturbing the Alasso residuals in the decomposition $y_i = \hat{y}_i + \hat{\epsilon}_i$, $i = 1, \dots, n$.

3. Assumptions. We first introduce some notation required for stating our assumptions and useful for the proofs later. We denote the true parameter vector as $\boldsymbol{\beta}_n = (\beta_{1,n}, \dots, \beta_{p,n})'$, where the subscript n emphasizes that the dimension $p := p_n$ may grow with the sample size n. Set $\mathcal{A}_n = \{j : \beta_{j,n} \ne 0\}$ and $p_0 := p_{0,n} = |\mathcal{A}_n|$. For simplicity, we shall suppress the subscript n in the notation p_n and p_{0n}. Without loss of generality, we shall assume that $\mathcal{A}_n = \{1, \dots, p_0\}$. Let $\boldsymbol{C}_n = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'$ and partition it according to $\mathcal{A}_n = \{1, \dots, p_0\}$ as

$$\boldsymbol{C}_n = \begin{bmatrix} \boldsymbol{C}_{11,n} & \boldsymbol{C}_{12,n} \\ \boldsymbol{C}_{21,n} & \boldsymbol{C}_{22,n} \end{bmatrix},$$

where $\boldsymbol{C}_{11,n}$ is of dimension $p_0 \times p_0$. Define $\tilde{\boldsymbol{x}}_i = \boldsymbol{C}_n^{-1}\boldsymbol{x}_i$ (when $p \le n$) and $\operatorname{sgn}(x) = -1, 0, 1$ according as $x < 0$, $x = 0$, $x > 0$, respectively. Suppose $\boldsymbol{D}_n$ is a known $q \times p$ matrix with $\operatorname{tr}(\boldsymbol{D}_n\boldsymbol{D}_n') = O(1)$ and q is not dependent on n. Let

$\boldsymbol{D}_n^{(1)}$ contain the first p_0 columns of $\boldsymbol{D}_n$. Define

$$\boldsymbol{S}_n = \begin{bmatrix} \boldsymbol{D}_n^{(1)} \boldsymbol{C}_{11,n}^{-1} \boldsymbol{D}_n^{(1)'} .\sigma^2 & \boldsymbol{D}_n^{(1)} \boldsymbol{C}_{11,n}^{-1} \bar{\boldsymbol{x}}_n^{(1)} .\mu_3 \\ \bar{\boldsymbol{x}}_n^{(1)'} \boldsymbol{C}_{11,n}^{-1} \boldsymbol{D}_n^{(1)'} .\mu_3 & (\mu_4 - \sigma^4) \end{bmatrix},$$

where $\bar{\boldsymbol{x}}_n = n^{-1} \sum_{i=1}^n \boldsymbol{x}_i = (\bar{\boldsymbol{x}}_n^{(1)\prime}, \bar{\boldsymbol{x}}_n^{(2)\prime})'$, $\sigma^2 = \mathbf{Var}(\epsilon_1) = \mathbf{E}(\epsilon_1^2)$, and where μ_3 and μ_4 are, respectively, the third and fourth central moments of ϵ_1. Define in addition the $q \times p_0$ matrix $\check{\boldsymbol{D}}_n^{(1)} = \boldsymbol{D}_n^{(1)} \boldsymbol{C}_{11,n}^{-1/2}$ and the $p_0 \times 1$ vector $\check{\boldsymbol{x}}_i^{(1)} = \boldsymbol{C}_{11,n}^{-1/2} \boldsymbol{x}_i^{(1)}$. Let K be a positive constant and r be a positive integer ≥ 3 unless otherwise specified. $\|\cdot\|$ and $\|\cdot\|_\infty$, respectively, denote the Euclidean norm and the Sup norm. $c \wedge d$ denotes $\min\{c, d\}$ for two real numbers c and d. $\mathbf{P}$ and $\mathbf{E}$, respectively, denote the usual probability and expectation, where as by $\mathbf{P}_*$ and $\mathbf{E}_*$ we denote, respectively, probability and expectation with respect to the distribution of G^* conditional upon the observed data.

We now introduce our assumptions.

(A.1) Let $\eta_{11,n}$ denote the smallest eigenvalue of the matrix $\boldsymbol{C}_{11,n}$:

(i) $\eta_{11,n} > Kn^{-a}$ for some $a \in [0, 1)$.

(ii) $\max\{n^{-1} \sum_{i=1}^n |x_{i,j}|^{2r} : 1 \leq j \leq p\} + \{n^{-1} \sum_{i=1}^n |(\boldsymbol{C}_{11,n}^{-1})_{j\cdot} \boldsymbol{x}_i^{(1)}|^{2r} : 1 \leq j \leq p_0\} = O(1)$.

(iii) $\max\{n^{-1} \sum_{i=1}^n |\tilde{x}_{i,j}|^{2r} : 1 \leq j \leq p\} = O(1)$, where $\tilde{x}_{i,j}$ is the jth element of $\tilde{\boldsymbol{x}}_i$. (when $p \leq n$)

(iii)$'$ $\max\{c_{11,n}^{j,j} : 1 \leq j \leq p_0\} = O(1)$, where $c_{11,n}^{j,j}$ is the (j, j)th element of $C_{11,n}^{-1}$ (when $p > n$).

(A.2) There exists a $\delta \in (0, 1)$ such that for all $n > \delta^{-1}$:

(i) $\sup\{\boldsymbol{x}' \check{\boldsymbol{D}}_n^{(1)} \check{\boldsymbol{D}}_n^{(1)'} \boldsymbol{x} : \boldsymbol{x} \in \mathcal{R}^q, \|\boldsymbol{x}\| = 1\} < \delta^{-1}$.

(ii) $n^{-1} \sum_{i=1}^n \|\check{\boldsymbol{D}}_n^{(1)} \check{\boldsymbol{x}}_i^{(1)} \check{\boldsymbol{x}}_i^{(1)'} \check{\boldsymbol{D}}_n^{(1)'}\|^r = O(1)$.

(iii) $\inf\{\boldsymbol{x}' \boldsymbol{S}_n \boldsymbol{x} : \boldsymbol{x} \in \mathcal{R}^{q+1}, \|\boldsymbol{x}\| = 1\} > \delta$.

(A.3) $\max\{|\beta_{j,n}| : j \in \mathcal{A}_n\} = O(1)$ and $\min\{|\beta_{j,n}| : j \in \mathcal{A}_n\} \geq Kn^{-b}$ for some $b \geq 0$ such that $4b < 1$ and $a + 2b \leq 1$, where a is defined as in (A.1)(i).

(A.4) (i) $\mathbf{E}|\epsilon_1|^r < \infty$. $\mathbf{E}\epsilon_1 = 0$.

(ii) $(\epsilon_1, \epsilon_1^2)$ satisfies Cramér's condition:

$$\limsup_{\|(t_1,t_2)\| \to \infty} \bigl|\mathbf{E}\bigl(\exp\bigl(i\bigl(t_1 \epsilon_1 + t_2 \epsilon_1^2\bigr)\bigr)\bigr)\bigr| < 1.$$

(A.5) (i) $\mathbf{E}_*(G_1^*)^r < \infty$. $\mathbf{Var}(G_1^*) = \sigma_{G^*}^2 = \mu_{G^*}^2$, $\mathbf{E}_*(G_1^* - \mu_{G^*})^3 = \mu_{G^*}^3$.

(ii) G_i^* and ϵ_i are independent for all $1 \leq i \leq n$.

(iii) $((G_1^* - \mu_{G^*}), (G_1^* - \mu_{G^*})^2)$ satisfies Cramér's condition:

$$\limsup_{\|(t_1,t_2)\| \to \infty} \bigl|\mathbf{E}_*\bigl(\exp\bigl(i\bigl(t_1\bigl(G_1^* - \mu_{G^*}\bigr) + t_2\bigl(G_1^* - \mu_{G^*}\bigr)^2\bigr)\bigr)\bigr)\bigr| < 1.$$

(A.6) There exists $\delta_1 \in (0,1)$ such that for all $n > \delta_1^{-1}$:

(i) $\frac{\lambda_n}{\sqrt{n}} \leq \delta_1^{-1} n^{-\delta_1} \min\{\frac{n^{-b\gamma}}{p_0}, \frac{n^{-b\gamma-a/2}}{\sqrt{p_0}}\}$.
(ii) $\frac{\lambda_n}{\sqrt{n}} n^{\gamma/2} \geq \delta_1 n^{\delta_1} p_0$.
(iii) $p_0 = o(n^{1/2}(\log n)^{-3/2})$.

(A.7) There exists $C \in (0,\infty)$ and $\delta_2 \in (0, \gamma^{-1}\delta_1)$, δ_1 being defined in the assumption (A.6), such that

$$\mathbf{P}\big(\max\{\big|\sqrt{n}(\tilde{\beta}_{j,n} - \beta_{j,n})\big| : 1 \leq j \leq p\} > C.n^{\delta_2}\big) = o\big(n^{-1/2}\big),$$
$$\mathbf{P}_*\big(\max\{\big|\sqrt{n}(\tilde{\beta}^*_{j,n} - \hat{\beta}_{j,n})\big| : 1 \leq j \leq p\} > C.n^{\delta_2}\big) = o_p\big(n^{-1/2}\big).$$

Now we explain the assumptions briefly. Assumption (A.1) describes the regularity conditions needed on the growth of the design vectors. Assumption (A.1)(i) is a restriction on the smallest eigenvalue of $\boldsymbol{C}_{11,n}$. Assumption (A.1)(i) is a weaker condition than assuming that $\boldsymbol{C}_{11,n}$ converges to a positive definite matrix. (A.1)(ii) and (iii) are needed to bound the weighted sums of types $[\sum_{i=1}^n \boldsymbol{x}_i \epsilon_i]$, $[\sum_{i=1}^n \tilde{\boldsymbol{x}}_i \epsilon_i]$, $[\boldsymbol{C}_{11,n}^{-1} \sum_{i=1}^n \boldsymbol{x}_i^{(1)} \epsilon_i]$ (second one only when $p \leq n$). For $r = 2$, (A.1)(iii) is equivalent to the condition that the diagonal elements of the matrix $\boldsymbol{C}_n^{-1}$ are uniformly bounded. Also for general value of r, (A.1)(ii) and (iii) are much weaker than conditioning on l_r-norms of the design vectors. Here, the value of r is specified by the underlying Edgeworth expansion. Assumption (A.1)(iii) requires $p \leq n$, and hence is not defined when $p > n$. Note that the condition (A.1)(iii)$'$ needs $p_0 \leq n$ which is true in our setup due to assumption (A.6)(iii).

Assumptions (A.2)(i) bounds the eigenvalues of the matrix $\boldsymbol{D}_n^{(1)} \boldsymbol{C}_{11,n}^{-1} \boldsymbol{D}_n^{(1)\prime}$ away from infinity. It is necessary to obtain bounds needed in the Studentized setup. Assumption (A.2)(ii) is a condition similar to the conditions in (A.1)(ii) and (iii); but involving the $q \times p$ matrix $\boldsymbol{D}_n$. This condition is needed for showing necessary closeness of the covariance matrix estimators $\breve{\boldsymbol{\Sigma}}_n$, $\tilde{\boldsymbol{\Sigma}}_n$ (defined in Section 5) to their population counterparts (for details see Lemma 8.5). Assumption (A.2)(iii) bounds the minimum eigenvalue of the matrix S_n away from 0. This condition along with the Cramér conditions given in (A.4) and (A.5) enable certain Edgeworth expansions.

Assumption (A.3) separates the relevant covariates from the nonrelevant ones. The condition on the minimum is needed to ensure that the nonzero regression coefficients cannot converge to zero faster than the error rate, that is, not faster than $O(n^{-1/2})$. We mention that one can assume $b < 1/2$ instead of assuming $b < 1/4$, but with the price of putting another restriction on the penalty parameter λ_n. We do not consider such a setting here. We also want to point out that it is not possible to relax this minimal signal condition by the bias correction, considered in Section 5. With further relaxation, the bias of the Alasso estimator will be larger than the estimation error which is of order $O_p(n^{-1/2})$, and hence second-order correctness cannot be achieved by perturbation bootstrap in more relaxed minimal signal condition.

Assumption (A.4)(i) is a moment condition on the error term needed for valid Edgeworth expansion. Assumption (A.4)(ii) is Cramér's condition on the errors, which is very common in the literature of Edgeworth expansions; it is satisfied when the distribution of $(\epsilon_1, \epsilon_1^2)$ has a nondegenerate component which is absolutely continuous with respect to the Lebesgue measure [cf. Hall (1992)]. Assumption (A.4)(ii) is only needed to get a valid Edgeworth expansion for the original Alasso estimator in the Studentized setup. Assumptions (A.5)(i) and (iii) are the analogous conditions that are needed on the perturbing random quantities to get a valid Edgeworth expansion in the bootstrap setting. Assumption (A.5)(ii) is natural, since the ϵ_i are present already in the data generating process, whereas G_i^* are introduced by the user. One can look for Generalized Beta and Generalized Gamma families for suitable choices of the distribution of G^*. The p.d.f. of Generalized Beta family of distributions is

$$\begin{aligned} &\mathrm{GB}(y;\, f, g, h, \omega, \rho) \\ &\quad = \begin{cases} \dfrac{|f| y^{f\omega-1}(1-(1-c)(y/g)^f)^{\rho-1}}{g^{f\omega} B(\omega,\rho)(1+c(y/g)^f)^{\omega+\rho}} & \text{for } 0 < y^f < \dfrac{g^f}{1-h}, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

where $0 \le h \le 1$ and other parameters are all positive. We interpret $1/0$ as ∞. The function $B(\omega, \rho)$ is the beta function. Choices of the distribution of G^* can be obtained by finding solution of (f, g, h, ω, ρ) from the following two equations:

$$\begin{aligned} &\frac{B(\omega+2/f, \rho)}{B(\omega,\rho)} {}_2F_1[\omega+2/f, 2/f; h; \omega+\rho+2/f] \\ &\qquad = 2\biggl[\frac{B(\omega+1/f, \rho)}{B(\omega,\rho)} {}_2F_1[\omega+1/f, 1/f; h; \omega+\rho+1/f]\biggr]^2 \end{aligned}$$

and

$$\begin{aligned} &\frac{B(\omega+3/f, \rho)}{B(\omega,\rho)} {}_2F_1[\omega+3/f, 3/f; h; \omega+\rho+3/f] \\ &\qquad = 5\biggl[\frac{B(\omega+1/f, \rho)}{B(\omega,\rho)} {}_2F_1[\omega+1/f, 1/f; h; \omega+\rho+1/f]\biggr]^3, \end{aligned}$$

where ${}_2F_1$ denotes hypergeometric series. The p.d.f. of Generalized Gamma family of distributions is given by

$$\mathrm{GG}(y; \omega, \rho, \nu) = \begin{cases} \dfrac{(\nu/\omega^\rho) y^{\rho-1} e^{(y/\omega)^\nu}}{\Gamma(\rho/\nu)} & \text{for } y > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where all the parameters are positive and $\Gamma(\cdot)$ denotes the gamma function. For this family, the suitable choices of the distribution of G^* can be obtained by considering any positive value of the parameter ω and solving the following two equations for (ρ, ν):

$$[\Gamma((\rho+2)/\nu)] * \Gamma(\rho/\nu) = 2[\Gamma((\rho+1)/\nu)]^2$$

and

$$[\Gamma((\rho+3)/\nu)] * [\Gamma(\rho/\nu)]^2 = 5[\Gamma((\rho+1)/\nu)]^3.$$

One immediate choice of the distribution of G^* from Generalized Beta family is the Beta(α, β) distribution with $3\alpha = \beta = 3/2$. We have utilized this distribution as the distribution of the perturbing quantities G_i^*'s in our simulations, presented in Section 6. Outside these two generalized family of distributions, one possible choice is the distribution of $(M_1 + M_2)$ where M_1 and M_2 are independent and M_1 is a Gamma random variable with shape and scale parameters 0.008652 and 2, respectively, and M_2 is a Beta random variable with both the parameters 0.036490. Another possible choice is the distribution of $(M_3 + M_4)$ where M_3 and M_4 are independent and M_3 is an Exponential random variable with mean $(79 - 15\sqrt{33})/16$ and M_4 is an Inverse Gamma random variable with both shape and scale parameters $(4 + \sqrt{11/3})$.

Assumptions (A.6)(i) and (ii) can be compared with the condition (c) $\lambda_n/\sqrt{n} \to 0$ and $n^{\gamma/2}\lambda_n/\sqrt{n} \to \infty$ [cf. Zou (2006), Caner and Fan (2010)]. Whereas (c) is ensuring the oracle normal approximation, (A.6)(i) and (ii) are required for obtaining Edgeworth expansions. Lastly, (A.6)(iii) limits how quickly the number of nonzero regression coefficients may grow. Though it would seem that $p_0 = O(n)$ with $p_0 \leq n$ should be a sufficient restriction on the growth rate of p_0 for approximating the distribution of the Alasso estimator, a careful analysis reveals that further reduction in the growth rate of p_0 is necessary for accommodating the Studentization. Clearly, it is difficult to comprehend what possible choices of p_0, λ_n, γ, a, b would satisfy the assumptions presented in (A.6). Thus it is better to present some possible choices of those parameters.

