The Annals of Statistics
2013, Vol. 41, No. 3, 1232–1259
DOI: 10.1214/13-AOS1106

RATES OF CONVERGENCE OF THE ADAPTIVE LASSO ESTIMATORS TO THE ORACLE DISTRIBUTION AND HIGHER ORDER REFINEMENTS BY THE BOOTSTRAP

BY A. CHATTERJEE[1] AND S. N. LAHIRI[2]

Indian Statistical Institute and North Carolina State University

Zou [*J. Amer. Statist. Assoc.* **101** (2006) 1418–1429] proposed the Adaptive LASSO (ALASSO) method for simultaneous variable selection and estimation of the regression parameters, and established its oracle property. In this paper, we investigate the rate of convergence of the ALASSO estimator to the oracle distribution when the dimension of the regression parameters may grow to infinity with the sample size. It is shown that the rate critically depends on the choices of the penalty parameter and the initial estimator, among other factors, and that confidence intervals (CIs) based on the oracle limit law often have poor coverage accuracy. As an alternative, we consider the residual bootstrap method for the ALASSO estimators that has been recently shown to be consistent; cf. Chatterjee and Lahiri [*J. Amer. Statist. Assoc.* **106** (2011a) 608–625]. We show that the bootstrap applied to a suitable studentized version of the ALASSO estimator achieves second-order correctness, even when the dimension of the regression parameters is unbounded. Results from a moderately large simulation study show marked improvement in coverage accuracy for the bootstrap CIs over the oracle based CIs.

1. Introduction.

Consider the regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where y_i is the response, $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})'$ is a p dimensional covariate vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the regression parameter and $\{\varepsilon_i : i = 1, \ldots, n\}$ are independent and identically distributed (i.i.d.) errors. Let $\widetilde{\boldsymbol{\beta}}_n$ denote a root-n consistent estimator of $\boldsymbol{\beta}$, such as the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$. The Adaptive Lasso (ALASSO) estimator of $\boldsymbol{\beta}$ is defined as the minimizer of the weighted ℓ_1-penalized least squares criterion function,

$$\widehat{\boldsymbol{\beta}}_n = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\mathbf{u}\right)^2 + \lambda_n \sum_{j=1}^{p} \frac{|u_j|}{|\tilde{\beta}_{j,n}|^{\gamma}}, \tag{1.2}$$

Received February 2012; revised January 2013.

[1]Supported in part by the VI-MSS program of Department of Science and Technology, Government of India, and the Statistical and Applied Mathematical Sciences Institute (SAMSI), NC, USA.

[2]Supported in part by NSF Grant DMS-10-07703 and NSA Grant H98230-11-1-0130. On leave from Texas A&M University.

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

Key words and phrases. Bootstrap, Edgeworth expansion, penalized regression.

where $\lambda_n > 0$ is a regularization parameter, $\gamma > 0$ and $\tilde{\beta}_{j,n}$ is the jth component of $\widetilde{\boldsymbol{\beta}}_n$. The ALASSO provides an improvement over the LASSO and related bridge estimators that often require strong regularity conditions on the design vectors $\mathbf{x}_i$'s for consistent variable selection and that have nontrivial bias in the selected nonzero components; cf. Knight and Fu (2000), Fan and Li (2001), Yuan and Lin (2007), Zhao and Yu (2006). To highlight some of the key properties of the ALASSO, suppose for the time being, that the first p_0 components of the true regression parameter $\boldsymbol{\beta}$ are nonzero and the last $(p - p_0)$ components are zero, where $1 \le p_0 < p$. Let $\widetilde{\mathcal{I}}_n = \{j : 1 \le j \le p, \widehat{\beta}_{j,n} \neq 0\}$ denote the variables selected by the ALASSO, where $\widehat{\beta}_{j,n}$ is the jth component of $\widehat{\boldsymbol{\beta}}_n$. Zou (2006) showed that under some mild regularity conditions, for fixed p, as $n \to \infty$,

$$\mathbf{P}(\widetilde{\mathcal{I}}_n = \mathcal{I}_n) \to 1 \quad \text{and} \quad \sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}^{(1)}) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{C}_{11}^{-1}), \tag{1.3}$$

where $\mathcal{I}_n = \{1, \dots, p_0\}$, $\widehat{\boldsymbol{\beta}}_n^{(1)} = (\widehat{\beta}_{1,n}, \dots, \widehat{\beta}_{p_0,n})$, $\boldsymbol{\beta}^{(1)} = (\beta_1, \dots, \beta_{p_0})$ and $\mathbf{C}_{11}$ is the upper left $p_0 \times p_0$ submatrix of $\mathbf{C} \equiv \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'$. Thus, the ALASSO method enjoys the *oracle property* [cf. Fan and Li (2001)], that is, it can correctly identify the set of nonzero components of $\boldsymbol{\beta}$, with probability tending to 1 and at the same time, estimate the nonzero components accurately, with the same precision as that of the OLS method, in the limit.

Although the oracle property of the ALASSO estimators allows one to carry out statistical inference on the nonzero regression parameters, following variable selection, accuracy of of the resulting inference remains unknown. In this paper, we investigate the rate of convergence of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}^{(1)})$ to the oracle limit and show that the penalization term in (1.2) induces a substantial amount of bias which, although vanishes asymptotically, can lead to a poor rate of convergence. As a result, large sample inference based on the oracle distribution is not very accurate. As an alternative, we consider the bootstrap method or more precisely, the residual bootstrap method [cf. Efron (1979), Freedman (1981)], that is, the most common version of the bootstrap in a regression model like (1.1). Recently, Chatterjee and Lahiri (2010, 2011a) showed that while the residual bootstrap drastically fails for the LASSO. Rather surprisingly, it provides a valid approximation to the distribution of the centered and scaled ALASSO-estimator. Notwithstanding its success in capturing the first order limit, the accuracy of the bootstrap for the ALASSO remains unknown. In this paper, we also study the rate of bootstrap approximation to the distribution of the ALASSO estimators, with and without studentization, and develop ways to improve it, all in the more general framework where the number of regression parameters $p = p_n$ is allowed to go to infinity with the sample size n.

To describe the main findings of the paper, consider (1.1) where p, $\mathbf{x}_i$'s and $\boldsymbol{\beta}$ are allowed to depend on n (but we often suppress the subscript n to ease notation) and let $\mathbf{T}_n = \sqrt{n}\mathbf{D}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$, where $\mathbf{D}_n$ is a known $q \times p$ matrix with $\operatorname{tr}(\mathbf{D}_n \mathbf{D}_n') = O(1)$ and $q \in \mathbb{N} = \{1, 2, \dots\}$ is an integer, *not* depending on n. Thus,

$\mathbf{T}_n$ is the vector of q linear functions of $n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. Under the regularity conditions of Section 3, $\{\mathbf{T}_n : n \geq 1\}$ is asymptotically normal with mean zero and $q \times q$ asymptotic variance $\boldsymbol{\Sigma}_n$ (say). We consider the error of oracle-based normal approximation,

$$\Delta_n \equiv \sup_{B \in \mathcal{C}_q} \left| \mathbf{P}(\mathbf{T}_n \in B) - \Phi(B; \boldsymbol{\Sigma}_n) \right|,$$

where, for $k \geq 1$, $\mathcal{C}_k$ is the collection of all convex measurable subsets of $\mathbb{R}^k$ and $\Phi(\cdot; \mathbf{A})$ is the Gaussian measure on $\mathbb{R}^k$ with mean zero and $k \times k$ covariance matrix $\mathbf{A}$. Theorem 3.1 below gives an upper bound on Δ_n,

$$\Delta_n \leq \text{const}\big[n^{-1/2} + \|\mathbf{b}_n\| + c_n\big], \tag{1.4}$$

where $\mathbf{b}_n$ is a bias term that results from the penalization scheme in (1.2) and where $c_n \in (0, \infty)$ is determined by the initial $\sqrt{n}$-consistent estimator $\widetilde{\boldsymbol{\beta}}_n$ and the tuning parameter γ in (1.2). The magnitude of both these terms critically depend on the choice of the penalization parameter λ_n and the exponent γ, and either of them can make the error rate sub-optimal, that is, worse than the rate $O(n^{-1/2})$ that is attained by the oracle based OLS estimator. Further, Theorem 3.2 shows that under some additional mild conditions, the rate in (1.4) is *exact*, that is, Δ_n is also *bounded below* by a constant multiple of the sum of the three terms on the right-hand side of (1.4). Therefore, it follows that although the ALASSO estimator converges to the oracle distribution in the limit, the convergence rate can be sub-optimal. A direct implication of this result is that large sample tests and CIs based on the normal limit law of the ALASSO estimator may perform poorly, depending on the choice of the regularization parameters λ_n and γ. The simulation results of Section 6 confirm this finite samples.

Next we consider properties of bootstrap approximations to the distributions of $\mathbf{T}_n$ and $\mathbf{R}_n$, a *computationally simple* studentized version of $\mathbf{T}_n$, given by $\mathbf{R}_n = \frac{\mathbf{T}_n}{\widehat{\sigma}_n}$, where $\widehat{\sigma}_n^2$ is the sample variance of the ALASSO based residuals. Here we use a scalar studentizing factor instead of the usual matrix factor [cf. Lahiri (1994)] to reduce the computational burden. Fortunately, this does not impact the accuracy of the bootstrap approximation as σ^2 is the only unknown population parameter in the limit distribution of $\mathbf{T}_n$. Theorem 4.1 below shows that under fairly general conditions, the rate of bootstrap approximation to the distribution of $\mathbf{T}_n$ is $O_p(n^{-1/2})$. Thus, the bootstrap corrects for the effects of $\|\mathbf{b}_n\|$ and c_n in (1.4), and produces a more "accurate" approximation to the distribution of $\mathbf{T}_n$ than the oracle based normal approximation. As a consequence, bootstrap percentile CIs based on the ALASSO have a better performance compared to the large sample normal CIs based on the oracle.

The results on the studentized statistic $\mathbf{R}_n$ are more encouraging. Theorem 4.2 shows that the bootstrap applied to $\mathbf{R}_n$ has an error rate of $o_p(n^{-1/2})$ which outperforms the best possible rate, namely $O(n^{-1/2})$ of normal approximation, irrespective of the order of the terms $\|\mathbf{b}_n\|$ and c_n in (1.4). Thus, the bootstrap applied to the studentized statistic $\mathbf{R}_n$ achieves *second order correctness*. In contrast, the normal approximation to the distribution of $\mathbf{R}_n$ has an error of the order $O(n^{-1/2} + \|\mathbf{b}_n\| + c_n)$, as in the case of $\mathbf{T}_n$. As a result, bootstrap percentile-t CIs based on $\mathbf{R}_n$ are significantly more accurate than their counterparts based on normal critical points. This observation is also corroborated by the simulation results of Section 6.

In Section 4.4, a further refinement is obtained. A more careful analysis of the $o_p(n^{-1/2})$-term in Theorem 4.2 shows that although it outperforms the normal approximation over the class $\mathcal{C}_q$, this rate does not always match the "optimal" level, namely $O_p(n^{-1})$ that is attained by the bootstrap in the more classical setting of estimation of regression parameters by the OLS method with a *fixed* p. Exploiting the higher order analysis in the proof of Theorem 4.2, we carefully construct a modified studentized version $\breve{\mathbf{R}}_n$ of $\widehat{\boldsymbol{\beta}}_n$. Theorem 4.3 shows that under slightly stronger regularity conditions (compared to those in Theorem 4.2), the rate of bootstrap approximation for the modified pivot $\breve{\mathbf{R}}_n$ is $O_p(n^{-1})$. This appears to be a remarkable result because, even with a diverging p and with the regularization step, the specially constructed pivotal quantity $\breve{\mathbf{R}}_n$ attains the same optimal rate $O_p(n^{-1})$ as in the classical set up of linear regression with a fixed p.

The key technical tool used in the proofs of the results in Sections 3 and 4 is an Edgeworth expansion (EE) result for the ALASSO estimator and its studentized version, given in Theorem 7.2 of Section 7, which may be of independent interest. The derivation of the EE critically depends on the choice of the initial estimator in (1.2). In Sections 3 and 4, the initial estimator is chosen to be the OLS, which necessarily requires $p \le n$. However, in many applications, it is important to allow $p > n$. In such situations, one may use a bridge estimator [cf. Knight and Fu (2000)] in place of the OLS as the initial estimator. In Section 5, we show that under some suitable regularity conditions, the bootstrap approximation to the distributions of $\mathbf{R}_n$ and $\breve{\mathbf{R}}_n$ continue to be second order correct even for $p > n$. Here, p is allowed to grow at polynomial rates in n. More precisely, we allow $p = O(n^a)$ for any given $a > 1$, provided (in addition to certain other conditions) $\mathbf{E}|\varepsilon_1|^r < \infty$ for a sufficiently large r, depending on a. Thus, the allowable growth rate of p depends on the rate of decay of the tails of the error distribution.

The rest of the paper is organized as follows. We conclude this section with a brief literature review. In Section 2, we introduce the theoretical framework and state the regularity conditions. Results on the rate of convergence to the oracle limit law is given in Section 3. The main results on the bootstrap are given in Section 4 for the $p \le n$ case and in Section 5 for the $p > n$ case. Section 6 presents the results from a moderately large simulation study and it also gives two real data examples. An outline of the proofs of the main results is given in Section 7 and their detailed proofs are relegated to a supplementary material file; cf. Chatterjee and Lahiri (2013).

The literature on penalized regression in high dimensions has been growing very rapidly in recent years; here we give only a modest account of the work that is most related to the present paper due to space limitation. In two important papers, Tibshirani (1996) introduced the LASSO, as an estimation and variable selection method and Zou (2006) introduced the ALASSO method as an improvement over the LASSO and established its oracle property. Other popular penalized estimation and variable selection methods are given by the SCAD [Fan and Li (2001)] and the Dantzig Selector [Candes and Tao (2007)]. Properties of the ALASSO and the related methods have been investigated by many authors, including Knight and Fu (2000), Meinshausen and Bühlmann (2006), Wainwright (2006), Bunea, Tsybakov and Wegkamp (2007), Bickel, Ritov and Tsybakov (2009), Huang, Ma and Zhang (2008), Huang, Horowitz and Ma (2008), Zhang and Huang (2008), Meinshausen and Yu (2009), Pötscher and Schneider (2009), Chatterjee and Lahiri (2011b), Gupta (2012) among others. Fan and Li (2001) introduced the important notion of "oracle property" in the context of penalized estimation and variable selection by the SCAD. Post model selection inference, including the bootstrap and its variants have been investigated by Bach (2009), Chatterjee and Lahiri (2010, 2011a), Minnier, Tian and Cai (2011) and Berk et al. (2013), among others.

2. Preliminaries and the regularity conditions.

2.1. *Theoretical set up.* For deriving the theoretical results, we consider a generalized version of (1.1), where $p = p_n$ is allowed to depend on the sample size n. To highlight this, we shall denote the true parameter value by $\boldsymbol{\beta}_n$ and redefine

$$\mathbf{T}_n = \sqrt{n}\mathbf{D}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n),$$

where, as in Section 1, $\mathbf{D}_n$ is a $q \times p_n$ (known) matrix satisfying $\operatorname{tr}(\mathbf{D}_n\mathbf{D}_n') = O(1)$, and q does not depend on n. Also, for the $p \le n$ case, that is, in Sections 3 and 4, we shall take the initial estimator $\tilde{\boldsymbol{\beta}}_n$ to be the OLS of $\boldsymbol{\beta}_n$, given by $\tilde{\boldsymbol{\beta}}_n = [\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i']^{-1}\sum_{i=1}^n \mathbf{x}_i y_i$.