First, consider $a = 0$ and $b = 0$, that is assume that the smallest eigenvalue of $\boldsymbol{C}_{11,n}$ and the smallest nonzero regression coefficients are bounded away from 0. In that case, it is easy to check that one set of possible choices are $p_0 = O(n^{\gamma/5})$ and $\lambda_n = C.n^{1/2-\gamma/4}$ for some constant $C > 0$, provided $\gamma \in (0, 2)$. In particular, if $\gamma = 1$ then the choices of p_0 and λ_n maybe respectively $p_0 = O(n^{1/5})$ and $\lambda_n = C.n^{1/4}$ when $a = b = 0$. Again p_0 can grow with n at the rate $o(n^{1/2}(\log n)^{-3/2})$, when $\gamma > 2$ and $\lambda_n = C.n^{(2-\gamma)/6}$ for some constant $C > 0$ whenever $a = b = 0$.

In general, if $a \in [0, 1/2)$ and $b < 1/4$, then it can be shown that the possible choices of γ, p_0 and λ_n are respectively $4a/(1-2b) < \gamma < 2/(1+2b)$, $p_0 = O(n^{[(1-2b)\gamma]/5})$ and $\lambda_n = C.n^{1/2-\gamma/4-b\gamma/2}$ for some constant $C > 0$. On the other hand, if $a \in [1/2, 1)$ and $a + 2b < 1$, one set of possible choices would be $\gamma \geq$

2, $p_0 = O(n^{2/3-(a+2b\gamma+4c)/3})$ and $\lambda_n = C.n^{1/6-(a+2b\gamma+c)/3}$ for some constants $c, C > 0$. With $a = 1/2$ and $b = 0$, clearly the choices of p_0 and λ_n reduce to $p_0 = O(n^{1/2-\delta})$ and $\lambda_n = C.n^{-\delta/4}$ for some $\delta, C > 0$.

Assumption (A.7) places deviation bounds on both the sample and bootstrap initial estimators which are needed to get valid Edgeworth expansions. These conditions are satisfied by OLS estimator in $p \le n$ case (cf. Lemma 8.2). Note that nonbootstrap part of (A.7) is satisfied if there exists a linear approximation of the type $\sum_{i=1}^{n} a_{i,j}\epsilon_i$ of $\sqrt{n}(\tilde{\beta}_{j,n} - \beta_{j,n})$, where $\max\{\sum_{i=1}^{n} |a_{i,j}|^r : 1 \le j \le p\} = o(p^{-1}n^{-1/2+r\delta_2})$ and $\mathbf{E}(|\epsilon_1|^r) < \infty$ for some $r \ge 3$. The bootstrap deviation bound corresponding to (A.7) holds provided similar approximation exists with $(G_1^* - \mu_{G^*})$ in place of ϵ_1. More precisely, for the Ridge estimator and for its perturbation bootstrap version defined in Section 2, if for some $r \ge 4$, the conditions

(a) $\mathbf{E}|\epsilon_1|^r + \mathbf{E}_*(G_1^*)^r < \infty$,

(b) $\max\{n^{-1}\sum_{i=1}^{n}(|\boldsymbol{x}_i|^{2r} + |\breve{\boldsymbol{x}}_i|^{2r}) : 1 \le j \le p\} = O(n^{\delta_2/2})$ for all $i \in \{1, \dots, n\}$,

(c) $\max\{\boldsymbol{e}_j'(\boldsymbol{C}_n + \tilde{\lambda}_n n^{-1}\boldsymbol{I}_p)^{-1}\boldsymbol{\beta}_n : 1 \le j \le p\} = O(n^{(1+\delta_2)/2}\tilde{\lambda}_n^{-1})$,

(d) $\sup\{\boldsymbol{e}_j'(\boldsymbol{C}_n + \tilde{\lambda}_n n^{-1}\boldsymbol{I}_p)^{-1}\boldsymbol{z}_n : \|\boldsymbol{z}_n\| \le 1\} = O(n^{(1+\delta_2)/2}\tilde{\lambda}_n^{-1})$ for all $j \in \{1, \dots, p\}$

are satisfied, then the assumption (A.7) holds. Here, $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_p\}$ is the standard basis of $\mathcal{R}^p$, $\breve{\boldsymbol{x}}_i = (\boldsymbol{C}_n + \tilde{\lambda}_n n^{-1}\boldsymbol{I}_p)^{-1}\boldsymbol{x}_i$ and $\tilde{\lambda}_n$ is the penalty parameter corresponding to the Ridge estimator (cf. Section 2). This follows analogously to Proposition 8.4 of Chatterjee and Lahiri (2013) after applying Lemma 8.1, stated in Section 8.

4. Impossibility of second-order correctness of the naive perturbation bootstrap. In this section, we describe the naive perturbation bootstrap as defined by MTC(11) for the Alasso and show that second-order correctness can not be achievable by their naive perturbation bootstrap method. When the objective function is the usual least squares criterion function, the naive perturbation bootstrap Alasso estimator $\boldsymbol{\beta}_n^{*N}$ is defined in MTC(11) as

$$\boldsymbol{\beta}_n^{*N} = \arg\min_{\boldsymbol{v}_n^*}\left[\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{v}_n^*)^2 G_i^* + \lambda_n^* \sum_{j=1}^{p}|\tilde{\beta}_{j,n}^{*N}|^{-\gamma}|v_{j,n}^*|\right], \tag{4.1}$$

where:

(i) $\lambda_n^* > 0$ is such that $\lambda_n^* n^{-1/2} \to 0$ and $\lambda_n^* \to \infty$ as $n \to \infty$;

(ii) the initial naive bootstrap estimator is defined as

$$\tilde{\boldsymbol{\beta}}_n^{*N} = \arg\min_{\boldsymbol{v}_n^*}\left[\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{v}_n^*)^2 G_i^*\right]$$

and $\tilde{\beta}_{j,n}^{*N}$ is the jth component of $\tilde{\boldsymbol{\beta}}_n^{*N}$;

(iii) $\{G_1^*, \dots, G_n^*\}$ is a set of i.i.d. nonnegative random quantities with mean and variance both equal to 1.

Note that the initial estimator $\tilde{\boldsymbol{\beta}}_n^{*N}$ is unique only when p is less than or equal to n. We now consider the quantity $\boldsymbol{u}_n^{*N} = \sqrt{n}(\boldsymbol{\beta}_n^{*N} - \hat{\boldsymbol{\beta}}_n)$, which we can show from (4.1) to be the minimizer

$$\begin{aligned}\boldsymbol{u}_n^{*N} = \arg\min_{\boldsymbol{w}_n^*}\Bigg[&\boldsymbol{w}_n^{*'}\boldsymbol{C}_n^*\boldsymbol{w}_n^* - 2\boldsymbol{w}_n^*\boldsymbol{W}_n^* \\ &+ \lambda_n^*\sum_{j=1}^{p}|\tilde{\beta}_{j,n}^{*N}|^{-\gamma}\left(\left|\hat{\beta}_{j,n} + \frac{w_{j,n}^*}{\sqrt{n}}\right| - |\hat{\beta}_{j,n}|\right)\Bigg],\end{aligned} \tag{4.2}$$

where $\hat{\beta}_{j,n}$ is the jth component of the Alasso estimator $\hat{\boldsymbol{\beta}}_n$, $\boldsymbol{C}_n^* = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i \times \boldsymbol{x}_i' G_i^*$, and $\boldsymbol{W}_n^* = n^{-1/2}\sum_{i=1}^n \hat{\epsilon}_i \boldsymbol{x}_i G_i^*$. To describe the solution of MTC(11), assume $\mathcal{A} = \{j : \beta_j \neq 0\} = \{1, \dots, p_0\}$. MTC(11) claimed that when $\gamma = 1$ and p is fixed, $((\boldsymbol{u}_{n1}^{*N})', \mathbf{0})'$ is a solution of (4.2) for sufficiently large n, where

$$\boldsymbol{u}_{n1}^{*N} = \boldsymbol{C}_{11,n}^{-1} n^{-1/2}\sum_{i=1}^n \epsilon_i \boldsymbol{x}_i^{(1)}(G_i^* - 1) \quad \text{and} \quad \|\boldsymbol{u}_n^{*N} - ((\boldsymbol{u}_{n1}^{*N})', \mathbf{0})'\|_\infty = o_{p_*}(1).$$

However, to achieve second-order correctness, we need to obtain a solution $((\boldsymbol{u}_{n2}^{*N})', \mathbf{0})'$ of (4.2) such that $\|\boldsymbol{u}_n^{*N} - ((\boldsymbol{u}_{n2}^{*N})', \mathbf{0})'\|_\infty = o_{p_*}(n^{-1/2})$. We show that such an $\boldsymbol{u}_{n2}^{*N}$ has the form

$$\boldsymbol{u}_{n2}^{*N} = \boldsymbol{C}_{11,n}^{*-1}\left[\boldsymbol{W}_n^{*(1)} - \frac{\lambda_n^*}{\sqrt{n}}\tilde{\boldsymbol{s}}_n^{*(1)}\right]$$

for sufficiently large n, where $\boldsymbol{W}_n^{*(1)}$ is the first p_0 components of $\boldsymbol{W}_n^*$ and the jth component of $\tilde{\boldsymbol{s}}_n^{*(1)}$ equals to $\operatorname{sgn}(\hat{\beta}_{j,n})\|\tilde{\beta}_{jn}^{*N}|^{-\gamma}$, $j \in \mathcal{A}$ (here, we drop the subscript n from the notation of true parameter values since we are considering p to be fixed in this section). We establish this fact by exploring the KKT condition corresponding to (4.2), which is given by

$$2\boldsymbol{C}_n^*\boldsymbol{w}_n^* - 2\boldsymbol{W}_n^* + \frac{\lambda_n^*}{\sqrt{n}}\boldsymbol{\Gamma}_n^*\boldsymbol{l}_n = \mathbf{0}, \tag{4.3}$$

for some $\boldsymbol{l}_n = (l_{1n}, \dots, l_{pn})'$ with $l_{j,n} \in [-1, 1]$ for $j = 1, \dots, p$ and $\boldsymbol{\Gamma}_n^* = \operatorname{diag}(|\tilde{\beta}_{1n}^{*N}|^{-\gamma}, \dots, |\tilde{\beta}_{pn}^{*N}|^{-\gamma})$. Since $\boldsymbol{C}_n^*$ is a nonnegative definite matrix, (4.2) is a convex optimization problem; hence (4.3) is both necessary and sufficient in solving (4.2).

Note that $\boldsymbol{W}_n^*$ is not centered and hence we need to adjust the solution $((\boldsymbol{u}_{n2}^{*N})', \mathbf{0})'$ for centering before investigating if the naive perturbation bootstrap can asymptotically correct the distribution of Alasso up to second order. Clearly, the centering adjustment term is $\boldsymbol{Ad}_n^* = (\boldsymbol{Ad}_n^{*(1)'}, \mathbf{0}')'$ where $\boldsymbol{Ad}_n^{*(1)} =$

$\boldsymbol{C}_{11,n}^{*-1} n^{-1/2} \sum_{i=1}^{n} \hat{\epsilon}_i \boldsymbol{x}_i^{(1)}$. It follows from the steps of the proofs of the results of Section 5 that we need $\|\boldsymbol{A}\boldsymbol{d}_n^*\| = o_{p_*}(n^{-1/2})$ to achieve second-order correctness. We show that this is indeed not the case even in the fixed p setting.

More precisely, we negate the second-order correctness of the naive perturbation bootstrap of MTC(11) by first showing that $((\boldsymbol{u}_{n2}^{*N})', \boldsymbol{0}')'$ satisfies the KKT condition (4.3) exactly with bootstrap probability converging to 1. Then we show that $\sqrt{n}\|\boldsymbol{A}\boldsymbol{d}_n^*\|$ diverges in bootstrap probability to ∞, which in turn implies that the conditional c.d.f. of $\boldsymbol{F}_n^{*N} = \sqrt{n}(\boldsymbol{\beta}_n^{*N} - \hat{\boldsymbol{\beta}}_n)$ can not approximate the c.d.f. of $\boldsymbol{F}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ with the uniform accuracy $O_p(n^{-1/2})$, needed for the validity of second-order correctness. We formalize these arguments in the following theorem.

THEOREM 4.1. *Let p be fixed and $\boldsymbol{C}_n \to \boldsymbol{C}$, a positive definite matrix. Define* $Z_n^{*-1} = \sqrt{n}\|\boldsymbol{A}\boldsymbol{d}_n^*\|$. *Suppose*, $(\log n/n)^{1/2}. \max\{\lambda_n, \lambda_n^*\} \to 0$ *and* $(\log n)^{-(\gamma+1)/2}. \min\{\lambda_n, \lambda_n^*\}. \min\{1, n^{(\gamma-1)/2}\} \to \infty$ *as* $n \to \infty$. *Also assume that* (A.1)(i), (ii) *and* (A.4)(i) *hold with* $r = 4$. *Then there exists a sequence of Borel sets* $\{\boldsymbol{A}_n\}_{n\geq 1}$ *with* $\mathbf{P}(\boldsymbol{\varepsilon}_n \in \boldsymbol{A}_n) \to 1$ *and given* $\boldsymbol{\varepsilon}_n = (\epsilon_1, \ldots, \epsilon_n)' \in \boldsymbol{A}_n$, *the following conclusions hold*:

(a) $\mathbf{P}_*(\boldsymbol{u}_n^{*N} = ((\boldsymbol{u}_{n2}^{*N})', \boldsymbol{0}')') = 1 - o(n^{-1/2})$.

(b) $\mathbf{P}_*(Z_n^* > \epsilon) = o(n^{-1/2})$ *for any* $\epsilon > 0$.

(c) $\sup_{\boldsymbol{x}\in\mathcal{R}^p} |\mathbf{P}_*(\boldsymbol{F}_n^{*N} \leq \boldsymbol{x}) - \mathbf{P}(\boldsymbol{F}_n \leq \boldsymbol{x})| \geq K.\dfrac{\lambda_n}{\sqrt{n}}$ *for some* $K > 0$.

REMARK 1. Theorem 4.1(a), (b) state that the naive perturbation bootstrap is incompetent in approximating the distribution of Alasso up to second order. The fundamental reason behind second-order incorrectness is the inadequate centering in the form of $\sqrt{n}(\boldsymbol{\beta}_n^{*N} - \hat{\boldsymbol{\beta}}_n)$. Although the adjustment term necessary for centering is $o_{p_*}(1)$, which essentially helps to establish distributional consistency in MTC(11), the term is coarser than $n^{-1/2}$, leading to second-order incorrectness. Additionally, it is worth mentioning that Studentization will also not help in achieving second-order correctness by naive perturbation bootstrap of MTC(11), since the necessary centering correction cannot be accomplished by any sort of Studentization. Part (c) conveys uniformly how far the naive bootstrap c.d.f. is from the original c.d.f.

5. Modified perturbation bootstrap and its higher order properties. This section is divided into two subsections. The first one describes briefly the motivation behind considering the perturbation bootstrap modification in Alasso. The second subsection describes higher order asymptotic properties of our modified perturbation bootstrap method.

5.1. *Motivation for the modified perturbation bootstrap*. Theorem 4.1 establishes that the naive perturbation bootstrap of MTC(11) does not provide a solution for approximating the distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$ up to second order. As it is mentioned earlier, the problem occurs because $\boldsymbol{W}_n^*$ is not centered. Let $\breve{\boldsymbol{W}}_n^*$ denotes the centered version of $\boldsymbol{W}_n^*$, that is $\breve{\boldsymbol{W}}_n^* = n^{-1/2}\sum_{i=1}^n \hat{\epsilon}_i \boldsymbol{x}_i(G_i^* - \mu_{G^*})$, and consider the vector equation

$$2\boldsymbol{C}_n^*\boldsymbol{w}_n^* - 2\breve{\boldsymbol{W}}_n^* + \frac{\lambda_n^*}{\sqrt{n}}\boldsymbol{\Gamma}_n^*\boldsymbol{l}_n = \boldsymbol{0}, \tag{5.1}$$

which is same as (4.3) after replacing $\boldsymbol{W}_n^*$ with $\breve{\boldsymbol{W}}_n^*$. Note that the solution to (5.1) is of the form $((\boldsymbol{u}_{n3}^{*(1)})', \boldsymbol{0}')'$, where $\boldsymbol{u}_{n3}^{*(1)} = \boldsymbol{C}_{11,n}^{*-1}[\breve{\boldsymbol{W}}_n^{*(1)} - \frac{\lambda_n^*}{\sqrt{n}}\tilde{\boldsymbol{s}}_n^{*(1)}]$. Although this form is adequate for achieving second-order correctness in fixed dimension, there are some computational and higher-dimensional issues that we now address.

Note that $\boldsymbol{C}_{11,n}^*$ is a matrix involving random quantities $\{G_1^*, \ldots, G_n^*\}$. Thus $\boldsymbol{C}_{11,n}^*$ will not remain same for each bootstrap iteration, and hence each bootstrap iteration will require computing the inverse of $\boldsymbol{C}_{11,n}^*$ afresh. This is computationally expensive and the expense increases as the number of nonzero regression parameters increases. Therefore, it will be computationally advantageous if we can replace $\boldsymbol{C}_{11,n}^*$ by $\boldsymbol{C}_{11,n}$ in the form of $\boldsymbol{u}_{n3}^{*(1)}$.

Now define, $\boldsymbol{u}_{n4}^{*(1)} = \boldsymbol{C}_{11,n}^{-1}[\breve{\boldsymbol{W}}_n^{*(1)} - \frac{\lambda_n^*}{\sqrt{n}}\tilde{\boldsymbol{s}}_n^{*(1)}]$. If we look closely at the bias term $-\frac{\lambda_n^*}{\sqrt{n}}\boldsymbol{C}_{11,n}^{-1}\tilde{\boldsymbol{s}}_n^{*(1)}$, then it is clear that the primary contribution of the bias toward $\boldsymbol{u}_{n4}^{*(1)}$ is $-\frac{\lambda_n^*}{\sqrt{n}}\boldsymbol{C}_{11,n}^{-1}\tilde{\boldsymbol{s}}_n^{(1)}$, where jth component of $\tilde{\boldsymbol{s}}_n^{(1)}$ is equal to $\operatorname{sgn}(\hat{\beta}_{j,n})\|\tilde{\beta}_{jn}|^{-\gamma}$, $j \in \mathcal{A}$, where $\tilde{\beta}_{j,n}$ is the jth component of the OLS estimator $\tilde{\boldsymbol{\beta}}_n$. By Taylor's expansion, $(\tilde{\boldsymbol{s}}_n^{*(1)} - \tilde{\boldsymbol{s}}_n^{(1)})$ depends on the OLS residuals. The OLS residuals again depend on all p estimated regression parameters, unlike Alasso residuals which depend only on the estimates of the p_0 nonzero components. Since it is needed to bound $\|\tilde{\boldsymbol{s}}_n^{*(1)} - \tilde{\boldsymbol{s}}_n^{(1)}\|_\infty$ for achieving valid edgeworth expansion, we will come up with an implicit bound on the dimension p, which we do not want to impose. On the other hand, if the difference depends on Alasso residuals instead of OLS ones, then the implicit condition will be on p_0 and this is reasonable as p_0 can be much smaller than p. Additionally, $\tilde{\boldsymbol{\beta}}_n^{*N}$ involves inversion of the random matrix $\boldsymbol{C}_n^*$ and hence it is computationally expensive. Thus if $\boldsymbol{C}_n^*$ can be replaced by some fixed matrix, say $\boldsymbol{C}_n$, then the bootstrap will be computationally advantageous.

However, if we implement the modification described in Section 2, then both the theoretical and computational shortcomings of the perturbation bootstrap method become resolved and the second-order correctness is achieved even in increasing dimension under some mild regularity conditions. Additionally, we also have the nice structure due to the modification, which enables us to employ existing computational algorithms, as pointed out in Proposition 2.1.

5.2. *Higher order results.* Define $\boldsymbol{T}_n = \sqrt{n}\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$. Without loss of generality, we assume that $\mathcal{A}_n = \{j : \beta_{j,n} \neq 0\} = \{1, \dots, p_0\}$. Hence, by Taylor's expansion it is immediate from the form of Alasso estimator that $\boldsymbol{\Sigma}_n = n^{-1}\sum_{i=1}^n(\boldsymbol{\xi}_i^{(0)} + \boldsymbol{\eta}_i^{(0)})(\boldsymbol{\xi}_i^{(0)} + \boldsymbol{\eta}_i^{(0)})'$ or $\bar{\boldsymbol{\Sigma}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{\xi}_i^{(0)}\boldsymbol{\xi}_i'^{(0)}$ can be considered as the asymptotic variance of $\boldsymbol{T}_n/\sigma$ at sample size n. Here, $\boldsymbol{\xi_i}^{(0)} = \boldsymbol{D}_n^{(1)}\boldsymbol{C}_{11,n}^{-1}\boldsymbol{x}_i^{(1)}$, $\boldsymbol{\eta}_i^{(0)} = \boldsymbol{D}_n^{(1)}\boldsymbol{C}_{11,n}^{-1}\boldsymbol{\eta}_i$. For each $i \in \{1, \dots, n\}$, $\boldsymbol{\eta}_i$ is a $p_0 \times 1$ vector with jth element $(\frac{\lambda_n}{2n}\tilde{x}_{i,j}\frac{\gamma}{|\beta_{j,n}|^{\gamma+1}}\operatorname{sgn}(\beta_{j,n}))$ where $\tilde{\boldsymbol{x}}_i = \boldsymbol{C}_n^{-1}\boldsymbol{x}_i$ (when $p \le n$) and $\operatorname{sgn}(x) = -1, 0, 1$ according as $x < 0$, $x = 0$, $x > 0$, respectively, as defined earlier. The bias corresponding to $\boldsymbol{T}_n$ is $-\boldsymbol{b}_n = -\boldsymbol{D}_n^{(1)}\boldsymbol{C}_{11,n}^{-1}\boldsymbol{s}_n^{(1)}\frac{\lambda_n}{2\sqrt{n}}$, where $\boldsymbol{D}_n^{(1)}$ and $\boldsymbol{C}_{11,n}$ are as defined earlier and $\boldsymbol{s}_n^{(1)}$ is a $p_0 \times 1$ vector with jth element $\operatorname{sgn}(\beta_{j,n})|\beta_{j,n}|^{-\gamma}$. Although $\bar{\boldsymbol{\Sigma}}_n$ is defined for all p, $\boldsymbol{\Sigma}_n$ is only defined when $p \le n$. $\bar{\boldsymbol{\Sigma}}_n$ is also the asymptotic variance of $[\boldsymbol{T}_n + \boldsymbol{b}_n]/\sigma$.

Define the set $\hat{\mathcal{A}}_n = \{j : \hat{\beta}_{j,n} \neq 0\}$ and $\hat{p}_{0,n} = |\hat{\mathcal{A}}_n|$, supposing, without loss of generality, that $\hat{\mathcal{A}}_n = \{1, \dots, \hat{p}_{0,n}\}$. We then partition the matrix $\boldsymbol{C}_n = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'$ as

$$\boldsymbol{C}_n = \begin{bmatrix} \hat{\boldsymbol{C}}_{11,n} & \hat{\boldsymbol{C}}_{12,n} \\ \hat{\boldsymbol{C}}_{21,n} & \hat{\boldsymbol{C}}_{22,n} \end{bmatrix},$$

where $\hat{\boldsymbol{C}}_{11,n}$ is of dimension $\hat{p}_{0,n} \times \hat{p}_{0,n}$. Similarly, we define $\hat{\boldsymbol{D}}_n^{(1)}$ as the matrix containing the first $\hat{p}_{0,n}$ columns of $\boldsymbol{D}_n$ and we define $\hat{\boldsymbol{x}}_i^{(1)}$ as the vector containing the first $\hat{p}_{0,n}$ entries of $\boldsymbol{x}_i$. Hence, the bias-correction term $\breve{\boldsymbol{b}}_n$ corresponding to $\boldsymbol{T}_n$ can be defined as

$$\breve{\boldsymbol{b}}_n = \hat{\boldsymbol{D}}_n^{(1)}\hat{\boldsymbol{C}}_{11,n}^{-1}\hat{\boldsymbol{s}}_n^{(1)}\frac{\lambda_n}{2\sqrt{n}},$$

where $\hat{\boldsymbol{s}}_n^{(1)}$ is the $\hat{p}_{0,n} \times 1$ vector with jth entry equal to $\operatorname{sgn}(\hat{\beta}_{j,n})|\tilde{\beta}_{j,n}|^{-\gamma}$, $j \in \hat{\mathcal{A}}_n$.

Therefore, the Studentized pivots can be constructed as

$$\mathbf{R}_n = \begin{cases} \hat{\sigma}_n^{-1}\hat{\boldsymbol{\Sigma}}_n^{-1/2}\boldsymbol{T}_n & \text{for } p \le n, \\ \hat{\sigma}_n^{-1}\check{\boldsymbol{\Sigma}}_n^{-1/2}\boldsymbol{T}_n & \text{for } p > n, \end{cases} \quad \text{and} \quad \check{\mathbf{R}}_n = \check{\sigma}_n^{-1}\check{\boldsymbol{\Sigma}}_n^{-1/2}[\boldsymbol{T}_n + \breve{\boldsymbol{b}}_n],$$

where the matrices $\hat{\boldsymbol{\Sigma}}_n$ and $\check{\boldsymbol{\Sigma}}_n$ have the form

$$\hat{\boldsymbol{\Sigma}}_n = n^{-1}\sum_{i=1}^n(\hat{\boldsymbol{\xi}}_i^{(0)} + \hat{\boldsymbol{\eta}}_i^{(0)})(\hat{\boldsymbol{\xi}}_i^{(0)} + \hat{\boldsymbol{\eta}}_i^{(0)})' \quad \text{and} \quad \check{\boldsymbol{\Sigma}}_n = n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i^{(0)}\hat{\boldsymbol{\xi}}_i^{(0)'} \tag{5.2}$$

and

$$\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n \hat{\epsilon}_i^2 \quad \text{and} \quad \check{\sigma}_n^2 = n^{-1}\sum_{i=1}^n \tilde{\epsilon}_i^2,$$

where $\hat{\epsilon}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n$, $\tilde{\epsilon}_i = y_i - \sum_{j\in\hat{\mathcal{A}}_n} x_{ij}\tilde{\beta}_{j,n}$, $\hat{\boldsymbol{\xi}}_i^{(0)} = \hat{\boldsymbol{D}}_n^{(1)}\hat{\boldsymbol{C}}_{11,n}^{-1}\hat{\boldsymbol{x}}_i^{(1)}$ and $\hat{\boldsymbol{\eta}}_i^{(0)} = \hat{\boldsymbol{D}}_n^{(1)}\hat{\boldsymbol{C}}_{11,n}^{-1}\hat{\boldsymbol{\eta}}_i$, with

$$\hat{\boldsymbol{\eta}}_i = \left(\frac{\lambda_n}{2n}\tilde{x}_{i,j}\frac{\gamma}{|\hat{\beta}_{j,n}|^{\gamma+1}}\operatorname{sgn}(\hat{\beta}_{j,n})\right)_{j\in\hat{\mathcal{A}}_n}.$$

We construct perturbation bootstrap versions $\boldsymbol{R}_n^*$ and $\check{\boldsymbol{R}}_n^*$ of $\boldsymbol{R}_n$ and $\check{\boldsymbol{R}}_n$ first by replacing $\boldsymbol{T}_n$ with $\boldsymbol{T}_n^* = \sqrt{n}\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n)$. We then replace $\hat{\boldsymbol{\Sigma}}_n$ and $\check{\boldsymbol{\Sigma}}_n$ with $\breve{\boldsymbol{\Sigma}}_n$ and $\tilde{\boldsymbol{\Sigma}}_n$, respectively, which we define by replacing $\hat{\boldsymbol{\xi}}_i^{(0)}$ with $\breve{\boldsymbol{\xi}}_i^{(0)} = \hat{\boldsymbol{\xi}}_i^{(0)}\hat{\epsilon}_i$ and $\hat{\boldsymbol{\eta}}_i^{(0)}$ with $\breve{\boldsymbol{\eta}}_i^{(0)} = \hat{\boldsymbol{\eta}}_i^{(0)}\hat{\epsilon}_i$ in (5.2). We replace $\check{\boldsymbol{b}}_n$ with $\check{\boldsymbol{b}}_n^* = \hat{\boldsymbol{D}}_n^{*(1)}\hat{\boldsymbol{C}}_{11,n}^{*-1}\hat{\boldsymbol{s}}_n^{*(1)}\lambda_n/(2\sqrt{n})$, where $\hat{\boldsymbol{s}}_n^{*(1)}$ is the $|\hat{\mathcal{A}}_n^*|\times 1$ vector with jth entry equal to $\operatorname{sgn}(\hat{\beta}_{j,n}^*)|\tilde{\beta}_{j,n}^*|^{-\gamma}$, $j\in\hat{\mathcal{A}}_n^* = \{j : \hat{\boldsymbol{\beta}}_{j,n}^* \neq 0\}$. The matrix $\hat{\boldsymbol{C}}_{11,n}^*$ is the $|\hat{\mathcal{A}}_n^*|\times|\hat{\mathcal{A}}_n^*|$ submatrix of $\boldsymbol{C}_n$ with rows and columns in $\hat{\mathcal{A}}_n^*$ and $\hat{\boldsymbol{D}}_n^{*(1)}$ is the $q\times|\hat{\mathcal{A}}_n^*|$ submatrix of $\boldsymbol{D}_n$ with columns in $\hat{\mathcal{A}}_n^*$. Lastly, we need

$$\hat{\sigma}_n^{*2} = n^{-1}\mu_{G^*}^{-2}\sum_{i=1}^n \hat{\epsilon}_i^{*2}(G_i^* - \mu_{G^*})^2 \quad\text{and}\quad \breve{\sigma}_n^{*2} = n^{-1}\mu_{G^*}^{-2}\sum_{i=1}^n \tilde{\epsilon}_i^{*2}(G_i^* - \mu_{G^*})^2,$$

where $\hat{\epsilon}_i^* = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n^*$, $\tilde{\epsilon}_i^* = y_i - \sum_{j\in\hat{\mathcal{A}}_n^*} x_{ij}\tilde{\beta}_{j,n}^*$. With these, we construct $\mathbf{R}_n^*$ and $\check{\mathbf{R}}_n^*$ as

$$\mathbf{R}_n^* = \begin{cases}\hat{\sigma}_n^{*-1}\hat{\sigma}_n\breve{\boldsymbol{\Sigma}}_n^{-1/2}\boldsymbol{T}_n^* & \text{for } p\le n,\\ \hat{\sigma}_n^{*-1}\hat{\sigma}_n\tilde{\boldsymbol{\Sigma}}_n^{-1/2}\boldsymbol{T}_n^* & \text{for } p> n,\end{cases} \quad\text{and}\quad \check{\mathbf{R}}_n^* = \breve{\sigma}_n^{*-1}\breve{\sigma}_n\tilde{\boldsymbol{\Sigma}}_n^{-1/2}[\boldsymbol{T}_n^* + \check{\boldsymbol{b}}_n^*].$$

We are motivated to look at these Studentized or pivot quantities by the fact that studentization improves the rate of convergence of bootstrap estimators in many settings [cf. Hall (1992)].

5.2.1. *Results for $p \le n$.*

THEOREM 5.1. *Let* (A.1)–(A.6) *hold with* $r = 6$. *Then*

$$\sup_{B\in\mathcal{C}_q}\left|\mathbf{P}_*(\mathbf{R}_n^*\in B) - \mathbf{P}(\mathbf{R}_n\in B)\right| = o_p(n^{-1/2}).$$

Theorem 5.1 shows that after proper Studentization the modified perturbation bootstrap approximation of the distribution of the Alasso estimator is second-order correct. The error rate reduces to $o_p(n^{-1/2})$ from $O(n^{-1/2})$, the best possible rate obtained by the oracle Normal approximation. This is a significant improvement from the perspective of inference. As a consequence, the precision of the percentile confidence intervals based on $\mathbf{R}_n^*$ will be greater than that of confidence intervals based on the oracle Normal approximation.

We point out that the error rate in Theorem 5.1 cannot be reduced to the optimal rate of $O_p(n^{-1})$, unlike in the fixed-dimension case. To achieve this optimal rate by our modified bootstrap method, we now consider a bias corrected pivot $\check{\mathbf{R}}_n$ and its modified perturbation bootstrap version $\check{\mathbf{R}}_n^*$. The following theorem states that it achieves the optimal rate.

THEOREM 5.2. *Let* (A.1)–(A.6) *hold with* $r = 8$. *Then*

$$\sup_{B \in \mathcal{C}_q} \left| \mathbf{P}_*\left(\check{\mathbf{R}}_n^* \in B\right) - \mathbf{P}(\check{\mathbf{R}}_n \in B)\right| = O_p\left(n^{-1}\right).$$

Theorem 5.2 suggests that the modified perturbation bootstrap achieves notable improvement in the error rate over the oracle Normal approximation irrespective of the order of the bias term. Thus Theorem 5.2 establishes the perturbation bootstrap method as an effective method for approximating the distribution of the Alasso estimator when $p \leq n$.

5.2.2. *Results for* $p > n$. We now present results for the quality of perturbation bootstrap approximation when the dimension p of the regression parameter can be much larger than the sample size n. We consider the initial estimator $\tilde{\boldsymbol{\beta}}_n$ to be some bridge estimator, for example, Lasso or Ridge estimator, in defining the Alasso estimator by (1.2). The bootstrap version of Lasso or Ridge is defined by (2.2). Higher order results are presented separately for two cases based on growth of p with sample size n. First, we consider the case when p can grow polynomially and then we move to the situation when p can grow exponentially.