Let $I_n = \{j : 1 \le j \le p_n, \beta_{j,n} \ne 0\}$ be the (population) set of nonzero regression coefficients, where $\beta_{j,n}$ is the jth component of $\boldsymbol{\beta}_n$. The ALASSO yields an estimator $\widehat{I}_n \equiv \{j : 1 \le j \le p_n, \widehat{\beta}_{j,n} \ne 0\}$ of I_n. For notational simplicity, we shall assume that $I_n = \{1, \dots, p_{0n}\}$ and also suppress the dependence on n in p_n, p_{0n}, etc., when there is no chance of confusion.

2.2. *Conditions.* Let $\mathbf{C}_n = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'$. Write $\mathbf{C}_n = ((c_{i,j,n}))$ and $\mathbf{C}_n^{-1} = ((c_n^{i,j}))$, when it exists. Partition $\mathbf{C}_n$ as

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

where $\mathbf{C}_{11,n}$ is $p_0 \times p_0$. Similarly, let $\mathbf{D}_n^{(1)}$ is the $q \times p_0$ submatrix of $\mathbf{D}_n$, consisting of the first p_0 columns of $\mathbf{D}_n$. Let $\bar{\mathbf{x}}_n = n^{-1}\sum_{i=1}^n \mathbf{x}_i$ and let $\bar{\mathbf{x}}_n^{(1)}$ denote the first p_0 components of $\bar{\mathbf{x}}_n$. Define

$$\mathbf{\Sigma}_n^{(0)} = \begin{bmatrix} \mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}(\mathbf{D}_n^{(1)})'\sigma^2 & \mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}\bar{\mathbf{x}}_n^{(1)} \cdot \mathbf{E}(\varepsilon_1^3) \\ (\bar{\mathbf{x}}^{(1)})'\mathbf{C}_{11,n}^{-1}(\mathbf{D}_n^{(1)})' \cdot \mathbf{E}(\varepsilon_1^3) & \mathrm{Var}(\varepsilon_1^2) \end{bmatrix},$$

which is used in condition (C.3) below. Let $\mathbf{A}_{i\cdot}$ and $\mathbf{A}_{\cdot j}$, respectively, denote the ith row and the jth column of a matrix $\mathbf{A}$, and let $\mathbf{A}'$ denote the transpose of $\mathbf{A}$. For $x, y \in \mathbb{R}$, let $x \vee y = \max\{x, y\}$, $x_+ = \max\{x, 0\}$ and $\mathrm{sgn}(x) = -1, 0, 1$ according as $x < 0$, $x = 0$ and $x > 0$. Let $\iota = \sqrt{-1}$. Unless otherwise stated, limits in the order symbols are taken by letting $n \to \infty$.

We shall make use of the following conditions:

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

$$(\mathbf{x}'\mathbf{C}_{12,n}\mathbf{y})^2 \le \delta^2(\mathbf{x}'\mathbf{C}_{11,n}\mathbf{x}) \cdot (\mathbf{y}'\mathbf{C}_{22,n}\mathbf{y}) \qquad \text{for all } \mathbf{x} \in \mathbb{R}^{p_0}, \mathbf{y} \in \mathbb{R}^{p-p_0}.$$

(C.2) Let η_n and $\eta_{11,n}$ denote the smallest eigen-values of $\mathbf{C}_n$ and $\mathbf{C}_{11,n}$, respectively.
- (i) $\eta_{11,n} > Kn^{-a}$ for some $K \in (0, \infty)$ and $a \in [0, 1]$.
- (ii) $\max\{n^{-1}\sum_{i=1}^n(|x_{i,j}|^r + |\tilde{x}_{i,j}|^r) : 1 \le j \le p\} = O(1)$, where $\tilde{x}_{i,j}$ is the jth element of $(\mathbf{x}_i'\mathbf{C}_n^{-1})$ (for $p \le n$) and $r \ge 3$ is an integer (to be specified in the statements of theorems).

(C.3) There exists a $\delta \in (0, 1)$ such that for all $n > \delta^{-1}$:
- (i) $\sup\{\mathbf{x}'\mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}(\mathbf{D}_n^{(1)})'\mathbf{x} : \mathbf{x} \in \mathbb{R}^q, \|\mathbf{x}\| = 1\} < \delta^{-1}$.
- (ii) $\inf\{\mathbf{x}'\mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}(\mathbf{D}_n^{(1)})'\mathbf{x} : \mathbf{x} \in \mathbb{R}^q, \|\mathbf{x}\| = 1\} > \delta$.
- (ii)′ $\inf\{\mathbf{t}'\mathbf{\Sigma}_n^{(0)}\mathbf{t} : \mathbf{t} \in \mathbb{R}^{q+1}, \|\mathbf{t}\| = 1\} > \delta$.

(C.4) $\max\{|\beta_{j,n}| : j \in I_n\} = O(1)$ and $\min\{|\beta_{j,n}| : j \in I_n\} \ge Kn^{-b}$, for some $K \in (0, \infty)$ and $b \in [0, 1/2)$, such that $a + 2b \le 1$, where a is as in (C.2)(i):

(C.5)
- (i) $\mathbf{E}(\varepsilon_1) = 0$, $\mathbf{E}(\varepsilon_1^2) = \sigma^2 \in (0, \infty)$ and $\mathbf{E}|\varepsilon_1|^r < \infty$, for some $r \ge 3$.
- (ii) ε_1 satisfies Cramér's condition: $\limsup_{|t|\to\infty} |\mathbf{E}(\exp(\iota t\varepsilon_1))| < 1$.
- (ii)′ $(\varepsilon_1, \varepsilon_1^2)$ satisfies Cramér's condition,

$$\limsup_{\|(t_1,t_2)\|\to\infty} |\mathbf{E}\exp(\iota \cdot (t_1\varepsilon_1 + t_2\varepsilon_1^2))| < 1.$$

(C.6) There exists $\delta \in (0, 1)$ such that for all $n \ge \delta^{-1}$,

$$\frac{\lambda_n}{\sqrt{n}} \le \delta^{-1}n^{-\delta}\min\left\{\frac{n^{-b\gamma}}{p_0}, \frac{n^{-b\gamma-a/2}}{\sqrt{p_0}}, n^{-a}\right\} \quad \text{and}$$

$$\frac{\lambda_n}{\sqrt{n}} \cdot n^{\gamma/2} \ge \delta n^\delta \max\{n^a p_0, p_0^{3/2}n^{b(1-\gamma)_+}\}.$$

We now comment on the conditions. Condition (C.1) is equivalent to saying that the multiple correlation between relevant variables (with $\beta_{j,n} \neq 0$) and the spurious variables ($\beta_{j,n} = 0$) is strictly less than one, in absolute value. This condition is weaker than assuming orthogonality of the two sets of variables. Variants of this condition has been used in the literature, particularly in the context of the Lasso; see Meinshausen and Yu (2009), Huang, Horowitz and Ma (2008), Chatterjee and Lahiri (2011a), and the references therein.