5.2.2.1. *p grows polynomially.*

THEOREM 5.3. *Let* (A.1)(i), (ii), (iii)$'$ *and* (A.2)–(A.6) *and* (A.7) *hold and* $p = O(n^{(r-3)/2})$ *for some positive integer* $r \geq 3$. *Now if* $b = 0$ [*cf. condition* (A.3) *in Section* 3] *and* $r \geq 8$, *then we have*

$$\sup_{B \in \mathcal{C}_q} \left| \mathbf{P}_*\left(\mathbf{R}_n^* \in B\right) - \mathbf{P}(\mathbf{R}_n \in B)\right| = o_p\left(n^{-1/2}\right),$$

$$\sup_{B \in \mathcal{C}_q} \left| \mathbf{P}_*\left(\check{\mathbf{R}}_n^* \in B\right) - \mathbf{P}(\check{\mathbf{R}}_n \in B)\right| = o_p\left(n^{-1/2}\right).$$

Theorem 5.3 states that our proposed modified perturbation bootstrap approximation is second-order correct, even when p grows polynomially with n. The error rate obtained by our proposed method is significantly better than $O(n^{-1/2})$, which is the best-attainable rate of the oracle Normal approximation. When p can grow at a polynomial rate with n, the validity of our method depends on the existence of some polynomial moment of the error distribution. To see why, note that it is essential to have

$$\mathbf{P}\Big(\max_{1\le j\le p}|\breve{W}_{j,n}| > K.\sqrt{\log n}\Big) = o\big(n^{-1/2}\big) \quad \text{and}$$
$$\mathbf{P}_*\Big(\max_{1\le j\le p}|\breve{W}^*_{j,n}| > K.\sqrt{\log n}\Big) = o_p\big(n^{-1/2}\big) \tag{5.3}$$

to obtain second-order correctness, as presented in Theorem 5.3. Here, $K \in (1,\infty)$ is a constant, $\breve{W}_{j,n} = n^{-1/2}\sum_{i=1}^n \epsilon_i x_{i,j}$ and $\breve{W}^*_{j,n} = n^{-1/2}\sum_{i=1}^n \hat{\epsilon}_i x_{i,j}(G^*_i - \mu_{G^*})$. In view of Lemma 8.1, the following bound is needed to conclude (5.3):

$$p.\left(\max_{1\le j\le p}\left[\sum_{i=1}^n |x_{i,j}|^{2r}\right]\right)\big(\mathbf{E}|\epsilon_1|^r\big)^2 = o\big(n^{(r-1)/2}(\log n)^{r/2}\big).$$

Clearly, under the assumption $\max\{n^{-1}\sum_{i=1}^n |x_{i,j}|^{2r} : 1 \le j \le p\} = O(1)$ [cf. condition (A.1) (ii)], we must have $p = o(n^{(r-3)/2}(\log n)^{r/2})$ provided $\mathbf{E}|\epsilon_1|^r < \infty$. Therefore, in view of condition (A.1)(ii), p can grow like $(a_n.n^l.(\log n)^{l+3/2})$ where $a_n \to 0$ as $n \to \infty$, provided $\mathbf{E}|\epsilon_1|^{2l+3} < \infty$. This implies that p can grow polynomially with n under the assumption that some polynomial moment of the error distribution exists.

5.2.2.2. *p grows exponentially.* When p grows exponentially with some fractional power of n, existence of polynomial moment of some order of regression errors ϵ_i's [cf. condition (A.4)(i)] is not enough to achieve higher order accuracy. Indeed, we need to have some control over the moment generating function of the error variable. Following two important cases are considered in this setting.

Errors are sub-Gaussian: Suppose error ϵ_1 is sub-Gaussian. This means that there exists $d > 0$ such that

$$\mathbf{E}\big[e^{\kappa\epsilon_1}\big] \le e^{\kappa^2 d^2/2} \qquad \text{for all } \kappa \in \mathcal{R}. \tag{5.4}$$

When the regression errors have sub-Gaussian tails, we need to choose the perturbing quantities G^*_i's effectively to have sub-Gaussian tails, that is, there exists $d^* > 0$ such that

$$\mathbf{E}_*\big[e^{\kappa(G^*_1-\mu_{G^*})}\big] \le e^{\kappa^2 d^{*2}/2} \qquad \text{for all } \kappa \in \mathcal{R}. \tag{5.5}$$

THEOREM 5.4. *Let* (A.1)(i), (ii), (iii)$'$ *and* (A.2)–(A.6) *and* (A.7) *hold with* $r = 8$ *and* $b = 0$. *Also assume that* (5.4) *and* (5.5) *hold and* $p = O(\exp(n^{(\delta_1-\gamma\delta_2)}))$ *where* δ_1 *and* δ_2 *are defined in assumptions* (A.6) *and* (A.7) *in Section* 3. *Then the conclusions of Theorem* 5.3 *hold.*

Errors are subexponential: Consider the regression errors to be sub-exponential, that is there exist positive parameters d, h such that

$$\mathbf{E}\big[e^{\kappa\epsilon_1}\big] \le e^{\kappa^2 d^2/2} \qquad \text{for all } |\kappa| < 1/h. \tag{5.6}$$

Similar to sub-Gaussian case, we need to choose the perturbing quantities G_i^*'s to be subexponential besides the errors being subexponential, that is, there exist positive parameters d^*, h^* such that

$$\mathbf{E}_*\big[e^{\kappa(G_1^*-\mu_{G^*})}\big] \le e^{\kappa^2 d^{*2}/2} \qquad \text{for all } |\kappa| < 1/h^*. \tag{5.7}$$

THEOREM 5.5. *Let* (A.1)(i), (ii), (iii)′ *and* (A.2)–(A.5), (A.6)(i), (ii) *and* (A.7) *hold with* $r = 8$ *and* $b = 0$. *Also assume that* (5.6) *and* (5.7) *hold.*

(a) *If* $p = O(\exp(n^{(\delta_1-\gamma\delta_2)}))$ *and* $p_0 = O(n^{(1-\delta_1+\gamma\delta_2)/2})$ *are satisfied where* δ_1 *and* δ_2 *are defined in assumptions* (A.6) *and* (A.7) *in Section* 3, *then the conclusions of Theorem* 5.3 *hold.*

(b) *If* $p = O(\exp(n))$, $n^{(-\delta_1+\gamma\delta_2)} = o(p_0^2/n)$ *and* $p_0/\sqrt{n} = o((\log n)^{-3/2})$ *are satisfied where* δ_1 *and* δ_2 *are defined in assumptions* (A.6) *and* (A.7) *in Section* 3, *then the conclusions of Theorem* 5.3 *hold.*

Theorem 5.4 and 5.5 show that our perturbation bootstrap method remains valid as a second-order correct method even when the dimension p grows exponentially with some fractional power of n. Moreover, we can achieve exponential growth of p in some situations when errors are sub-exponential, as stated in part (b) of Theorem 5.5. To obtain higher order results stated in Theorem 5.4 and Theorem 5.5, we need to relax (5.3) a bit for $j = p_0 + 1, \dots, p$. It follows from the proofs and condition (A.6)(ii) that we can relax (5.3) for $j = p_0 + 1, \dots, p$, to the following:

$$\begin{aligned} &\mathbf{P}\Big(\max_{p_0+1\le j\le p} |\breve{W}_{j,n}| > K.n^{(\delta_1-\gamma\delta_2)}.p_0\Big) = o\big(n^{-1/2}\big) \quad \text{and} \\ &\mathbf{P}_*\Big(\max_{p_0+1\le j\le p} |\breve{W}_{j,n}^*| > K.n^{(\delta_1-\gamma\delta_2)}.p_0\Big) = o_p\big(n^{-1/2}\big), \end{aligned} \tag{5.8}$$

keeping higher order results valid. Now consider using Hoeffding's inequality in the sub-Gaussian case and Bernstein's inequality in the subexponential case. As a result, the following two bounds are needed respectively in sub-Gaussian and subexponential case to conclude (5.8)

$$\begin{aligned} &p.\exp\Big(-\frac{C_1.n^{1+2(\delta_1-\gamma\delta_2)}.p_0^2}{2.\max_{1\le j\le p}[\sum_{i=1}^n(|x_{i,j}|^2+|x_{i,j}|^4)]}\Big) = o\big(n^{-1/2}\big) \quad \text{and} \\ &p.\exp\Big(-\frac{C_2.n^{1+2(\delta_1-\gamma\delta_2)}.p_0^2}{2(\max_{1\le j\le p}[\sum_{i=1}^n(|x_{i,j}|^2+|x_{i,j}|^4)]+C_3.n^{1/2+(\delta_1-\gamma\delta_2)}.p_0)}\Big) \\ &\quad = o\big(n^{-1/2}\big). \end{aligned}$$

C_1, C_2, C_3 are some positive constants. In view of the assumption $\max\{n^{-1} \times \sum_{i=1}^n |x_{i,j}|^{2r} : 1 \le j \le p\} = O(1)$ [cf. condition (A.1)(ii)], the first bound is implied by $p = o(\exp(C.n^{2(\delta_1-\gamma\delta_2)}.p_0^2).n^{-1/2})$, whereas $p = o(\exp((C.n^{2(\delta_1-\gamma\delta_2)}.p_0^2)/(1+p_0.n^{-1/2+(\delta_1-\gamma\delta_2)})).n^{-1/2})$ is required to obtain the second bound. Here,

C is some positive constant. These requirements on the growth of p are implying the growth conditions stated in Theorem 5.4 and Theorem 5.5.

REMARK 2. Note that the matrices $\breve{\mathbf{\Sigma}}_n$ and $\tilde{\mathbf{\Sigma}}_n$ used in defining the bootstrap pivots do not depend on $G_1^*, \ldots, G_n^*$. Hence it is not required to compute the negative square roots of these matrices for each Monte Carlo bootstrap iteration; these must only be computed once. This is a notable feature of our modified perturbation bootstrap method from the perspective of computational complexity.

REMARK 3. When the dimension p is increasing exponentially, then it is important to choose the distribution of G_i^*'s appropriately depending on whether the regression errors are sub-Gaussian or subexponential. Note that if a random variable W_1 has distribution Beta(a_1, b_1), then by Hoeffding's inequality,

$$\mathbf{E}\big[e^{\kappa(W_1-\mathbf{E}W_1)}\big] \le e^{\kappa^2/8} \qquad \text{for all } \kappa \in \mathcal{R}$$

and hence W_1 is sub-Gaussian with parameter value $1/4$, for any choice of (a_1, b_1). On the other hand, if W_2 has Gamma distribution with shape parameter a_2 and scale parameter b_2 then

$$\begin{aligned}
\log \mathbf{E}\big[e^{\kappa(W_2-\mathbf{E}W_2)}\big] &= -a_2 b_2 \kappa - a_2 \log(1 - b_2\kappa) \qquad \text{for } |\kappa| < 1/b_2 \\
&\le \frac{a_2 b_2^2 \kappa^2}{2(1 - b_2\kappa)} \qquad \text{for } |\kappa| < 1/b_2 \\
&\le a_2 b_2^2 \kappa^2 \qquad \text{for } |\kappa| < 1/2b_2,
\end{aligned}$$

where the first inequality follows from the fact that $-\log(1-u) \le u + \frac{u^2}{2(1-u)}$ for $0 \le u < 1$. Therefore, W_2 is subexponential with parameters $(b_2\sqrt{2a_2}, 2b_2)$, and hence $W_1 + W_2$ is also subexponential with parameters $(\sqrt{1/4 + 2a_2 b_2^2}, 2b_2)$ when W_1 and W_2 are independent. These observations imply that Beta(1/2, 3/2) is an appropriate choice for the distribution of G_i^*'s when the errors are sub-Gaussian and the distribution of $(M_1 + M_2)$ is an appropriate choice for the distribution of G_i^*'s when the errors are sub-exponential where M_1 and M_2 are independent and M_1 is a Gamma random variable with shape and scale parameters 0.008652 and 2, respectively, and M_2 is a Beta random variable with both the parameters 0.036490.

REMARK 4. Let us consider the problem of simultaneous inference. Suppose we want to make inference simultaneously for the parameters $\theta_{j,n}$ for all j in the index set $\mathcal{J}_n$, where $\theta_{j,n} = \boldsymbol{d}'_{j,n}\boldsymbol{\beta}_n$ with $\boldsymbol{d}_{j,n}$ being a $p \times 1$ vector satisfying all the conditions of $\boldsymbol{D}_n$ assumed in Section 3.

First, suppose that $|\mathcal{J}_n|$, the cardinality of $\mathcal{J}_n$, is fixed. Then assuming without loss of generality that $|\mathcal{J}_n| = \{1, \ldots, l\}$ and taking $\boldsymbol{D}_n = (\boldsymbol{d}_{1,n}, \boldsymbol{d}_{2,n}, \ldots, \boldsymbol{d}_{l,n})'$, we can use Theorems 5.1, 5.2, 5.3, 5.4, 5.5 to make simultaneous inference. Obviously, we need to utilize the fact that the perturbation bootstrap approximation holds uniformly over all convex sets of $\mathcal{R}^l$.

Now suppose that $|\mathcal{J}_n|$ is increasing with n. In this scenario, simultaneous inference is not possible with a mere choice of the matrix $\boldsymbol{D}_n$. There are two possible ways out. One way out is to establish the validity of the bootstrap in approximating the distribution of $\max\{\sqrt{n}|\hat{\beta}_{j,n} - \beta_{j,n}| : j \in \mathcal{J}_n\}$. The Edgeworth expansion theory used in this paper is a well-developed technique in fixed-dimensional settings; however, its validity in increasing dimension, more precisely how the error rate depends on the dimension, is still unknown, and hence future investigation is necessary. The second way out is to use componentwise bootstrap approximations dictated by Theorems 5.1, 5.2, 5.3, 5.4, 5.5 and then combine them using the well-known Bonferroni correction procedure. For example, suppose we want to construct a $100(1-\alpha)\%$ confidence region for $(\theta_{1,n}, \dots, \theta_{l_n,n})$ where $\theta_{j,n} = \boldsymbol{d}'_{j,n}\boldsymbol{\beta}_n$ with $\boldsymbol{d}_{j,n}$ being a $p \times 1$ vector satisfying all the conditions of $\boldsymbol{D}_n$ assumed in Section 3, for all $j \in \{1, \dots, l_n\}$. Now, $|\mathcal{J}_n| = l_n$ is increasing with n and α is the family wise error rate (FWER) of the region. FWER of a confidence region Λ_n ($\subseteq \mathcal{R}^{l_n}$) of $(\theta_{1,n}, \dots, \theta_{l_n,n})$ is defined as $\mathbf{P}(\theta_{j,n} \notin \Lambda_n$ for at least one $j)$. Define $\check{R}_{j,n} = \check{\mathbf{R}}_n$ and $\check{R}^*_{j,n} = \check{\mathbf{R}}^*_n$ when $\boldsymbol{D}_n = \boldsymbol{d}'_{j,n}$, $j \in \{1, \dots, l_n\}$. Define $\hat{u}^j_\Omega$ as the $(1-\Omega)$th quantile of the bootstrap distribution of $|\check{R}^*_{j,n}|$ for $j \in \{1, \dots, l_n\}$ for $\Omega \in (0, 1)$. Also, for $I \subseteq \{1, \dots, p\}$, define $\boldsymbol{d}^I_{j,n}$ as the $|I| \times 1$ subvector of $\boldsymbol{d}_{j,n}$ with the entries in the index set I, for all $j \in \{1, \dots, l_n\}$. Similarly, define $\boldsymbol{C}_n^{I,I}$ as the submatrix of $\boldsymbol{C}_n$ with row and column indices in I. Then one can have the following corollary.

COROLLARY 5.1. *Define*, $\hat{\mathcal{A}}_n = \{j : \hat{\beta}_{j,n} \neq 0\}$. *Assume that* $[\boldsymbol{d}^{\hat{\mathcal{A}}_n}_{j,n}]'\boldsymbol{C}_n^{\hat{\mathcal{A}}_n,\hat{\mathcal{A}}_n} \times [\boldsymbol{d}^{\hat{\mathcal{A}}_n}_{j,n}] \geq c$ *for some constant* $c > 0$, *all* $j \in \{1, \dots, l_n\}$ *and that* $l_n = O(n^{1/2})$. *If* $\Omega < \alpha/l_n$, *then* $\{(\theta_{1,n}, \dots, \theta_{l_n,n}) : |\check{R}_{j,n}| \leq \hat{u}^j_\Omega, j = 1, \dots, l_n\}$ *is a confidence region for* $(\theta_{1,n}, \dots, \theta_{l_n,n})$ *with FWER* $\leq \alpha$ *for sufficiently large* n.

Proof of this corollary is presented in Section 8. The confidence region of Corollary 5.1 can be utilized for testing $\theta_{j,n} = 0$ simultaneously for $j \in \{1, \dots, l_n\}$ with FWER $\leq \alpha$. Construction of confidence regions and multiple testing can similarly be carried out with $R_{j,n}$ and $R^*_{j,n}$ instead of $\check{R}_{j,n}$ and $\check{R}^*_{j,n}$ for $j \in \{1, \dots, l_n\}$.

We want to point out that the condition $[\boldsymbol{d}^{\hat{\mathcal{A}}_n}_{j,n}]'\boldsymbol{C}_n^{\hat{\mathcal{A}}_n,\hat{\mathcal{A}}_n}[\boldsymbol{d}^{\hat{\mathcal{A}}_n}_{j,n}] \geq c$ for all $j \in \{1, \dots, l_n\}$ along with the variable selection consistency of Alasso ensure condition (A.2)(iii) of Section 3. (A.2)(iii) is essential to obtain valid Edgeworth expansions in the original and bootstrap settings. In particular, this condition is implying that we can construct a valid confidence region for $(\theta_{1,n}, \dots, \theta_{l_n,n})$ utilizing the perturbation bootstrap only when $\theta_{j,n}$ involves some nonzero component of the regression parameter vector $\boldsymbol{\beta}_n$ for all $j \in \{1, \ldots, l_n\}$. To check if this condition can be dropped or not, future investigation is required. For simultaneous inference regarding $\boldsymbol{\beta}_n$ using the de-biased Lasso, one can also consider the recent work of Dezeure, Bühlmann and Zhang (2017).

6. Simulation results. We study through simulation the coverage of one-sided and two-sided 95% confidence intervals for individual nonzero regression coefficients constructed via the pivot quantities $\boldsymbol{R}_n$ and $\check{\boldsymbol{R}}_n$ as well as via their modified perturbation bootstrap versions $\mathbf{R}_n^*$ and $\check{\boldsymbol{R}}_n^*$. To make further comparisons, we also construct confidence intervals based on a Normal approximation to the distribution of a local quadratic approximation pivot $\boldsymbol{R}_n^{\mathrm{LQA}}$, which uses the estimator of $\mathrm{Cov}((\beta_j, j \in \hat{A}_n)')$ proposed in the original Alasso paper by Zou (2006). We also consider the confidence interval from the oracle Normal approximation, which is based on the closeness in distribution of $\boldsymbol{T}_n$ to a $\mathrm{Normal}(0, \sigma^2 \boldsymbol{D}^{(1)} \boldsymbol{C}_{11,n}^{-1} \boldsymbol{D}^{(1)})$ random variable, where we use the true active set of covariates $\mathcal{A}_n$ to compute $\boldsymbol{C}_{11,n}^{-1}$. We denote this by $\boldsymbol{R}_n^{\mathrm{oracle}}$. For the sake of comparison, we also consider the confidence intervals based on the naive perturbation bootstrap from MTC(11) which in that paper are denoted by CN^{*Q} and CN^{*N}.