Condition (C.2) gives the regularity conditions on the design matrix that are needed for establishing an $(r-2)$th order EE for the ALASSO estimator and its bootstrap versions. (C.2)(i) requires a lower bound on the smallest eigen-value of the submatrix $\mathbf{C}_{11,n}$ corresponding to the relavent variables (with $\beta_{j,n} \neq 0$), in the increasing dimensional case. When p is bounded, $\mathbf{C}_n \to \mathbf{C}$ (elementwise) and $\mathbf{C}$ is nonsingular, this condition holds with $a = 0$. Condition (C.2)(ii) is a uniform bound on the ℓ_r-norms of the sequences $\{x_{i,j}\}_{i=1}^n$, $\{\tilde{x}_{i,j}\}_{i=1}^n$, that are needed for obtaining a uniform bound on the rth order moments of the weighted sums $\sum_{i=1}^n x_{i,j}\varepsilon_i$ and $\sum_{i=1}^n \tilde{x}_{i,j}\varepsilon_i$, for $1 \leq j \leq p$. Note that for $r = 2$, the condition $\max\{n^{-1}\sum_{i=1}^n |x_{i,j}|^r : 1 \leq j \leq p\} = O(1)$ is equivalent to requiring that the diagonal elements of the $p \times p$ matrix $\mathbf{C}_n$ be uniformly bounded. Similarly, for $r = 2$,

$$\begin{aligned} n^{-1}\sum_{i=1}^n |\tilde{x}_{i,j}|^r &= (\mathbf{C}_n^{-1})_{j\cdot}\left(n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{C}_n^{-1})_{\cdot j} \\ &= (\mathbf{C}_n^{-1})_{j\cdot}\mathbf{C}_n(\mathbf{C}_n^{-1})_{\cdot j} = (\mathbb{I}_p)_{j\cdot}(\mathbf{C}_n^{-1})_{\cdot j} = c_n^{j,j}, \end{aligned}$$

where $\mathbb{I}_p$ denotes the identity matrix of order p. Thus, for $r = 2$,

$$\max\left\{n^{-1}\sum_{i=1}^n |\tilde{x}_{i,j}|^r : 1 \leq j \leq p\right\} = O(1), \tag{2.1}$$

if and only if the diagonal elements of $\mathbf{C}_n^{-1}$ are uniformly bounded. Condition (C.2)(ii) is a stronger version of these conditions with $r \geq 3$, dictated by the order of the EE one is interested in.

Conditions (C.3)(i) and (C.3)(ii) require that the maximum and the minimum eigen-values of the $q \times q$ matrix $\mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}(\mathbf{D}_n^{(1)})'$ be bounded away from zero and infinity, respectively. A sufficient condition is the existence of a nonsingular limit of $\mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}(\mathbf{D}_n^{(1)})'$, which we do *not* assume. (C.3)(ii)$'$ is a stronger form of (C.3)(ii) that is needed for the studentized case only. Note that (C.3) rules out inference on individual zero components of $\boldsymbol{\beta}_n$ (as $\mathbf{D}_n^{(1)} = \mathbf{0}$ in this case). The main results of the paper are valid only for linear combinations of the ALASSO estimator that put nontrivial weights on at least one nonzero component of $\boldsymbol{\beta}_n$.

Next consider condition (C.4) which makes it possible to separate out the signal from the noise by the ALASSO. It requires the minimum of the nonzero coefficients to be of coarser order than $O(n^{-1/2})$, so that the coefficients are not masked by the estimation error, which is of the order $O_p(n^{-1/2})$. It is worth pointing out that the results of the paper remain valid if the requirement $a + 2b \le 1$ in condition (C.4) is replaced by a somewhat weaker condition $n^{a+2b} = O(np_0)$. Condition (C.5) is a moment and smoothness condition on the error variables. These are required for the validity of an $(r-2)$th order EE, $r \ge 3$, where (C.5)(ii) is used for $\mathbf{T}_n$ and its stronger version (C.5)(ii)$'$ for the studentized cases, respectively.

Finally, consider condition (C.6). When p_0, the number of nonzero components of $\boldsymbol{\beta}_n$ is fixed (but the total number of parameters p may tend to ∞), we may suppose that $\boldsymbol{\beta}_n = \boldsymbol{\beta}$ for all $n \ge 1$ and hence, the nonzero components of $\boldsymbol{\beta}_n$ are bounded away from zero. If, in addition, the submatrix $\mathbf{C}_{11,n}$ converges elementwise to a $p_0 \times p_0$ nonsingular matrix $\mathbf{C}$, then $a = b = 0$. In this case, condition (C.6) is equivalent to

$$\frac{\lambda_n}{\sqrt{n}} + \left[\frac{\lambda_n}{\sqrt{n}} \cdot n^{\gamma/2}\right]^{-1} = O(n^{-\delta})$$

for some $\delta > 0$. This condition may be compared to the condition

$$\frac{\lambda_n}{\sqrt{n}} + \left[\frac{\lambda_n}{\sqrt{n}} \cdot n^{\gamma/2}\right]^{-1} = o(1),$$

that was imposed by Zou (2006) to establish the asymptotic distribution (and the *oracle* property) of the ALASSO, further assuming that p itself is fixed. Thus, for a regression problem with finitely many nonzero regression parameters and a *nice* design matrix, the EE results hold under a slight strengthening of the Zou (2006) conditions on λ_n and γ. It is interesting to note that the growth rate of the zero components $(p - p_0)$ (or p itself) does not have a direct impact on λ_n and γ in condition (C.6). However, when either $p_0 \to \infty$ or some of the nonzero components of $\boldsymbol{\beta}_n$ become small, the choices of λ_n and γ start to depend on the associated rates. A similar behavior ensues for a nearly singular submatrix $\mathbf{C}_{11,n}$. Further, note that for any given values of $a \in [0, 1]$ and $b \in [0, 1/2)$, we may allow $p_0 = O(n)$ (with $p_0 \le n$), by choosing λ_n and γ^{-1} suitably small. See Remark 1 in Section 3 for more details on the implications of these conditions.

3. Rates of convergence to the oracle distribution. The main results of this section give upper and lower bounds on the accuracy of approximation by the limiting oracle distribution for the ALASSO. To describe the terms in the bounds, let $\mathbf{b}_n = \mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} \mathbf{s}_n^{(1)} \cdot \frac{\lambda_n}{\sqrt{n}}$, where $\mathbf{s}_n^{(1)}$ is a $p_0 \times 1$ vector with jth component $s_{j,n} = \operatorname{sgn}(\beta_{j,n})|\beta_{j,n}|^{-\gamma}$, $1 \le j \le p_0$. Also let $\boldsymbol{\Gamma}_n = \mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} \boldsymbol{\Lambda}_n^{(1)} \mathbf{C}_{11,n}^{-1} (\mathbf{D}_n^{(1)})^{-1}$ where $\boldsymbol{\Lambda}_n^{(1)}$ is a diagonal matrix with (j, j)th element given by $\operatorname{sgn}(\beta_{j,n})|\beta_{j,n}|^{-(\gamma+1)}$, $1 \le j \le p_0$. Also, for a $k \times k$ nonnegative definite matrix $\boldsymbol{\Sigma}$, let $\Phi(\cdot : \boldsymbol{\Sigma})$ denote the Gaussian measure on $\mathbb{R}^k$ with zero mean and covariance matrix $\boldsymbol{\Sigma}$.

Then we have the following result:

THEOREM 3.1. *Suppose that conditions* (C.1)–(C.6) *hold with* $r = 4$ *and that* $\tilde{\boldsymbol{\beta}}_n$ *is the OLS of* $\boldsymbol{\beta}_n$. *Then*

$$\Delta_n \equiv \sup_{B \in \mathcal{C}_q} \left|\mathbf{P}(\mathbf{T}_n \in B) - \Phi\big(B : \sigma^2 \mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} (\mathbf{D}_n^{(1)})'\big)\right|$$

$$= O\left(n^{-1/2} + \|\mathbf{b}_n\| + \frac{\lambda_n}{n} \cdot n^{a+b(\gamma+1)}\right).$$

Theorem 3.1 gives a precise description of the quantities that determine the rate of convergence to the normal limit. In particular, the ALASSO estimator has a bias that may lead to an inferior rate of convergence to the limiting normal distribution [compared to the standard $O(n^{-1/2})$ rate], depending on the choice of the penalty constant λ_n, the exponent γ and the rate of decay of the smallest of the regression parameters. In addition, there is a third term, of the order $a_{3,n} \equiv \lambda_n \cdot n^{-1+a+b(\gamma+1)}$ that results from the use of the initial estimator $\widetilde{\boldsymbol{\beta}}_n$ in the ALASSO penalization scheme and that can also lead to a sub-$n^{-1/2}$-rate of convergence to the normal limit.

We next show that under some mild conditions, the bound given in Theorem 3.1 is precise in the sense that, in general, it cannot be improved upon.

THEOREM 3.2. *Suppose that the conditions of* 3.1 *hold and that* $\mathbf{E}\varepsilon_1^3 \neq 0$, $\liminf_{n\to\infty} \sum_{|\boldsymbol{\alpha}|=3} |(\mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} \bar{\mathbf{x}}_n^{(1)})^{\boldsymbol{\alpha}}| \neq 0$, $n^{a+b(\gamma+1)} = O(\mathrm{tr}(\boldsymbol{\Gamma}_n))$ *and* $n^{b\gamma} = O(\|\mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} \mathbf{s}_n^{(1)}\|)$. *Then*

$$\Delta_n \asymp \left[n^{-1/2} + \frac{\lambda_n}{\sqrt{n}} \cdot n^{b\gamma} + \frac{\lambda_n}{n} \cdot n^{a+b(\gamma+1)}\right],$$

where we write $a_n \asymp b_n$ *if* $a_n = O(b_n)$ *and* $b_n = O(a_n)$ *as* $n \to \infty$.

Note that under the additional conditions of Theorem 3.2, the co-efficients of the first and the third terms on the right-hand side of the display above are non-negligible in the limit and $\|\mathbf{b}_n\| \geq K \frac{\lambda_n}{\sqrt{n}} \cdot n^{b\gamma}$ for some constant $K \in (0, \infty)$. As a result, the leading terms in the EE for $\mathbf{T}_n$ that determine the upper bound in Theorem 3.1 are also bounded from below by constant multiples of the three factors appearing in Theorem 3.2. As a consequence, the exact rate of approximation by the oracle distribution to the centered and scaled ALASSO estimator $\mathbf{T}_n$ is given by the maximum of these three terms. In Remark 1 below, we discuss in more details the effects of the choices of the penalty constant λ_n, the exponent γ, etc. on the accuracy of the oracle based normal approximation.

REMARK 1. Suppose that $\lambda_n \sim K n^c$ for some $K \in (0, \infty)$ and $c \in \mathbb{R}$ and let $\|\mathbf{C}_{11,n}^{-1/2} \mathbf{s}_n^{(1)}\| = O(n^{\gamma b})$. Then $\|\mathbf{b}_n\| \leq \|\mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1/2}\| \cdot \|\mathbf{C}_{11,n}^{-1/2} \mathbf{s}_n^{(1)}\| \lambda_n / \sqrt{n} =$

$O(\lambda_n n^{-1/2+\gamma b})$. Hence, under the conditions of Theorem 3.1, the rate of normal approximation for $\mathbf{T}_n$ is given by

$$\max\{n^{-1/2}, n^{c+b\gamma-1/2}, n^{a+b(\gamma+1)+c-1}\}.$$

Here, a sub-optimal rate results if either $b\gamma + c > 0$ or $a + b(1+\gamma) + c > 1/2$. Further, the bias term is the leading sub-optimal term whenever

$$a + b < 1/2 \quad \text{and} \quad b\gamma + c > 0. \tag{3.1}$$

In this case, using the EE results from Section 7 [cf. Theorem 7.2(a)], one can conclude that, for a linear function of $\boldsymbol{\beta}_n$ (i.e., for a $1 \times p$ vector $\mathbf{D}_n$ with $q = 1$), the errors in coverage probabilities of *both one and two-sided* confidence intervals (CIs) based on the oracle normal critical points are $O(n^{-1/2+(b\gamma+c)})$. This rate is much worse than the available optimal rates, particularly in the two-sided case.

By a similar reasoning, the third term is the dominant sub-optimal term whenever

$$a + b > 1/2 \quad \text{and} \quad a + b(\gamma + 1) + c \in (1/2, 1). \tag{3.2}$$

In this case, Theorem 7.2(a) shows that one-sided CIs based on the oracle distribution r has a sub-optimal error. However, as the corresponding term in the EE for $\mathbf{T}_n$ is even, it no longer contributes to the error of coverage probability in the two-sided case.

Finally the optimal rate of convergence in Theorem 3.2 holds, provided

$$c + b\gamma \le 0 \quad \text{and} \quad a + b(\gamma + 1) + c \le 1/2.$$

Since $a \ge 0$, $b \ge 0$ and $\gamma > 0$, the first inequality requires $c \le 0$, that is, $\lambda_n = O(1)$. Further, for $ab > 0$, that is, when both the smallest eigen-value $\eta_{11,n}$ of $\mathbf{C}_{11,n}$ and the minimum of the nonzero components (say $\beta_{1n}^{\min}$) of the regression vector $\boldsymbol{\beta}_n$ tend to zero, these inequalities require that c be chosen to be a sufficiently big negative number (and thus, λ_n to be a *small* positive number). This in turn leads to an inferior performance of the ALASSO for variable selection. In the next section, we show that the bootstrap attains the optimal rate of approximation to the distribution of $\mathbf{T}_n$ without requiring such unreasonable conditions on the choice of λ_n.

4. Accuracy of the bootstrap.

4.1. *The residual bootstrap.* For the sake of completeness, we now briefly describe the residual bootstrap [cf. Freedman (1981)]. Let $e_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_n$, $i = 1, \dots, n$ denote the residuals based on the ALASSO estimator, and let $\check{e}_i = e_i - \bar{e}_n$, $i = 1, \dots, n$, where $\bar{e}_n = n^{-1}\sum_{i=1}^n e_i$. Next, select a random sample of size n with replacement from $\{\check{e}_1, \dots, \check{e}_n\}$, and denote it by $\{e_1^*, \dots, e_n^*\}$. Define the residual bootstrap observations

$$y_i^* = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_n + e_i^*, \qquad i = 1, \dots, n.$$

Note that the centering step ensures the model requirement $\mathbf{E}\varepsilon_1 = 0$ for the bootstrap error variable e_1^*. The bootstrap version of a statistic is defined by replacing $\{(y_i, \mathbf{x}_i') : i = 1, \ldots, n\}$ with $\{(y_i^*, \mathbf{x}_i') : i = 1, \ldots, n\}$ and $\boldsymbol{\beta}_n$ with $\widehat{\boldsymbol{\beta}}_n$. For example, the bootstrap version ALASSO estimator is given by

$$\boldsymbol{\beta}_n^* = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i^* - \mathbf{x}_i'\mathbf{u})^2 + \lambda_n \sum_{j=1}^{p} \frac{|u_j|}{|\tilde{\beta}_{j,n}^*|^{\gamma}}, \tag{4.1}$$

where $\widetilde{\boldsymbol{\beta}}_n^* = (\tilde{\beta}_{1,n}^*, \ldots, \tilde{\beta}_{p,n}^*)'$ is the bootstrap version of the initial estimator $\tilde{\boldsymbol{\beta}}_n$ (which is given by the OLS in this section), obtained by replacing the y_i's with y_i^*'s. The bootstrap version of $\mathbf{T}_n$ is then defined as $\mathbf{T}_n^* = \sqrt{n}\mathbf{D}_n(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n)$. Similarly, define $\mathbf{R}_n^*$ and $\check{\mathbf{R}}_n^*$.