Under the settings

$$(n, p, p_0) \in \{(200, 80, 4), (150, 250, 6), (200, 500, 8)\}$$

[more settings are treated in the Supplementary Material (Das, Gregory and Lahiri (2019))], we generate n independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ from the model $Y = X'\boldsymbol{\beta} + \epsilon$, where ϵ is a standard normal random variable, $X = (X_1, \ldots, X_p)'$ is a mean-zero multivariate normal random vector such that

$$\mathrm{Cov}(X_j, X_k) = \mathbf{1}(j = k) + 0.3^{|j-k|} \mathbf{1}(j \le p_0) \mathbf{1}(k \le p_0) \mathbf{1}(j \ne k)$$

for $1 \le j, k \le p$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ with β_j defined as $\beta_j = (1/2) j (-1)^j \times \mathbf{1}(j \le p_0)$ for $j = 1, \ldots, p$.

We compute the empirical coverage over 500 simulated data sets of one- and two-sided confidence intervals for each nonzero regression coefficient under cross validation-selected values of $\tilde{\lambda}_n$ and λ_n, where $\tilde{\lambda}_n$ is the value of the tuning parameter used to obtain the preliminary Lasso estimate $\tilde{\boldsymbol{\beta}}_n$ and λ_n is the value of the tuning parameter used to obtain the Alasso estimate $\hat{\boldsymbol{\beta}}_n$. We use $\gamma = 1$ throughout. For each of the 500 simulated data sets, 1000 Monte Carlo draws of the independent random variables $G_1^*, \ldots, G_n^* \sim \mathrm{Beta}(1/2, 3/2)$ were drawn in order to create 1000 Monte Carlo draws of the bootstrap pivots.

When $p \le n$, we set $\tilde{\lambda}_n = 0$, whereby we use the ordinary least squares estimate for the preliminary estimator $\tilde{\boldsymbol{\beta}}_n$. When $p > n$, the value of $\tilde{\lambda}_n$ is chosen via 10-fold cross validation and $\tilde{\boldsymbol{\beta}}_n$ is computed under the selected value of $\tilde{\lambda}_n$. Once $\tilde{\boldsymbol{\beta}}_n$ is obtained, 10-fold cross validation is used to select λ_n. The values $\tilde{\lambda}_n$ and λ_n are thereafter held fixed for all bootstrap computations on the same dataset. In each cross

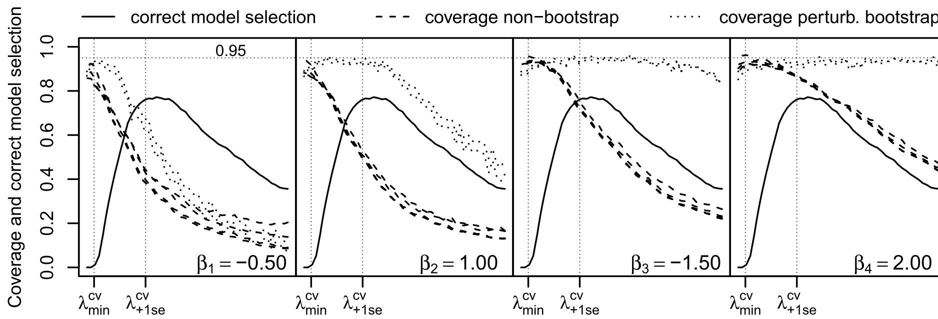

FIG. 1. *Coverage of* β_1, β_2, β_3 *and* β_4 *over* 500 *simulation runs of the confidence intervals based on* $\mathbf{R}_n^{\mathrm{LQA}}$, $\mathbf{R}_n^{\mathrm{oracle}}$, $\mathbf{R}_n$, $\check{\mathbf{R}}_n$ (*dashed curves*), $\mathbf{R}_n^*$, *and* $\check{\mathbf{R}}_n^*$ (*dotted curves*) *along with the frequency of correct model selection* (*solid curve*) *over a grid of fifty* λ_n *values in the* $(n, p, p_0) = (200, 80, 4)$ *case. Vertical lines show median choices of* λ_n *over* 500 *simulation runs when selected by minimizing the cross validation estimate of prediction error* ($\lambda_{\mathrm{min}}^{\mathrm{cv}}$) *or under the* 1-*standard error rule* ($\lambda_{+1\mathrm{se}}^{\mathrm{cv}}$).

validation procedure, the largest value of the tuning parameter for which the cross validation prediction error lies within one standard error of its minimum is used so that greater penalization is preferred; see Friedman, Hastie and Tibshirani (2010).

We begin our discussion of the simulation results with Figure 1, which presents for the case $(n, p, p_0) = (200, 80, 4)$ a study of how the coverages of the confidence intervals based on the various pivots are affected by the choice of λ_n and by the magnitude of the regression coefficients. Each panel of Figure 1 corresponds to one of the $p_0 = 4$ nonzero regression coefficients, where the magnitude of the coefficients increases from left to right. Each panel shows the coverage over 500 simulated data sets of the confidence intervals based on the pivots $\mathbf{R}_n^{\mathrm{LQA}}$, $\mathbf{R}_n^{\mathrm{oracle}}$, $\mathbf{R}_n$, $\check{\mathbf{R}}_n$ (dashed curves), $\mathbf{R}_n^*$ and $\check{\mathbf{R}}_n^*$ (dotted curves) plotted against 50 choices of the tuning parameter λ_n, increasing from left to right. Also appearing in each panel is a solid curve tracing the proportion of times the true model was selected by the Alasso estimator. The two vertical lines in each panel are positioned at the median choices of λ_n when it is selected as the minimizer of the cross validation estimate of the prediction error and when the one-standard-error rule is used. We do not show curves for the CN^{*Q} and CN^{*N} intervals in Figure 1, as they exhibited poorer performance and gave the plots a cluttered appearance.

We see that for small values of λ_n the confidence intervals based on all the pivots achieve close-to-nominal coverage. For such small values of λ_n, however, model selection scarcely occurs. As larger values of λ_n are chosen, the coverage of the confidence intervals tends to drop, the drop being more gradual the larger in magnitude the regression coefficient. The confidence intervals based on the perturbation bootstrap pivots $\mathbf{R}_n^*$ and $\check{\mathbf{R}}_n^*$, however, are able to sustain nominal coverage for much larger values of λ_n than the others, such that they are able to achieve close-to-nominal coverage for the model-selection-optimal choice of λ_n for all but the smallest regression coefficient.

TABLE 1

Empirical coverage of 95% confidence intervals for nonzero regression coefficients by Alasso under $(n, p, p_0) = (200, 80, 4)$ using $\tilde{\lambda}_n = 0$ and cross validation choice of λ_n. The median λ_n choice was $0.987 \cdot n^{1/4}$. One-sided intervals are bounded in the $\mathrm{sgn}(\beta_j)$ direction

β_j	$\mathbf{R}_n^{\mathbf{LQA}}$	$\mathbf{R}_n^{\mathbf{oracle}}$	CN^{*Q}	CN^{*N}	$\mathbf{R}_n$	$\check{\mathbf{R}}_n$	$\mathbf{R}_n^*$	$\check{\mathbf{R}}_n^*$
Coverage and *(avg. width)* of two-sided 95% CIs: $(n, p, p_0) = (200, 80, 4)$								
−0.50	0.42	0.31	0.10	0.45	0.31	0.42	0.61	0.68
	(0.44)	*(0.30)*	*(0.28)*	*(0.31)*	*(0.30)*	*(0.29)*	*(0.26)*	*(0.31)*
1.00	0.54	0.49	0.16	0.77	0.49	0.57	0.95	0.96
	(0.37)	*(0.31)*	*(0.47)*	*(0.48)*	*(0.31)*	*(0.30)*	*(0.39)*	*(0.44)*
−1.50	0.75	0.73	0.36	0.89	0.74	0.76	0.93	0.93
	(0.34)	*(0.31)*	*(0.46)*	*(0.47)*	*(0.32)*	*(0.30)*	*(0.37)*	*(0.41)*
2.00	0.86	0.86	0.59	0.93	0.86	0.87	0.92	0.92
	(0.32)	*(0.30)*	*(0.39)*	*(0.39)*	*(0.31)*	*(0.29)*	*(0.31)*	*(0.34)*
Coverage of one-sided 95% CIs								
−0.50	0.29	0.23	0.08	0.36	0.23	0.36	0.63	0.70
1.00	0.44	0.41	0.12	0.65	0.41	0.50	0.96	0.98
−1.50	0.64	0.61	0.29	0.82	0.62	0.68	0.95	0.96
2.00	0.79	0.78	0.48	0.88	0.78	0.80	0.95	0.96

Table 1 displays the coverage results for the $n > p$ case $(n, p, p_0) = (200, 80, 4)$ under the cross validation choice of λ_n using the one-standard-error rule and Tables 2 and 3 for the $n \le p$ cases $(n, p, p_0) \in \{(150, 250, 6), (200, 500, 8)\}$ under cross validation choices of $\tilde{\lambda}_n$ and λ_n, where both are chosen using the one-standard-error rule. The median values of the cross validation selections of $\tilde{\lambda}_n$ and λ_n under each setting are provided in the table captions in the forms $c_1 \cdot n^{1/2}$ and $c_2 \cdot n^{1/4}$ where c_1 and c_2 are constants. These correspond to the forms of the theoretical choices of $\tilde{\lambda}_n$ and λ_n under the choice of $\gamma = 1$.

In Table 1, we see that under $(n, p, p_0) = (200, 80, 4)$ the modified perturbation bootstrap intervals based on $\mathbf{R}_n^*$ and $\check{\mathbf{R}}_n^*$ achieve the closest-to-nominal coverage. The two-sided $\check{\mathbf{R}}_n^*$ interval achieves subnominal coverage for the smallest regression coefficient $\beta_j = -0.50$, as this coefficient was occasionally estimated to be zero, but achieves close-to-nominal coverage for the larger regression coefficients. The coverage of the other intervals is much more dramatically effected by the magnitude of the regression coefficient β_j, a phenomenon which is even more pronounced in the one-sided coverages; for example, the coverage of the $\check{\mathbf{R}}_n$ interval rises from 0.36 for $\beta_1 = -0.50$ to 0.80 for $\beta_4 = 2.00$. Given that the modified perturbation bootstrap distributions of $\mathbf{R}_n^*$ and $\check{\mathbf{R}}_n^*$ result in much closer-to-nominal coverages than the Normal approximations to the distributions of $\mathbf{R}_n$ and $\check{\mathbf{R}}_n$, we may conclude that the sample size is too small for the asymptotically-Normal pivots to have sufficiently approached their limiting distribution; the second-order correctness of the modified perturbation bootstrap is thus apparent.

TABLE 2

Empirical coverage of 95% *confidence intervals for nonzero regression coefficients by Alasso under* $(n, p, p_0) = (150, 250, 6)$ *using cross validation choices of* $\tilde{\lambda}_n$ *and* λ_n. *The median* $\tilde{\lambda}_n$ *and* λ_n *choices were* $0.014 \cdot n^{1/2}$ *and* $0.119 \cdot n^{1/4}$. *One-sided intervals are bounded in the* $\operatorname{sgn}(\beta_j)$ *direction*

β_j	$\mathbf{R}_n^{\mathbf{LQA}}$	$\mathbf{R}_n^{\mathbf{oracle}}$	CN^{*Q}	CN^{*N}	$\mathbf{R}_n$	$\check{\mathbf{R}}_n$	$\mathbf{R}_n^*$	$\check{\mathbf{R}}_n^*$
	Coverage and *(avg. width)* of two-sided 95% CIs: $(n, p, p_0) = (150, 250, 6)$							
−0.50	0.76	0.67	0.52	0.78	0.67	0.80	0.81	0.83
	(0.71)	*(0.31)*	*(0.41)*	*(0.45)*	*(0.32)*	*(0.38)*	*(0.39)*	*(0.50)*
1.00	0.82	0.75	0.62	0.87	0.76	0.86	0.90	0.93
	(0.52)	*(0.32)*	*(0.42)*	*(0.43)*	*(0.33)*	*(0.40)*	*(0.42)*	*(0.56)*
−1.50	0.89	0.86	0.78	0.90	0.87	0.92	0.91	0.95
	(0.53)	*(0.32)*	*(0.40)*	*(0.40)*	*(0.33)*	*(0.40)*	*(0.40)*	*(0.53)*
2.00	0.88	0.84	0.82	0.87	0.85	0.91	0.87	0.94
	(0.46)	*(0.33)*	*(0.38)*	*(0.39)*	*(0.33)*	*(0.40)*	*(0.38)*	*(0.50)*
−2.50	0.90	0.86	0.85	0.88	0.87	0.91	0.88	0.94
	(0.42)	*(0.32)*	*(0.37)*	*(0.37)*	*(0.33)*	*(0.40)*	*(0.36)*	*(0.48)*
3.00	0.89	0.85	0.85	0.87	0.86	0.92	0.87	0.93
	(0.45)	*(0.31)*	*(0.34)*	*(0.34)*	*(0.32)*	*(0.38)*	*(0.33)*	*(0.43)*
	Coverage of one−sided 95% CIs							
−0.50	0.71	0.63	0.45	0.75	0.64	0.73	0.84	0.88
1.00	0.76	0.69	0.53	0.81	0.69	0.81	0.91	0.95
−1.50	0.84	0.81	0.72	0.86	0.82	0.88	0.90	0.94
2.00	0.85	0.82	0.76	0.84	0.83	0.86	0.89	0.92
−2.50	0.87	0.83	0.82	0.86	0.84	0.88	0.89	0.92
3.00	0.86	0.83	0.82	0.84	0.84	0.88	0.87	0.92

In the $p > n$ settings, the modified perturbation bootstrap interval based on $\check{\mathbf{R}}_n^*$ continues to perform well. Under the $(n, p, p_0) = (150, 250, 6)$ setting, for which Table 2 shows the results, the $\check{\mathbf{R}}_n^*$ interval achieves the nominal coverage across all regression coefficients except for the smallest in magnitude for both two- and one-sided intervals. Here, also we see a difference between the performance of the confidence intervals based on $\mathbf{R}_n^*$ and $\check{\mathbf{R}}_n^*$, owing to the bias correction; the coverage of the $\mathbf{R}_n^*$ interval tends to be subnominal for both one- and two-sided intervals. The confidence intervals based on the asymptotic normality of the respective pivot all have subnominal coverage for most of the regression coefficients, and their coverages are dramatically affected by the magnitude of the true regression coefficient.

The results are similar for the $(n, p, p_0) = (200, 500, 8)$ case, for which Table 3 shows the results. The only confidence interval which reliably achieves close-to-nominal coverage is the modified perturbation bootstrap interval based on $\check{\mathbf{R}}_n^*$. We note that the width of the $\check{\mathbf{R}}_n^*$ interval seems to adapt more to the magnitude of the regression coefficient than the widths of the Normal-based confidence inter-

TABLE 3

Empirical coverage of 95% *confidence intervals for nonzero regression coefficients by Alasso under* $(n, p, p_0) = (200, 500, 8)$ *using cross validation choices of* $\tilde{\lambda}_n$ *and* λ_n. *The median* $\tilde{\lambda}_n$ *and* λ_n *choices were* $0.01 \cdot n^{1/2}$ *and* $0.30 \cdot n^{1/4}$. *One-sided intervals are bounded in the* $\mathrm{sgn}(\beta_j)$ *direction*

β_j	$\mathbf{R}_n^{\mathrm{LQA}}$	$\mathbf{R}_n^{\mathrm{oracle}}$	CN^{*Q}	CN^{*N}	$\mathbf{R}_n$	$\check{\mathbf{R}}_n$	$\mathbf{R}_n^*$	$\check{\mathbf{R}}_n^*$
			Coverage and *(avg. width)* of two-sided 95% CIs: $(n, p, p_0) = (200, 500, 8)$					
−0.50	0.79	0.68	0.56	0.86	0.70	0.81	0.87	0.92
	(0.69)	*(0.26)*	*(0.38)*	*(0.42)*	*(0.27)*	*(0.33)*	*(0.36)*	*(0.46)*
1.00	0.84	0.75	0.65	0.86	0.77	0.87	0.88	0.94
	(0.54)	*(0.27)*	*(0.34)*	*(0.35)*	*(0.28)*	*(0.35)*	*(0.34)*	*(0.44)*
−1.50	0.90	0.85	0.83	0.88	0.85	0.91	0.86	0.94
	(0.45)	*(0.27)*	*(0.31)*	*(0.31)*	*(0.28)*	*(0.35)*	*(0.31)*	*(0.41)*
2.00	0.89	0.84	0.86	0.85	0.85	0.92	0.86	0.95
	(0.44)	*(0.28)*	*(0.30)*	*(0.30)*	*(0.28)*	*(0.35)*	*(0.30)*	*(0.40)*
−2.50	0.93	0.89	0.87	0.88	0.89	0.91	0.89	0.93
	(0.46)	*(0.28)*	*(0.30)*	*(0.30)*	*(0.28)*	*(0.35)*	*(0.30)*	*(0.39)*
3.00	0.91	0.85	0.87	0.86	0.86	0.91	0.86	0.92
	(0.46)	*(0.27)*	*(0.30)*	*(0.30)*	*(0.28)*	*(0.35)*	*(0.29)*	*(0.39)*
−3.50	0.91	0.86	0.87	0.87	0.87	0.92	0.87	0.95
	(0.48)	*(0.27)*	*(0.30)*	*(0.30)*	*(0.28)*	*(0.35)*	*(0.29)*	*(0.39)*
4.00	0.89	0.86	0.87	0.87	0.86	0.90	0.84	0.92
	(0.45)	*(0.26)*	*(0.28)*	*(0.28)*	*(0.27)*	*(0.33)*	*(0.28)*	*(0.36)*
			Coverage of one−sided 95% CIs					
−0.50	0.72	0.62	0.48	0.82	0.63	0.75	0.89	0.94
1.00	0.79	0.70	0.59	0.80	0.71	0.81	0.87	0.94
−1.50	0.87	0.79	0.79	0.82	0.80	0.87	0.87	0.92
2.00	0.85	0.80	0.81	0.82	0.82	0.86	0.85	0.91
−2.50	0.89	0.84	0.84	0.86	0.85	0.88	0.86	0.91
3.00	0.86	0.79	0.83	0.82	0.81	0.86	0.84	0.90
−3.50	0.88	0.82	0.85	0.85	0.83	0.88	0.85	0.91
4.00	0.89	0.83	0.84	0.84	0.85	0.87	0.84	0.90

vals, which remain, with the exception of the $\mathbf{R}_n^{\mathrm{LQA}}$ interval, fairly constant across all magnitudes of β_j, resulting in poorer coverage for smaller regression coefficients. In contrast, the $\check{\mathbf{R}}_n^*$ interval is able to achieve nominal coverage even for the smallest values of β_j by producing suitably wider confidence intervals.