4.2. *Rates of bootstrap approximation for* $\mathbf{T}_n$. The following result shows that the bootstrap approximation to the distribution of $\mathbf{T}_n$ attains the rate $O_p(n^{-1/2})$ under regularity conditions (C.1)–(C.6).

THEOREM 4.1. *If conditions* (C.1)–(C.6) *hold with* $r = 4$, *then*

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

A comparison of Theorem 4.1 and the results of Section 3 shows that the bootstrap approximation attains the optimal rate $O_p(n^{-1/2})$, *irrespective* of the order of magnitudes of the bias term $\|\mathbf{b}_n\|$ and of the third term $a_{3,n}$ in Theorem 3.1. In particular, this rate is attainable even when the smallest eigen-value $\eta_{11,n}$ of $\mathbf{C}_{11,n}$ or the minimum of the nonzero components (say $\beta_{1n}^{\min}$) of the regression vector $\boldsymbol{\beta}_n$ tend to zero. Most importantly, the bootstrap approximation to the ALASSO estimator attains the same level of accuracy in increasing dimensions as in the simpler case of the OLS of regression parameters when the dimension p of the regression parameter is fixed and no penalization is used. Thus, the bootstrap approximation for $\mathbf{T}_n$ is in a way immune to the effects of high dimensions.

4.3. *Rates of bootstrap approximation for* $\mathbf{R}_n$. As is well known in the fixed p case [cf. Hall (1992)], the bootstrap gives a more accurate approximation when it is applied to a pivotal quantity, such as a studentized version of a statistic, rather than to its nonpivotal version, like $\mathbf{T}_n$. Here we consider the following studentized version of the ALASSO estimator:

$$\mathbf{R}_n = \mathbf{T}_n/\widehat{\sigma}_n,$$

where $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n} \check{e}_i^2$ and $\check{e}_1, \ldots, \check{e}_n$ are the centered residuals (cf. Section 4.1). As explained in Section 1, this differs from the standard version of the studentized statistic $\tilde{\mathbf{R}}_n = \widehat{\mathbf{V}}_n^{-1/2}\mathbf{T}_n$ where $\widehat{\mathbf{V}}_n$ is an estimator of the asymptotic covariance matrix $\mathbf{V}_n = \sigma^2 \mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}(\mathbf{D}_n^{(1)})'$ of $\mathbf{T}_n$ given by the oracle limit distribution; cf. Theorem 3.1. Note that this studentized version of $\mathbf{T}_n$ can be computationally highly demanding, particularly for repeated bootstrap computation, when p_0 is large. In comparison, the proposed studentized version of $\mathbf{T}_n$ that we consider here is based only on a scalar factor and hence, computationally simpler.

The following result gives the rate of bootstrap approximation to the distribution of $\mathbf{R}_n$. For notational compactness, in the rest of this section, we shall write (C.1)′–(C.6)′, to denote conditions (C.1)–(C.6), when (C.3) and (C.6) are defined with part (ii)′ instead of part (ii).

THEOREM 4.2. *If conditions* (C.1)′–(C.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 4.2 shows that under conditions (C.1)′–(C.6)′, the bootstrap approximation to the distribution of $\mathbf{R}_n$ is second-order-correct, as it corrects for the effects of the leading terms in the EE of $\mathbf{R}_n$. From the proof of Theorem 7.2, it follows that the bootstrap not only captures the usual $O(n^{-1/2})$ term in the EE, but it also corrects for the effects of the second and the third terms in the upper bound of Theorem 3.1 that result from the penalization step in the definition of the ALASSO. The accuracy level $o_p(n^{-1/2})$ for the bootstrap holds even when the actual magnitudes of these terms are coarser than $n^{-1/2}$ which, in turn, leads to a poor rate of approximation by the limiting normal distribution. A practical implication of this result is that percentile-t bootstrap CIs based on $\mathbf{R}_n$ will be more accurate than the CIs based on the large sample normal critical points. Indeed, the finite sample simulation results presented in Section 6 show that the CIs based on normal critical points are practically useless in moderate samples and improvements in the coverage accuracy achieved by the bootstrap CIs based on $\mathbf{R}_n$ are spectacular.

4.4. *A modified pivot and higher order correctness.* Although the residual bootstrap approximation for the studentized statistic $\mathbf{R}_n$ is second order correct, a more careful analysis shows that it may fail to achieve the same *optimal* rate, namely, $O_p(n^{-1})$ as in the traditional fixed and finite dimensional regression problems. The main reason behind this is the effect of the bias term $\|\mathbf{b}_n\|$ in Theorem 3.1, which can be coarser than $n^{-1/2}$. While the second order correctness is a desirable property for the one-sided CIs, the higher level of accuracy, namely $O_p(n^{-1})$, is important for two-sided CIs; cf. Hall (1992). To that end, we now define a modified pivotal quantity

$$\breve{\mathbf{R}}_n = \frac{\sqrt{n}\mathbf{D}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) + \breve{\mathbf{b}}_n}{\breve{\sigma}_n}, \tag{4.2}$$

where $\breve{\mathbf{b}}_n = \breve{\mathbf{D}}_n^{(1)} \breve{\mathbf{C}}_{11,n}^{-1} \breve{\mathbf{s}}_n^{(1)} \cdot \frac{\lambda_n}{\sqrt{n}}$, $\breve{\mathbf{D}}_n^{(1)}$ and $\breve{\mathbf{C}}_{11,n}^{(1)}$ are, respectively, $q \times |\widehat{I}_n|$ and $|\widehat{I}_n| \times |\widehat{I}_n|$ submatrices of $\mathbf{D}_n$ and $\mathbf{C}_n$ with columns (and also rows, in case of $\breve{\mathbf{C}}_{11,n}$) in $\widehat{I}_n = \{j : 1 \leq j \leq p, \widehat{\beta}_{j,n} \neq 0\}$, and similarly, $\breve{\mathbf{s}}_n^{(1)}$ is the $|\widehat{I}_n| \times 1$ vector with jth element $\operatorname{sgn}(\widehat{\beta}_{j,n})|\tilde{\beta}_{j,n}|^{-\gamma}$, $j \in \widehat{I}_n$. Here $\breve{\sigma}_n^2$ is defined as

$$\breve{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (\breve{\varepsilon}_i - \bar{\breve{\varepsilon}}_n)^2,$$

where $\breve{\varepsilon}_i = y_i - \mathbf{x}_i' \breve{\boldsymbol{\beta}}_n$, and$\breve{\beta}_{j,n} = \tilde{\beta}_{j,n} \cdot \mathbf{1}(j \in \widehat{I}_n)$, $1 \leq j \leq p$. Note that $\breve{\mathbf{R}}_n$ is obtained by applying a specially designed bias-correction term to $\mathbf{T}_n$ and by a suitable rescaling, which are suggested by the form of the third order EE of Theorem 7.2. Also, it is interesting to note that for both of these estimators, we only use the sub-vectors of the design vectors $\mathbf{x}_i$'s and components of the initial estimator that correspond to the (random) set of variables selected by the ALASSO. Next, define $\breve{\mathbf{R}}_n^*$, the bootstrap version of $\breve{\mathbf{R}}_n$, by replacing $\{y_1, \dots, y_n\}$ and $\boldsymbol{\beta}$ by $\{y_1^*, \dots, y_n^*\}$ and $\widehat{\boldsymbol{\beta}}_n$, respectively. Then we have the following result:

THEOREM 4.3. *If conditions* (C.1)′–(C.6)′ *hold with* $r = 8$, *then*

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

Theorem 4.3 asserts that under appropriate regularity conditions, the rate of bootstrap approximation to the modified pivotal quantity $\breve{\mathbf{R}}_n$ attains the the "optimal" level of accuracy irrespective of the magnitude of $\|\mathbf{b}_n\|$. An immediate consequence of this result is that symmetric bootstrap confidence regions based on the modified pivot attains the higher rate $O(n^{-1})$ of convergence accuracy even when the magnitude of $\|\mathbf{b}_n\|$ is coarser than $n^{-1/2}$. As explained in Remark 1, the coarser magnitude of $\|\mathbf{b}_n\|$ can occur quite naturally in a variety of situations whenever a combination of values of the underlying regression parameters, the design matrix and the choice of the penalty constant satisfy (3.1). In such cases, bootstrap CIs based on $\breve{\mathbf{R}}_n$ gives a marked improvement over normal critical points based CIs where the accuracy is sub-$O(n^{-1/2})$ for both one- and two-sided CIs.