We see that the modified perturbation bootstrap is able to produce reliable confidence intervals for regression coefficients in the high-dimensional setting under data-based choices of the tuning parameter, and importantly, under levels of penalization large enough for model selection to occur.

7. Data analysis. To illustrate the construction of confidence intervals for regression coefficients in the high-dimensional linear regression model using the modified perturbation bootstrap, we present an analysis of the `riboflavin` data

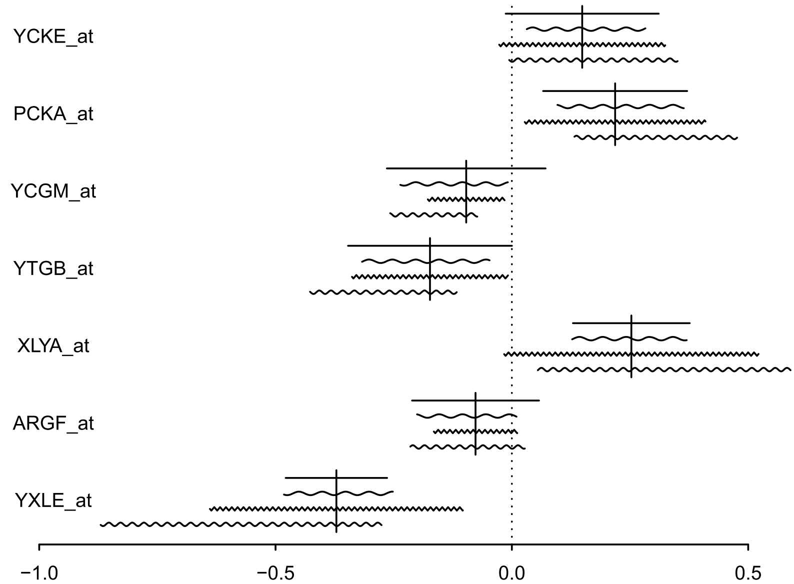

FIG. 2. *Confidence intervals based on* $\mathbf{R}_n^{\mathrm{LQA}}$ (*straight*), $\check{\mathbf{R}}_n$ (*wavy*), CN^{*N} (*jagged*) *and* $\check{\mathbf{R}}_n^*$ (*wiggly*) *for each of the Alasso selected genes from the* `riboflavin` *data set.*

set considered in Bühlmann, Kalisch and Meier (2014), which those authors make publicly available in their Supplementary Material. The data contains $n = 71$ independent records consisting of a response variable which is the logarithm of the riboflavin production rate and of 4088 gene expression levels in batches of *Bacillis subtilis* bacteria. Of the 4088, we preselect 200 genes by sorting them in order of decreasing empirical variance and keeping the first 200. We then fit the linear regression model to the data set with $n = 71$ and $p = 200$ and compute confidence intervals for the regression coefficients selected by the Alasso procedure. The variables selected by our methods were different from those discovered in Bühlmann, Kalisch and Meier (2014). We choose $\tilde{\lambda}_n$ and λ_n using 10-fold crossvalidation. Figure 2 displays the confidence intervals for the Alasso-selected covariates obtained from the $\mathbf{R}_n^{\mathrm{LQA}}$, $\check{\mathbf{R}}_n$, CN^{*N}, and $\check{\mathbf{R}}_n^*$ pivots, where 1000 bootstrap replicates were used for the bootstrap-based intervals.

The interval based on the $\mathbf{R}_n^{\mathrm{LQA}}$ pivot (straight line) and the CN^{*N} interval (jagged), are symmetric around the estimated value of the regression coefficient (the CN^{*N} interval is formed by adding and subtracting an upper quantile of a Normal distribution with a bootstrap-estimated variance). The intervals based on $\check{\mathbf{R}}_n$ are asymmetric owing to the bias correction (which is quite small in this example) and, in the case of the $\check{\mathbf{R}}_n^*$ interval, owing to the bias correction and to the asymmetry of the bootstrap distribution of $\check{\mathbf{R}}_n^*$. For some of the coefficients, the $\check{\mathbf{R}}_n^*$

interval is highly asymmetric, suggesting that the distribution of the pivot $\check{\mathbf{R}}_n$ may still be far from Normal.

8. Proofs. Only an outline of the proofs are presented in this section. Details are in the Supplementary Material [Das, Gregory and Lahiri (2019)]. First, we define some additional notation and recall some notation that are defined earlier. Define $\check{\boldsymbol{W}}_n^* = n^{-1/2}\sum_{i=1}^n \hat{\epsilon}_i \boldsymbol{x}_i(G_i^* - \mu_{G^*})$. Write $\check{\boldsymbol{W}}_n^{*(0)} = \check{\boldsymbol{W}}_n^*$, $p^{(0)} = p$, $p^{(1)} = p_0$, and $p^{(2)} = p - p_0$. Define $\tilde{b}_n = \sigma^{-1}\boldsymbol{\Sigma}_n^{-1/2}\boldsymbol{b}_n$ when $p \le n$ and $\tilde{b}_n = \sigma^{-1}\bar{\boldsymbol{\Sigma}}_n^{-1/2}\boldsymbol{b}_n$ when $p > n$. Recall that $\boldsymbol{b}_n = \boldsymbol{D}_n^{(1)}\boldsymbol{C}_{11,n}^{-1}\boldsymbol{s}_n^{(1)}\frac{\lambda_n}{2\sqrt{n}}$, where $\boldsymbol{D}_n^{(1)}$ and $\boldsymbol{C}_{11,n}$ are as defined earlier and $\boldsymbol{s}_n^{(1)}$ is a $p_0 \times 1$ vector with jth element $\operatorname{sgn}(\beta_{j,n})|\beta_{j,n}|^{-\gamma}$. Note that under conditions (A.2)(i), (A.3) and (A.6)(i), $\|\boldsymbol{\Sigma}_n\| = O(1)$, $\|\bar{\boldsymbol{\Sigma}}_n\| = O(1)$, $\|\boldsymbol{D}_n^{(1)}\boldsymbol{C}_{11,n}^{-1/2}\| = O(1)$ and $\|\boldsymbol{s}_n^{(1)}\| \le K\sqrt{p_0}.n^{b\gamma}$. Hence, $\|\tilde{\boldsymbol{b}}_n\| = O(n^{-\delta_1})$. Also define $\check{\boldsymbol{b}}_n = \check{\boldsymbol{\Sigma}}_n^{-1/2}\boldsymbol{D}_n^{(1)}\boldsymbol{C}_{11,n}^{-1}\check{\boldsymbol{s}}_n^{(1)}\frac{\lambda_n}{2\sqrt{n}}$ when $p \le n$ and $\check{\boldsymbol{b}}_n = \tilde{\boldsymbol{\Sigma}}_n^{-1/2}\boldsymbol{D}_n^{(1)}\boldsymbol{C}_{11,n}^{-1}\check{\boldsymbol{s}}_n^{(1)}\frac{\lambda_n}{2\sqrt{n}}$ when $p > n$, where $\check{\boldsymbol{s}}_n^{(1)} = (\check{s}_{1n}, \dots, \check{s}_{p_0 n})'$ and $\check{s}_{j,n} = \operatorname{sgn}(\hat{\beta}_{j,n})|\hat{\beta}_{j,n}|^{-\gamma}$.

We denote by $\|\cdot\|$ and $\|\cdot\|_\infty$, respectively, the L^2 and L^∞ norm. For a nonnegative integer-valued vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_l)'$ and a function $f = (f_1, f_2, \dots, f_l) : \mathscr{R}^l \to \mathscr{R}^l$, $l \ge 1$, write $|\boldsymbol{\alpha}| = \alpha_1 + \dots + \alpha_l$, $\boldsymbol{\alpha}! = \alpha_1! \cdots \alpha_l!$, $f^{\boldsymbol{\alpha}} = (f_1^{\alpha_1}) \cdots (f_l^{\alpha_l})$ and $D^{\boldsymbol{\alpha}} f_1 = D_1^{\alpha_1} \cdots D_l^{\alpha_l} f_1$, where $D_j f_1$ denotes the partial derivative of f_1 with respect to the jth component of the argument, $1 \le j \le l$. For $\boldsymbol{t} = (t_1, \dots, t_l)' \in \mathscr{R}^l$ and $\boldsymbol{\alpha}$ as above, define $t^{\boldsymbol{\alpha}} = t_1^{\alpha_1} \cdots t_l^{\alpha_l}$. Let $\boldsymbol{\Phi}_V$ denote the multivariate Normal distribution with mean $\mathbf{0}$ and dispersion matrix $\boldsymbol{V}$ having jth row $\boldsymbol{V}_{j.}$ and let $\boldsymbol{\phi}_V$ denote the density of $\boldsymbol{\Phi}_V$. We write $\boldsymbol{\Phi}_V = \boldsymbol{\Phi}$ and $\boldsymbol{\phi}_V = \boldsymbol{\phi}$ when $\boldsymbol{V}$ is the identity matrix.

Define $\boldsymbol{A}_{1n} = \{\{\|\check{\boldsymbol{W}}_n^{(1)}\|_\infty \le K\sqrt{\log n}\} \cap \{\|\check{\boldsymbol{W}}_n^{(2)}\|_\infty \le K\sqrt{\log n}\} \cap \{\|\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)\|_\infty \le K\sqrt{\log n}\}\}$ for $p \le n$ and $\boldsymbol{A}_{1n} = \{\{\|\check{\boldsymbol{W}}_n^{(1)}\|_\infty \le K\sqrt{\log n}\} \cap \{\|\check{\boldsymbol{W}}_n^{(2)}\|_\infty \le K\sqrt{\log n}\} \cap \{\|\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)\|_\infty \le C.n^{\delta_2}\}\}$ for $p > n$. We have assumed $\mathcal{A}_n = \{1, \dots, p_0\}$. $\check{\boldsymbol{W}}_n^{(1)}$ and $\check{\boldsymbol{W}}_n^{(2)}$ are respectively first p_0 and last $(p - p_0)$ components of $\check{\boldsymbol{W}}_n$. Note that $\mathbf{P}(\boldsymbol{A}_{1n}) \ge 1 - O(p.n^{-(r-2)/2})$ for $p \le n$ and $\mathbf{P}(\boldsymbol{A}_{1n}) \ge 1 - o(n^{-1/2})$ for $p > n$ [cf. Lemma 8.1 of Chatterjee and Lahiri (2013)].

Note that, $\check{\boldsymbol{b}}_n = O_p(n^{-\delta_1})$, by Lemma 8.4 and 8.5, described below. Suppose, $r_1 = \min\{a \in \mathcal{N} : \|\check{\boldsymbol{b}}_n\|^{a+1} = o_p(n^{-1/2})\}$, $\mathcal{N}$ being the set of natural numbers. Define the conditional Lebesgue density of two-term Edgeworth expansion of $\boldsymbol{R}_n^*$ as

$$\begin{aligned}\xi_n^*(\boldsymbol{x}) = {}&\phi(\boldsymbol{x})\Bigg[1 + \sum_{k=1}^{r_1}\frac{1}{k!}\Bigg\{\sum_{|\boldsymbol{\alpha}|=k}\check{\boldsymbol{b}}_n^{\alpha} H_{\boldsymbol{\alpha}}(\boldsymbol{x})\Bigg\} + \frac{1}{\sqrt{n}}\Bigg[\frac{1}{6}\sum_{|\boldsymbol{\alpha}|=3}\boldsymbol{t}^{\boldsymbol{\alpha}}\bar{\boldsymbol{\xi}}_n^{*(1)}(\boldsymbol{\alpha})H_{\boldsymbol{\alpha}}(\boldsymbol{x})\\ &- \frac{1}{2\hat{\sigma}_n^2}\Bigg\{\sum_{|\boldsymbol{\alpha}|=1}\boldsymbol{t}^{\boldsymbol{\alpha}}\bar{\boldsymbol{\xi}}_n^{*(3)}(\boldsymbol{\alpha})H_{\boldsymbol{\alpha}}(\boldsymbol{x})\end{aligned}$$

$$+\sum_{|\boldsymbol{\alpha}|=1}\sum_{|\boldsymbol{\zeta}|=2} \boldsymbol{t}^{\boldsymbol{\alpha}+\boldsymbol{\zeta}} \bar{\boldsymbol{\xi}}_n^{*(3)}(\boldsymbol{\alpha})\bar{\boldsymbol{\xi}}_n^{*(1)}(\boldsymbol{\zeta})H_{\boldsymbol{\alpha}+\boldsymbol{\zeta}}(\boldsymbol{x})\Big\}\Big]\Big],$$

where $x \in \mathcal{R}^q$, $\bar{\boldsymbol{\xi}}_n^{*(j)}(\boldsymbol{\alpha}) = n^{-1}\sum_{i=1}^n (\breve{\boldsymbol{\xi}}_i^{(0)} \hat{\epsilon}_i^j)^{\boldsymbol{\alpha}}$, $j = 0, 1, \dots$ and $H_{\boldsymbol{\alpha}}(\boldsymbol{x}) = (-D)^{\boldsymbol{\alpha}}\phi(\boldsymbol{x})$, where $\phi(\cdot)$ is the standard normal density on $\mathcal{R}^q$.

LEMMA 8.1. *Suppose $Y_1, \dots, Y_n$ are zero mean independent r.v.s and $\mathbf{E}(|Y_i|^t) < \infty$ for $i = 1, \dots, n$ and $\sum_{i=1}^n \mathbf{E}(|Y_i|^t) = \sigma_t$; $S_n = \sum_{i=1}^n Y_i$. Then, for any $t \geq 2$ and $x > 0$*

$$P\big[|S_n| > x\big] \leq C\big[\sigma_t x^{-t} + \exp(-x^2/\sigma_2)\big].$$

Proof of Lemma 8.1. This inequality was proved in Fuk and Nagaev (1971).

LEMMA 8.2. *Under assumptions* (A.1), (A.3), (A.4)(i) *and* (A.5)(i), (ii) *with* $r = 3$:

(i) $\mathbf{P}_*(\|\breve{\boldsymbol{W}}_n^{*(1)}\| > K\sqrt{p_0 \log n}) = O_p(p_0.n^{-(r-2)/2})$.
(ii) $\mathbf{P}_*(\|\breve{\boldsymbol{W}}_n^{*(l)}\|_\infty > K\sqrt{\log n}) = O_p(p^{(l)}.n^{-(r-2)/2})$, *for* $l = 0, 1, 2$.
(iii) $\mathbf{P}_*(\|\sqrt{n}(\tilde{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n)\|_\infty > K\sqrt{\log n}) = O_p(p.n^{-(r-2)/2})$, *when* $p \leq n$.

PROOF. This lemma follows through the same line of Lemma 8.1 of Chatterjee and Lahiri (2013) and employing Lemma 8.1, stated above. □

LEMMA 8.3. *Suppose p is fixed. Then under condition* (A.1)(ii) *and* (A.4)(i) *with r=2,*

$$\mathbf{P}_*\big(\|\tilde{\boldsymbol{\beta}}_n^{*N} - \tilde{\boldsymbol{\beta}}_n\| = o\big(n^{-1/2}(\log n)^{1/2}\big)\big) \geq 1 - o_p\big(n^{-1/2}\big).$$

PROOF. This lemma is proved in Proposition 4.1 of Das and Lahiri (2019). □

LEMMA 8.4. *Suppose assumptions* (A.1)–(A.3), (A.4)(i), (A.5)(i), (ii) *and* (A.6) *hold with $r = 4$. Then*

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\|_\infty = O_p\big(n^{-1/2}\big) \quad \textit{and}$$

$$\textit{on the set } \boldsymbol{A}_{1n}, \qquad \|\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n\|_\infty = O_{p_*}\big(n^{-1/2}\big).$$

PROOF. See the Supplementary Material [Das, Gregory and Lahiri (2019)]. □

LEMMA 8.5. *Under the assumptions* (A.1)–(A.3), (A.4)(i) *and* (A.6)(i) *and* (iii) *with* $r = 6$, *we have*

$$\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_n\| = o_p(n^{-(1+\delta_1)/2}), \qquad \|\check{\mathbf{\Sigma}}_n - \bar{\mathbf{\Sigma}}_n\| = o_p(n^{-1})$$

$$\|\check{\mathbf{\Sigma}}_n - \sigma^2\mathbf{\Sigma}_n\|, \|\tilde{\mathbf{\Sigma}}_n - \sigma^2\bar{\mathbf{\Sigma}}_n\| = O_p(n^{-1/2}),$$

where δ_1 *is as defined in assumption* (A.6).

PROOF. See the Supplementary Material [Das, Gregory and Lahiri (2019)]. □

LEMMA 8.6. *Let* $p \le n$ *and suppose that* (A.1)–(A.6) *hold with* $r = 6$. *Then on a set* A_{2n} *with* $\mathbf{P}(\boldsymbol{\varepsilon} \in \boldsymbol{A}_{2n}) \to 1$, *when* $\boldsymbol{\varepsilon} \in \boldsymbol{A}_{2n}$, *we have*:

(a) *if* $p \le n$, *then*

$$\sup_{B \in \mathcal{C}_q} \left| \mathbf{P}_*(\boldsymbol{R}_n^* \in B) - \int_B \xi_n^*(\boldsymbol{x})\, d\boldsymbol{x} \right| = o(n^{-1/2});$$

(b) *if* $p > n$, $b = 0$ *and additionally conditions* (A.7) *and* (A.1)(iii)$'$ [*in place of* (A.1)(iii)] *hold*, *then*

$$\sup_{B \in \mathcal{C}_q} \left| \mathbf{P}_*(\boldsymbol{R}_n^* \in B) - \int_B \xi_n^*(\boldsymbol{x})\, d\boldsymbol{x} \right| = o(n^{-1/2}).$$

PROOF. See the Supplementary Material [Das, Gregory and Lahiri (2019)]. □

PROOF OF PROPOSITION 2.1. Note that for any $\boldsymbol{t} \in \mathcal{R}^p$ and for each $i \in \{1, \dots, n\}$,

$$\sum_{i=1}^n (y_i - \boldsymbol{x}_i'\boldsymbol{t}^*)^2 (G_i^* - \mu_{G^*}) = \sum_{i=1}^n [\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{t})]^2 (G_i^* - \mu_{G^*}) - 2(\boldsymbol{t} - \hat{\boldsymbol{\beta}}_n)'\check{\boldsymbol{W}}_n^* + \sum_{i=1}^n \hat{\epsilon}_i^2 (G_i^* - \mu_{G^*})^2$$

and

$$\sum_{i=1}^n (z_i - \boldsymbol{x}_i'\boldsymbol{t})^2 = \sum_{i=1}^n [\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{t})]^2 - 2\mu_{G^*}^{-1}(\boldsymbol{t} - \hat{\boldsymbol{\beta}}_n)'\check{\boldsymbol{W}}_n^* + \mu_{G^*}^{-2} \sum_{i=1}^n [\hat{\epsilon}_i (G_i^* - \mu_{G^*})]^2.$$

Therefore, Proposition 2.1 follows. For details, see the Supplementary Material [Das, Gregory and Lahiri (2019)]. □

PROOF OF THEOREM 4.1. The KKT condition (4.3) corresponding to the Alasso criterion function, defined in MTC(11), can be rewritten through the vector

$\boldsymbol{w}^* = (\boldsymbol{w}_n^{*(1)\prime}, \boldsymbol{w}_n^{*(2)\prime})'$ as

$$2\boldsymbol{C}^*_{11,n}\boldsymbol{w}_n^{*(1)} + 2\boldsymbol{C}^*_{12,n}\boldsymbol{w}_n^{*(2)} - 2\boldsymbol{W}_n^{*(1)} + \frac{\lambda_n^*}{\sqrt{n}}\boldsymbol{\Gamma}_n^{*(1)}\boldsymbol{l}_n^{(1)} = \boldsymbol{0} \tag{8.1}$$

and for each $j \in \{p_0 + 1, \dots, p\}$

$$\begin{aligned} -\frac{\lambda_n^*}{2\sqrt{n}}|\tilde{\beta}_{j,n}^{*N}|^{-\gamma} &\le \left[(\boldsymbol{C}^*_{21,n})_{j\cdot}\boldsymbol{w}_n^{*(1)} + (\boldsymbol{C}^*_{22,n})_{j\cdot}\boldsymbol{w}_n^{*(2)} - W^*_{j,n}\right] \\ &\le \frac{\lambda_n^*}{2\sqrt{n}}|\tilde{\beta}_{j,n}^{*N}|^{-\gamma}. \end{aligned} \tag{8.2}$$

Here, $\boldsymbol{W}_n^* = n^{-1/2}\sum_{i=1}^n \hat{\epsilon}_i \boldsymbol{x}_i G_i^*$, $\boldsymbol{W}_n^{*(1)}$ is the vector of the first p_0 components of $\boldsymbol{W}_n^*$, $W^*_{j,n}$ is the jth component of $\boldsymbol{W}_n^*$ for $j \in \{1, \dots, p\}$, $\boldsymbol{l}_n^{(1)} = (l_{1n}, \dots, l_{p_0 n})'$ with $l_{k,n} \in [-1, 1]$ for $k = 1, \dots, p_0$ and $\boldsymbol{\Gamma}_n^{*(1)} = \operatorname{diag}(|\tilde{\beta}_{1n}^{*N}|^{-\gamma}, \dots, |\tilde{\beta}_{p_0 n}^{*N}|^{-\gamma})$ and

$$\boldsymbol{C}_n^* = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i' G_i^* = \begin{bmatrix} \boldsymbol{C}^*_{11,n} & \boldsymbol{C}^*_{12,n} \\ \boldsymbol{C}^*_{21,n} & \boldsymbol{C}^*_{22,n} \end{bmatrix},$$

where $\boldsymbol{C}^*_{11,n}$ is of dimension $p_0 \times p_0$. $(\boldsymbol{C}^*_{21,n})_{j\cdot}$ is the jth row of $\boldsymbol{C}^*_{21,n}$, $j \in \{p_0 + 1, \dots, p\}$.

Now, to prove part (a) of Theorem 4.1, it is enough to show that $(\boldsymbol{u}_{n2}^{*N\prime}, \boldsymbol{0}')'$ satisfies (8.1) and (8.2) separately with bootstrap probability $1 - o_p(n^{-1/2})$. The vector $\boldsymbol{u}_{n2}^{*N}$ is defined as $\boldsymbol{u}_{n2}^{*N} = \boldsymbol{C}^{*-1}_{11,n}[\boldsymbol{W}_n^{*(1)} - \frac{\lambda_n^*}{\sqrt{n}}\tilde{\boldsymbol{s}}_n^{*N(1)}]$, where the jth component of $\tilde{\boldsymbol{s}}_n^{*N(1)}$ is equal to $\operatorname{sgn}(\hat{\beta}_{j,n})|\tilde{\beta}_{jn}^{*N}|^{-\gamma}$, $j \in \{1, \dots, p_0\}$.

Note that $(\boldsymbol{u}_{n2}^{*N\prime}, \boldsymbol{0}')'$ exactly satisfies (8.1) if $\boldsymbol{l}_n^{(1)} = (\operatorname{sgn}(\hat{\beta}_{1,n}), \dots, \hat{\beta}_{p_0,n}))$. Thus we can conclude that $(\boldsymbol{u}_{n2}^{*N\prime}, \boldsymbol{0}')'$ satisfies (8.1) with bootstrap probability $1 - o_p(n^{-1/2})$, if we can show that $\|\boldsymbol{C}^{*-1}_{11,n}[\boldsymbol{W}_n^{*(1)} - \frac{\lambda_n^*}{\sqrt{n}}\boldsymbol{\Gamma}_n^{*(1)}\boldsymbol{l}_n^{(1)}]\| = o(n^{1/2})$ with bootstrap probability $1 - o_p(n^{-1/2})$. This follows from the facts that

$$\mathbf{P}_*\big(\|\boldsymbol{C}^*_{11,n} - \boldsymbol{C}_{11,n}\mu_{G^*}\| > K.p_0.n^{-1/2}.(\log n)^{1/2}\big) = o_p\big(n^{-1/2}\big), \tag{8.3}$$

on the set $\boldsymbol{A}_{1n}$, $|\tilde{\beta}_{j,n}|^{-\gamma}$ is bounded for all $j \in \{1, \dots, p_0\}$ and $n^{-1/2}\lambda_n \to 0$. Now, note that

$$\begin{aligned} &\mathbf{P}_*\Big(\max_j\big\{\|(\boldsymbol{C}^*_{21,n})_{j\cdot} - (\boldsymbol{C}_{21,n})_{j\cdot}\mu_{G^*}\| : j \in \{p_0 + 1, \dots, p\}\big\} \\ &\quad > K.p_0^{1/2}.n^{-1/2}.(\log n)^{1/2}\Big) \\ &\quad \le \sum_{k=1}^{p_0}\sum_{j=p_0+1}^{p}\mathbf{P}_*\left(\left|\sum_{i=1}^n x_{ij}x_{ik}\big(G_i^* - \mu_{G^*}\big)\right| > K.n^{-1/2}.(\log n)^{1/2}\right) \\ &\quad = o_p\big(n^{-1/2}\big). \end{aligned}$$

This fact, Lemma 8.3 and $\mathbf{P}_*(|W^*_{jn}| > K.(\log n)^{1/2}) = 1 - o_p(n^{-1/2})$ imply that

$$\mathbf{P}_*\big((\boldsymbol{u}^{*N\prime}_{n2}, \mathbf{0}')' \text{ satisfies } (8.2)\big) = 1 - o_p\big(n^{-1/2}\big),$$

due to $\max\{\lambda_n, \lambda^*_n\}.(\log n/n)^{1/2} \to 0$. Therefore, part (a) of Theorem 4.1 follows.

Now for part (b), note that due to (8.3) and the fact that $n^{-1/2}.(\log n)^{1/2}.\lambda_n \to 0$, it follows that on the set $\boldsymbol{A}_{1n}$,

$$\mathbf{P}_*\left(\sqrt{n}\left\|(\boldsymbol{C}^{*-1}_{11,n} - \boldsymbol{C}^{-1}_{11,n}\mu^{-1}_{G^*})n^{-1/2}\sum_{i=1}^n \boldsymbol{x}^{(1)}_i \hat{\epsilon}_i\right\|_\infty = o(1)\right) \tag{8.4}$$
$$= 1 - o\big(n^{-1/2}\big).$$

Since $\boldsymbol{C}_n \to \boldsymbol{C}$ for some $p \times p$ positive definite matrix $\boldsymbol{C}$ and $\mathbf{P}(\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| = O(n^{-1/2}(\log n)^{1/2})) \geq 1 - o(n^{-1/2})$, so part (b) follows due to (8.4).

To prove part (c), it is enough to show

$$\sup_{\boldsymbol{x}\in\mathcal{R}^{p_0}} \big|\mathbf{P}_*\big(\boldsymbol{F}^{*(1)}_n \leq \boldsymbol{x}\big) - \mathbf{P}\big(\boldsymbol{F}^{(1)}_n \leq \boldsymbol{x}\big)\big| \geq K.\frac{\lambda_n}{\sqrt{n}} \qquad \text{for some } K > 0, \tag{8.5}$$

where $\boldsymbol{F}^{*(1)}_n$ and $\boldsymbol{F}^{(1)}_n$ are subvectors of $\boldsymbol{F}^*_n$ and $\boldsymbol{F}_n$, respectively, comprising of first p_0 components. Note that

$$\begin{aligned}
\boldsymbol{F}^{*(1)}_n &= \boldsymbol{C}^{*-1}_{11,n}\left[\boldsymbol{W}^{*(1)}_n - \frac{\lambda^*_n}{\sqrt{n}}\tilde{\boldsymbol{s}}^{*N(1)}_n\right] \\
&= \boldsymbol{C}^{-1}_{11,n}\mu^{-1}_{G^*}\left[\boldsymbol{W}^{*(1)}_n - \frac{\lambda^*_n}{\sqrt{n}}\tilde{\boldsymbol{s}}^{*N(1)}_n\right] \\
&\quad + \big(\boldsymbol{C}^{*-1}_{11,n} - \boldsymbol{C}^{-1}_{11,n}\mu^{-1}_{G^*}\big)\left[\boldsymbol{W}^{*(1)}_n - \frac{\lambda^*_n}{\sqrt{n}}\tilde{\boldsymbol{s}}^{*N(1)}_n\right] \\
&= \breve{F}^{*(1)}_n + \breve{\boldsymbol{R}}^*_{1n} \qquad \text{(say)} \\
&= \tilde{\boldsymbol{F}}^{*(1)}_n + \tilde{\boldsymbol{Ad}}^{(1)}_n + \breve{\boldsymbol{R}}^*_{1n},
\end{aligned}$$

where $\tilde{\boldsymbol{Ad}}^{(1)}_n = \boldsymbol{C}^{-1}_{11,n} n^{-1/2}\sum_{i=1}^n \hat{\epsilon}_i \boldsymbol{x}^{(1)}_i$ and it follows from Lemma 8.1 that

$$\mathbf{P}_*\big(\|\breve{\boldsymbol{R}}^*_{1n}\| \leq c_n.n^{-1/2}\big) = 1 - o_p(1),$$

where $\{c_n\}$ is a sequence of positive constants increasing to ∞ with $c_n = o(\sqrt{\log n})$.

Again for sufficiently large n,

$$\boldsymbol{F}^{(1)}_n = n^{-1/2}\sum_{i=1}^n (\tilde{\boldsymbol{\xi}}^{(0)}_i + \tilde{\boldsymbol{\eta}}^{(0)}_i)\epsilon_i + \tilde{\boldsymbol{R}}_{2n},$$

$$\tilde{\boldsymbol{F}}^{*(1)}_n = \mu^{-1}_{G^*} n^{-1/2}\sum_{i=1}^n (\tilde{\boldsymbol{\xi}}^{(0)}_i \hat{\epsilon}_i + \tilde{\boldsymbol{\eta}}^{(0)}_i \bar{\epsilon}_i)(G^*_i - \mu_{G^*}) + \tilde{\boldsymbol{R}}^*_{2n},$$

where $\mathbf{P}(\|\tilde{\boldsymbol{R}}_{2n}\| = o(n^{-1/2})) = 1 - o(1)$ and $\mathbf{P}_*(\|\tilde{\boldsymbol{R}}^*_{2n}\| = o(n^{-1/2})) = 1 - o_p(1)$. Here, $\tilde{\boldsymbol{\xi}}_i^{(0)} = \boldsymbol{C}_{11,n}^{-1}\boldsymbol{x}_i^{(1)}$, $\tilde{\boldsymbol{\eta}}_i^{(0)} = \boldsymbol{C}_{11,n}^{-1}\tilde{\boldsymbol{\eta}}_i$ with jth component ($j \in \mathcal{A} = \{k : \beta_j \neq 0\}$) of $\tilde{\eta}_i$ is $(\frac{\lambda_n}{2n}\tilde{x}_{i,j}\frac{\gamma}{|\tilde{\beta}_{j,n}|^{\gamma+1}}\operatorname{sgn}(\hat{\beta}_{j,n}))$. Here, we have assumed without loss of generality that $\mathcal{A} = \{1, \dots, p_0\}$. and $\hat{\epsilon}_i$ and $\bar{\epsilon}_i$ are respectively Alasso and OLS residuals. Then by the Berry–Essen theorem and Lemma 3.1 of Bhattacharya and Ranga Rao (1986), we have

$$\begin{aligned} &\sup_{\boldsymbol{x}\in\mathcal{R}^{p_0}} \left|\mathbf{P}\left(\boldsymbol{F}_n^{(1)} \le \boldsymbol{x}\right) - \boldsymbol{\Phi}_{\boldsymbol{V}_n}(\boldsymbol{x})\right| = O\left(n^{-1/2}\right) \\ \text{and} \quad &\sup_{\boldsymbol{x}\in\mathcal{R}^{p_0}} \left|\mathbf{P}_*\left(\tilde{\boldsymbol{F}}_n^{*(1)} + \breve{\boldsymbol{R}}^*_{1n} \le \boldsymbol{x}\right) - \boldsymbol{\Phi}_{\tilde{\boldsymbol{V}}_n}(\boldsymbol{x})\right| = O_p\left(c_n.n^{-1/2}\right), \end{aligned} \tag{8.6}$$

where $\boldsymbol{V}_n = n^{-1}\sum_{i=1}^n(\tilde{\boldsymbol{\xi}}_i^{(0)} + \tilde{\boldsymbol{\eta}}_i^{(0)})'(\tilde{\boldsymbol{\xi}}_i^{(0)} + \tilde{\boldsymbol{\eta}}_i^{(0)})\sigma^2$ and $\tilde{\boldsymbol{V}}_n = n^{-1}\sum_{i=1}^n(\tilde{\boldsymbol{\xi}}_i^{(0)}\hat{\epsilon}_i + \tilde{\boldsymbol{\eta}}_i^{(0)}\bar{\epsilon}_i)'(\tilde{\boldsymbol{\xi}}_i^{(0)}\hat{\epsilon}_i + \tilde{\boldsymbol{\eta}}_i^{(0)}\bar{\epsilon}_i)$ and $\|\tilde{\boldsymbol{V}}_n - \boldsymbol{V}_n\| = o_p(c_n.n^{-1/2})$. Hence by Turnbull (1930) and noting (14.66) of Lemma 14.6 of Bhattacharya and Ranga Rao (1986) and the facts that $\tilde{\boldsymbol{V}}_n = O_p(1)$ & $\boldsymbol{V}_n = O(1)$, we have

$$\sup_{\boldsymbol{x}\in\mathcal{R}^{p_0}} \left|\boldsymbol{\Phi}_{\tilde{\boldsymbol{V}}_n}(\boldsymbol{x}) - \boldsymbol{\Phi}_{\boldsymbol{V}_n}(\boldsymbol{x}))\right| \le \|\tilde{\boldsymbol{V}}_n - \boldsymbol{V}_n\| = o_p\left(c_n.n^{-1/2}\right). \tag{8.7}$$

Therefore, by (8.6) and (8.7) and noting that $c_n = o(\sqrt{\log n})$, we have

$$\sup_{\boldsymbol{x}\in\mathcal{R}^{p_0}} \left|\mathbf{P}_*\left(\tilde{\boldsymbol{F}}_n^{*(1)} + \breve{\boldsymbol{R}}^*_{1n} \le \boldsymbol{x}\right) - \mathbf{P}\left(\boldsymbol{F}_n^{(1)} \le \boldsymbol{x}\right)\right| = o_p\left(\lambda_n.n^{-1/2}\right). \tag{8.8}$$

Now by (8.6), (8.8) and Taylor's expansion, it can be shown that for any $\boldsymbol{x} \in \mathcal{R}^{p_0}$,

$$\begin{aligned} \mathbf{P}_*\left(\boldsymbol{F}_n^{*(1)} \le \boldsymbol{x}\right) = \mathbf{P}\left(\boldsymbol{F}_n^{(1)} \le \boldsymbol{x}\right) - \frac{\lambda_n}{2\sqrt{n}}\left[\tilde{\boldsymbol{s}}_n^{(1)\prime}\boldsymbol{C}_{11,n}^{-1}(D_1, \dots, D_p)'\Phi_{\boldsymbol{V}_n}(\tilde{\boldsymbol{x}})\right] \\ + o_p(\lambda_n/\sqrt{n}) \end{aligned}$$

for some $\tilde{\boldsymbol{x}}$ with $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \le \|\boldsymbol{Ad}_n^{(1)}\|$ where $\boldsymbol{Ad}_n^{(1)} = \boldsymbol{C}_{11,n}^{-1}\frac{\lambda_n}{2\sqrt{n}}\boldsymbol{s}_n^{(1)}$.