5. Results for the $p > n$ case. In many applications, p is much larger than n, and post variable selection inference on the regression parameters is an even more challenging problem. In this section, we study properties of the bootstrap approximation to the studentized ALASSO estimator in the $p > n$ case. Note that for $p > n$, the $p \times p$ matrix $n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'$ is always singular and hence the OLS of $\boldsymbol{\beta}_n$ is no longer uniquely defined. In the literature, a popular choice of the initial root-n consistent estimator $\widetilde{\boldsymbol{\beta}}_n$ for $p > n$ is the LASSO estimator, although other bridge estimators of $\boldsymbol{\beta}_n$ [cf. Knight and Fu (2000)] can also be used. Let $\widehat{\boldsymbol{\beta}}_n$ be the ALASSO estimator defined by (1.2), with a root-n consistent initial estimator $\widetilde{\boldsymbol{\beta}}_n$.

Also define the studentized version of $\hat{\boldsymbol{\beta}}_n$ (cf. Section 4.3) by $\mathbf{R}_n = \widehat{\sigma}_n^{-1}\mathbf{T}_n$ where $\widehat{\sigma}_n^2$ is the average of squared centered residuals $\breve{e}_1, \ldots, \breve{e}_n$, from the ALASSO fit, and define the bias corrected version $\breve{\mathbf{R}}_n$ as in (4.2).

To prove the results in the $p > n$ case, we need the following condition:

(C.7) There exists $K \in (0, \infty)$ such that

$$\begin{aligned} &\mathbf{P}\Big(\max_{1\le j\le p}\big|\sqrt{n}(\widetilde{\beta}_{j,n} - \beta_{j,n})\big| > K\sqrt{\log n}\Big) = o\big(n^{-1/2}\big), \\ &\mathbf{P}_*\Big(\max_{1\le j\le p}\big|\sqrt{n}(\widetilde{\beta}^*_{j,n} - \widehat{\beta}_{j,n})\big| > K\sqrt{\log n}\Big) = o_p\big(n^{-1/2}\big). \end{aligned} \tag{5.1}$$

We also need the following modified version of (C.2)(ii):

(C.2)(ii)$'$

$$\max_{1\le j\le p}\left\{n^{-1}\sum_{i=1}^{n}|x_{i,j}|^r\right\} + \max_{1\le j\le p_0}\{c_{11,n}^{j,j}\} = O(1),$$

where $c_{11,n}^{j,j}$ is the (j, j)th element of $\mathbf{C}_{11,n}^{-1}$.

We now briefly discuss the conditions. Condition (C.7) is a high-level condition that requires the initial estimator $\widetilde{\boldsymbol{\beta}}_n$ and its bootstrap version not only to be $\sqrt{n}$-consistent, but also to satisfy a suitable form of moderate deviation bound. For estimators $\widetilde{\boldsymbol{\beta}}_n$, such that $\sqrt{n}(\widetilde{\beta}_{j,n} - \beta_{j,n})$ can be closely approximated by $\sum_{i=1}^n h_{j,i,n}\varepsilon_i$ for some $\{h_{j,i,n}\} \subset \mathbb{R}$ with $\sum_{i=1}^n h_{j,i,n}^2 = O(1)$, (C.7) holds if $\mathbf{E}\varepsilon_1^4 < \infty$ and $\sum_{i=1}^n h_{j,i,n}^4 = o(n^{-1/2})$. See Proposition 8.4 [Chatterjee and Lahiri (2013)] for an example. Condition (C.2)(ii)$'$ drops the condition $\max\{n^{-1}\sum_{i=1}^n |\tilde{x}_{i,j}|^r : 1 \le j \le p\} = O(1)$, in (C.2)(ii), which can no longer hold in the $p > n$ case, as $\mathbf{C}_n^{-1}$ does not exist. Instead, it requires existence of $\mathbf{C}_{11,n}^{-1}$, which is of dimension $p_0 \times p_0$. Thus, we must have $p_0 \le n$ (in addition to other conditions) for the validity of the results in the $p > n$ case.

Let $\mathbf{R}_n^*$ and $\breve{\mathbf{R}}_n^*$ denote the (residual) bootstrap versions of $\mathbf{R}_n$ and $\breve{\mathbf{R}}_n$, respectively. Then, we have the following result:

THEOREM 5.1. *Suppose that $p > n$ and conditions* (C.1), (C.2)(i), (C.2)(ii)$'$, (C.3)–(C.7) *hold with $b = 0$. Then*

$$\sup_{B\in\mathcal{C}_q}\big|\mathbf{P}(\mathbf{R}_n \in B) - \mathbf{P}_*\big(\mathbf{R}_n^* \in B\big)\big| = o_p\big(n^{-1/2}\big) \quad \textit{and}$$

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

Thus, under the conditions of Theorem 5.1, the bootstrap approximations based on the pivots $\mathbf{R}_n$ and $\breve{\mathbf{R}}_n$ are both second-order accurate, even in the case where $p > n$. In comparison, the oracle based normal approximation admits the sub-optimal bounds of Section 3, and therefore, it is significantly less accurate than the bootstrap approximations. This conclusion is also supported by the finite sample simulation results of Section 6 for the $p > n$ cases considered therein.

REMARK 2. Note that in Theorem 5.1, the bound on the accuracy of the bootstrap approximations to $\breve{\mathbf{R}}_n$ is just $o_p(n^{-1/2})$ for the $p > n$ case. This is not as precise as the bound in the $p \leq n$ case where it is $O_p(n^{-1})$. It would be possible to derive a similar bound for the $p > n$ case for $\breve{\mathbf{R}}_n$ if we are willing to make some strong additional assumptions on the initial estimator [e.g., existence of an EE for the joint distribution of $\mathbf{T}_n, n^{-1}\sum_{i=1}^{n}(\varepsilon_i^k - \mathbf{E}\varepsilon_i^k)$, with $k = 1, 2$ and suitable linear combinations of $\sqrt{n}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$, which are not known at this stage]. As a result, we do not pursue such refinements here.

REMARK 3. Although we do not explicitly impose any growth conditions on p as a function of n, there is, however, an implicit requirement through condition (C.7). Indeed, if the leading terms in $\sqrt{n}(\tilde{\beta}_{j,n} - \beta_{j,n})$ can be expressed as $\sum_{i=1}^{n} h_{ji,n}\varepsilon_i$ for some $h_{j1,n}, \dots, h_{jn,n} \in \mathbb{R}$ with $\sum_{i=1}^{n} h_{ji,n}^2 = O(1)$, then for (C.7) to hold, arguments in the proof of Lemma 7.1(iii) require that, for some integer $r \geq 3$, $\mathbf{E}|\varepsilon_1|^r < \infty$ and $p \cdot n^{-(r-2)/2} = o(n^{-1/2})$. This implies that p can grow at a polynomial rate $p \sim Kn^a$, for some $K > 0$ and $a > 1$, provided $\mathbf{E}|\varepsilon_1|^r < \infty$ for some $r > 2a + 3$. Thus, the allowable growth rate of p depends on the lightness of the tails of the error distribution.