Therefore, (8.5) follows from the triangle inequality and the fact that $\sup_{\boldsymbol{x}\in\mathcal{R}^{p_0}}[f(\boldsymbol{x}) + g(\boldsymbol{x})] \le \sup_{\boldsymbol{x}\in\mathcal{R}^{p_0}} f(\boldsymbol{x}) + \sup_{\boldsymbol{x}\in\mathcal{R}^{p_0}} g(\boldsymbol{x})$. For details, see the Supplementary Material [Das, Gregory and Lahiri (2019)]. $\square$

PROOF OF THEOREM 5.1. By Lemma 8.6, we have

$$\sup_{B\in\mathcal{C}_q}\left|\mathbf{P}_*\left(\boldsymbol{R}_n^* \in B\right) - \int_B \xi_n^*(\boldsymbol{x})\,d\boldsymbol{x}\right| = o_p\left(n^{-1/2}\right).$$

Now, retracting the steps of Lemma 8.6 and using the fact that $\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_n\| = o_p(n^{-(1+\delta_1)/2})$ (cf. Lemma 8.5), it can be shown that

$$\sup_{B\in\mathcal{C}_q}\left|\mathbf{P}(\boldsymbol{R}_n \in B) - \int_B \xi_n(\boldsymbol{x})\,d\boldsymbol{x}\right| = o\left(n^{-1/2}\right),$$

where

$$\xi_n(\boldsymbol{x}) = \phi(\boldsymbol{x})\Bigg[1 + \sum_{k=1}^{r}\frac{1}{k!}\Bigg\{\sum_{\boldsymbol{\alpha}=k}\tilde{\boldsymbol{b}}_n^{\alpha} H_{\boldsymbol{\alpha}}(\boldsymbol{x})\Bigg\} + \frac{1}{\sqrt{n}}\Bigg[-\frac{\mu_3}{2\sigma^3}\sum_{|\boldsymbol{\alpha}|=1}\boldsymbol{t}^{\boldsymbol{\alpha}}\bar{\boldsymbol{\xi}}_n(\boldsymbol{\alpha})H_{\boldsymbol{\alpha}}(\boldsymbol{x})$$

$$+ \frac{\mu_3}{6\sigma^3}\Bigg\{\sum_{|\boldsymbol{\alpha}|=3}\boldsymbol{t}^{\boldsymbol{\alpha}}\bar{\boldsymbol{\xi}}_n(\boldsymbol{\alpha})H_{\boldsymbol{\alpha}}(\boldsymbol{x})$$

$$- 3\sum_{|\boldsymbol{\alpha}|=3}\sum_{|\boldsymbol{\zeta}|=1}\boldsymbol{t}^{\boldsymbol{\alpha}+\boldsymbol{\zeta}}\bar{\boldsymbol{\xi}}_n(\boldsymbol{\alpha})\bar{\boldsymbol{\xi}}_n(\boldsymbol{\zeta})H_{\boldsymbol{\alpha}+\boldsymbol{\zeta}}(\boldsymbol{x})\Bigg\}\Bigg]\Bigg],$$

where $x \in \mathcal{R}^q$, $\bar{\boldsymbol{\xi}}_n(\boldsymbol{\alpha}) = n^{-1}\sum_{i=1}^{n}(\boldsymbol{\Sigma}_n^{-1/2}\boldsymbol{\xi}_i^{(0)})^{\boldsymbol{\alpha}}$. For details, see the proof of Theorem 8.2 of Chatterjee and Lahiri (2013). Now due to assumption (A.6)(i), Lemma 8.4 and Lemma 8.5 and the facts that $\|\boldsymbol{b}_n\| = O(n^{-\delta_1})$ and $\|\check{\boldsymbol{b}}_n\| = O_p(n^{-\delta_1})$, the coefficients of $n^{-1/2}$ in $\xi_n^*(\boldsymbol{x})$ converge to those of $\xi_n(\boldsymbol{x})$ in probability and $\|\tilde{\boldsymbol{b}}_n^{\boldsymbol{\alpha}} - \check{\boldsymbol{b}}_n^{\boldsymbol{\alpha}}\| = o(n^{-1/2})$, for all $\boldsymbol{\alpha}$ such that $|\boldsymbol{\alpha}| \le r_1$. Therefore, Theorem 5.1 follows. □

PROOF OF THEOREM 5.2. By Lemma 8.6, on the set $\boldsymbol{A}_{1n}$, we have for large n

$$\boldsymbol{T}_n^* + \check{\boldsymbol{b}}_n^* = \mu_{G^*}^{-1}\boldsymbol{D}_n^{(1)}\boldsymbol{C}_{11,n}^{-1}\check{\boldsymbol{W}}_n^{*(1)} + Q_{4n}^*,$$

where $\mathbf{P}_*(\|Q_{4n}^*\| \neq 0) = o(n^{-1})$. Therefore, we have

$$\begin{aligned}\check{\boldsymbol{R}}_n^* &= \mu_{G^*}^{-1}\tilde{\boldsymbol{\Sigma}}_n^{-1/2}\boldsymbol{D}_n^{(1)}\boldsymbol{C}_{11,n}^{-1}\check{\boldsymbol{W}}_n^{*(1)}\Bigg[1 - \frac{1}{2\hat{\sigma}_n^2}(\hat{\sigma}_n^* - \hat{\sigma}_n) + \frac{3}{4\hat{\sigma}_n^4}\frac{(\hat{\sigma}_n^* - \hat{\sigma}_n)^2}{2}\Bigg] + Q_{5,n}^* \\ &= \boldsymbol{R}_{3n}^* + Q_{5n}^* \qquad \text{(say)},\end{aligned}$$

where $\mathbf{P}_*(\|Q_{5n}^*\| = o(n^{-1})) = o_p(n^{-1})$.

Thus by Corollary 2.6 of Bhattacharya and Ranga Rao (1986), the Edgeworth expansions of $\boldsymbol{R}_{3n}^*$ and $\check{\boldsymbol{R}}_n^*$ agree up to order $o(n^{-1})$. Now, similar to Lemma 8.6, using the transformation technique of Bhattacharya and Ghosh (1978), one can obtain the three-term Edgeworth expansion of $\boldsymbol{R}_{3n}^*$, say $\pi_n^*(\boldsymbol{x})$. Similarly, one can obtain the three-term Edgeworth expansion of $\check{\boldsymbol{R}}_n$, say $\pi_n(\boldsymbol{x})$. Theorem 5.2 now follows by comparing terms of $\pi_n(\boldsymbol{x})$ and $\pi_n^*(\boldsymbol{x})$, similar to as in Theorem 5.1. For details, see the Supplementary Material [Das, Gregory and Lahiri (2019)]. □

PROOF OF THEOREM 5.3. The first part follows by Lemma 8.6(b) and retracing the proof of Theorem 5.1. The second part follows analogously to the proof of Theorem 5.2. □

PROOFS OF THEOREMS 5.4 AND 5.5. The first part follows by Lemma 8.6(b) with the use of Hoeffding's and Bernstein's inequality in place of Lemma 8.1 and retracing the proof of Theorem 5.1. The second part follows analogously to the proof of Theorem 5.2. □

PROOF OF COROLLARY 5.1. We need to show $\mathbf{P}(|\check{R}_{j,n}| > \hat{u}^j_\Omega$ for at least one $j) \le \alpha$. Define the set $\mathcal{A}_n = \{j : \beta_{j,n} \neq 0\}$. Now note that

$$\begin{aligned}
&\mathbf{P}\big(|\check{R}_{j,n}| > \hat{u}^j_\Omega \text{ for at least one } j\big)\\
&\quad \le \sum_{j=1}^{l_n} \mathbf{P}\big(|\check{R}_{j,n}| > \hat{u}^j_\Omega\big)\\
&\quad \le \sum_{j=1}^{l_n} \mathbf{E}\big[\mathbf{P}_*\big(|\check{R}^*_{j,n}| > \hat{u}^j_\Omega\big)\big] + \sum_{j=1}^{l_n} \mathbf{E}\big[\big|\mathbf{P}_*\big(|\check{R}^*_{j,n}| > \hat{u}^j_\Omega\big) - \mathbf{P}\big(|\check{R}_{j,n}| > \hat{u}^j_\Omega\big)\big|\big].
\end{aligned}$$

Now note that $\mathbf{P}_*(|\check{R}^*_{j,n}| > \hat{u}^j_\Omega)$ is a nonnegative random variable that is bounded above by Ω, and hence $\mathbf{E}[\mathbf{P}_*(|\check{R}^*_{j,n}| > \hat{u}^j_\Omega)] \le \Omega$, for every $j \in \{1, \dots, l_n\}$. In case of the second term, for any $j \in \{1, \dots, l_n\}$ we have

$$\begin{aligned}
&\mathbf{E}\big[\big|\mathbf{P}_*\big(|\check{R}^*_{j,n}| > \hat{u}^j_\Omega\big) - \mathbf{P}\big(|\check{R}_{j,n}| > \hat{u}^j_\Omega\big)\big|\big]\\
&\quad = \mathbf{E}\big[\big(\big|\mathbf{P}_*\big(|\check{R}^*_{j,n}| > \hat{u}^j_\Omega\big) - \mathbf{P}\big(|\check{R}_{j,n}| > \hat{u}^j_\Omega\big)\big|\big)\mathbb{1}(\boldsymbol{A}_{1n})\big]\\
&\qquad + \mathbf{E}\big[\big(\big|\mathbf{P}_*\big(|\check{R}^*_{j,n}| > \hat{u}^j_\Omega\big) - \mathbf{P}\big(|\check{R}_{j,n}| > \hat{u}^j_\Omega\big)\big|\big)\mathbb{1}(\boldsymbol{A}^c_{1n})\big]\\
&\quad = L^{(1)}_j + L^{(2)}_j \qquad \text{(say)},
\end{aligned}$$

where the set $\boldsymbol{A}_{1n}$ is as defined at the beginning of this section. Since $\mathbf{P}(\boldsymbol{A}_{1n}) \ge 1 - o(n^{-1/2})$ and $l_n = O(n^{1/2})$, $\sum_{j=1}^{l_n} L^{(2)}_j = o(1)$. Now $\sum_{j=1}^{l_n} L^{(1)}_j = o(1)$ follows from theorems 5.2, 5.3, 5.4 and 5.5 and due to the fact that $\hat{\mathcal{A}}_n = \mathcal{A}_n$ on the set A_{1n} for sufficiently large n. Therefore, Corollary 5.1 follows. □

9. Conclusion. Second-order results of the perturbation bootstrap method in Alasso are established. It is shown that the naive perturbation bootstrap of Minnier, Tian and Cai (2011) is not sufficient for correcting the distribution of the Alasso estimator up to second order. Novel modification is proposed in the bootstrap objective function to achieve second-order correctness even in high dimension. The modification is also shown to be computationally efficient. Thus, in a way the results in this paper establish the perturbation bootstrap method as a significant refinement of the approximation of the exact distribution of the Alasso estimator over oracle normal approximation. This is an important finding from the perspective of valid inferences regarding the regression parameters based on adaptive Lasso estimator.

SUPPLEMENTARY MATERIAL

Supplement to "Perturbation bootstrap in adaptive Lasso" (DOI: 10.1214/18-AOS1741SUPP; .pdf). Details of the proofs and additional simulation studies are provided.

REFERENCES

BHATTACHARYA, R. N. and GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6** 434–451. MR0471142

BHATTACHARYA, R. N. and RANGA RAO, R. (1986). *Normal Approximation and Asymptotic Expansions*. John Wiley & Sons, New York.

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575

BÜHLMANN, P., KALISCH, M. and MEIER, L. (2014). High-dimensional statistics with a view towards applications in biology. *Annu. Rev. Statist. Appl.* **1** 255–278.

CAMPONOVO, L. (2015). On the validity of the pairs bootstrap for Lasso estimators. *Biometrika* **102** 981–987. MR3431568

CANER, M. and FAN, Q. (2010). The adaptive Lasso method for instrumental variable selection. Working paper, North Carolina State Univ., Raleigh, NC.

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

CHATTERJEE, A. and LAHIRI, S. N. (2010). Asymptotic properties of the residual bootstrap for Lasso estimators. *Proc. Amer. Math. Soc.* **138** 4497–4509. MR2680074

CHATTERJEE, A. and LAHIRI, S. N. (2011). Bootstrapping Lasso estimators. *J. Amer. Statist. Assoc.* **106** 608–625. MR2847974

CHATTERJEE, A. and LAHIRI, S. N. (2013). Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *Ann. Statist.* **41** 1232–1259. MR3113809

DAS, D., GREGORY, K. and LAHIRI, S. N. (2019). Supplement to "Perturbation bootstrap in adaptive Lasso." DOI:10.1214/18-AOS1741SUPP.

DAS, D. and LAHIRI, S. N. (2019). Second order correctness of perturbation bootstrap M-estimator of multiple linear regression parameter. *Bernoulli* **25** 654–682. MR3892332

DEZEURE, R., BÜHLMANN, P. and ZHANG, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *TEST* **26** 685–719. Available at arXiv:1606.03940. MR3713586

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

FUK, D. H. and NAGAEV, S. V. (1971). Probabilistic inequalities for sums of independent random variables. *Teor. Veroyatn. Primen.* **16** 660–675. MR0293695

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

HALL, P., LEE, Y. K., PARK, B. U. and PAUL, D. (2009). Tie-respecting bootstrap methods for estimating distributions of sets and functions of eigenvalues. *Bernoulli* **15** 380–401. MR2543867

JIN, Z., YING, Z. and WEI, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88** 381–390. MR1844838

KNIGHT, K. and FU, W. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

MINNIER, J., TIAN, L. and CAI, T. (2011). A perturbation method for inference on regularized regression estimates. *J. Amer. Statist. Assoc.* **106** 1371–1382. MR2896842

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TURNBULL, H. W. (1930). A matrix form of Taylor's theorem. *Proc. Edinb. Math. Soc.* (2) **33** 33–54.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

WANG, X. and SONG, L. (2011). Adaptive Lasso variable selection for the accelerated failure models. *Comm. Statist. Theory Methods* **40** 4372–4386. MR2864162

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. Roy. Statist. Soc. Ser. B* **76** 217–242. MR3153940

ZHOU, Q. M., SONG, P. X.-K. and THOMPSON, M. E. (2012). Information ratio test for model misspecification in quasi-likelihood inference. *J. Amer. Statist. Assoc.* **107** 205–213. MR2949352

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

D. DAS
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN—MADISON
1300 UNIVERSITY AVN.
MADISON, WISCONSIN 53706
USA
E-MAIL: ddas25@wisc.edu

K. GREGORY
DEPARTMENT OF STATISTICS
UNIVERSITY OF SOUTH CAROLINA
1523 GREENE ST
COLUMBIA, SOUTH CAROLINA 29201
USA
E-MAIL: gregorkb@stat.sc.edu

S. N. LAHIRI
DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
2311 STINSON DR
RALEIGH, NORTH CAROLINA 27607
USA
E-MAIL: snlahiri@ncsu.edu