REMARK 4. As pointed out by a referee, the use of $\widetilde{\boldsymbol{\beta}}_n$ in place of $\widetilde{\boldsymbol{\beta}}_n^*$ in the bootstrap computation of the ALASSO estimator in (4.1) will yield a computationally more efficient algorithm. It can be shown that with this modification, conclusions of Theorems 4.2, 4.3 and 5.1 remain valid, with the error bound $o_p(n^{-1/2})$ only.

6. Simulation results. In this section we study the finite sample performance of the proposed bootstrap methods. The following cases corresponding to different choices of $\boldsymbol{\beta}_n$ were studied:

(a) $(n, p) = (60, 10)$: with $p_0 = 5$ and $\boldsymbol{\beta}_n = (4, -1.5, -8, 0.9, -3, 0, \dots, 0)'$.

(b) $(n, p) = (60, 100)$: with $p_0 = 5$ and $\boldsymbol{\beta}_n$ same as in case (a) above, except that last 95 components are zeros.

(c) $(n, p) = (200, 80)$: with $p_0 = 10$ and with the last 70 components being zeros,

$$\boldsymbol{\beta}_n = (4, 2.5, 0.8, -1.5, -2, -5, -7.5, 5, 1.5, -3, 0, \dots, 0)'.$$

(d) $(n, p) = (200, 500)$: with $p_0 = 10$ and $\boldsymbol{\beta}_n$ same as in case (c) above, except that the last 490 components are zeros.

TABLE 1
Comparison of empirical coverage probabilities and average lengths (in parentheses) for 90% *CIs for the underlying parameter* $\beta_1 (= 4)$ *in cases* (a)–(d). *In all cases* $\lambda_{2,n} = 2n^{1/4}$ *and in cases* (b) *and* (d), $\lambda_{1,n} = 0.5n^{1/2}$

	One-sided			Two-sided (with average lengths)		
Case	$\mathbf{R}_n$	$\check{\mathbf{R}}_n$	**Oracle**	$\mathbf{R}_n$	$\check{\mathbf{R}}_n$	**Oracle**
(a)	0.898	0.904	0.668	0.918 (0.407)	0.900 (0.392)	0.158 (0.05)
(b)	0.894	0.930	0.740	0.894 (0.536)	0.894 (0.530)	0.154 (0.064)
(c)	0.912	0.844	0.518	0.928 (0.252)	0.994 (0.247)	0.064 (0.017)
(d)	0.892	0.878	0.622	0.880 (0.253)	0.890 (0.261)	0.098 (0.017)

Cases (b) and (d) correspond to the $p > n$ case. In all cases, the design vectors $(x_{i,1}, \dots, x_{i,p_0})'$ are independently generated from a normal population with mean $\mathbf{0}$ and covariance matrix $((\eta_{i,j}))$ with $\eta_{i,j} = (0.3)^{|i-j|}$ and the remaining $(p - p_0)$ covariates are i.i.d. $N(0, 1)$. The errors $\{\varepsilon_i\}$ are i.i.d. $N(0, 1)$. We fix $\gamma = 1$. In the high-dimensional case, since there is no unique least squares estimator, we have used the LASSO estimator as the initial estimator $\widetilde{\boldsymbol{\beta}}_n$, with associated tuning parameter $\lambda_{1,n}$. In the ALASSO step, the penalty parameter is $\lambda_{2,n}$ and to avoid division by zero, we used weights $(|\widetilde{\beta}_{j,n}| + a_n)^{-1}$ with $a_n = n^{-1/2}$, to define the weighted ℓ_1 penalty in (1.2).

6.1. *Comparison of oracle based normal CIs and bootstrap CIs.* As suggested from Table 1, in all cases when the underlying true parameter value is large enough, the bootstrap based CIs clearly superior to the oracle based method. For moderately small underlying true parameters, results in Table 2 suggest that the bootstrap-based methods are still better than the Oracle method for both one and two-sided CIs, even when $p > n$. The improvement is most significant for the 2-sided CIs.

6.2. *Comparison with a perturbation based method.* In the $p \leq n$ case, Minnier, Tian and Cai (2011) suggested a perturbation-based approach for construction of CIs of underlying regression parameters, including the zero parameters. We compare the performance of our proposed bootstrap-based method with their approach. We use ($n = 100$, $p = 10$). The design vectors $\mathbf{x}_i$ are independently selected from a normal population with mean $\mathbf{0}$, unit variances and pairwise covariances equal to 0.2. The errors ε_i are i.i.d. $N(0, \sigma^2)$. We considered two choices,

TABLE 2

Comparison of empirical coverage probabilities and average lengths (in parentheses) for 90% CIs for the underlying parameter $\beta_4 (= 0.9)$ in cases (a) and (b). In both cases $\lambda_{2,n} = 2n^{1/4}$ and in case (b), $\lambda_{1,n} = 0.5n^{1/2}$

	One-sided			Two-sided (with average lengths)		
Case	$\mathbf{R}_n$	$\breve{\mathbf{R}}_n$	**Oracle**	$\mathbf{R}_n$	$\breve{\mathbf{R}}_n$	**Oracle**
(a)	0.868	0.946	0.840	0.902 (0.598)	0.944 (0.529)	0.086 (0.061)
(b)	0.908	0.944	0.904	0.886 (0.607)	0.942 (0.652)	0.072 (0.058)

$\sigma = 1$ and 5. The true regression parameter is $\boldsymbol{\beta} = (2, -2, 0.5, -0.5, 0, \dots, 0)'$. This is very similar to the setup used in Minnier, Tian and Cai (2011). Among the different types of CIs they proposed, we focus on (i) the usual normal type CI (which has been modified by a thresholding approach to handle underlying zero parameters) and denoted by $\mathrm{CR}^{*\mathrm{N}}$ and (ii) CIs directly based on the quantiles of the perturbed regression estimates, denoted by $\mathrm{CR}^{*\mathrm{Q}}$. As suggested in their paper, we used a BIC-based choice for $\lambda_{2,n}$ for the simulations; cf. Minnier, Tian and Cai (2011).

As shown in Table 3 and somewhat contrary to the findings of Minnier, Tian and Cai (2011), we found that the $\mathrm{CR}^{*\mathrm{N}}$ based CIs have poor coverage for both zero and nonzero regression parameters. However, the $\mathrm{CR}^{*\mathrm{Q}}$ method performs much better, particularly when the error variance is high. In comparison, the bootstrap-based methods are uniformly superior in all cases. We also noted that compared to the the $\mathrm{CR}^{*\mathrm{Q}}$ method, the coverage accuracy of the bootstrap CIs is more sensitive to the choice of the smoothing parameter for the zero parameters; see Section 6.3 below.

TABLE 3

Comparison of empirical coverage probabilities for 90% two-sided CIs using the perturbation based approach by Minnier, Tian and Cai (2011), the oracle and the bootstrap based methods. For the Oracle and Bootstrap methods, the penalty parameter is $\lambda_{2,n} = 0.5 \cdot n^{1/4}$ and for the perturbation based approach the BIC based choice of $\lambda_{2,n}$ was used

		Perturbation			Bootstrap	
Parameter	σ	$\mathbf{CR}^{*\mathbf{N}}$	$\mathbf{CR}^{*\mathbf{Q}}$	**Oracle**	$\mathbf{R}_n$	$\breve{\mathbf{R}}_n$
$\beta_1 = 4$	1	0.012	0.306	0.132	0.916	0.898
	5	0.122	0.876	0.124	0.916	0.914
$\beta_5 = 0$	1	1.0	1.0	0	0.894	0.936
	5	0.288	0.902	0	0.932	0.918

TABLE 4
Comparison of empirical coverage probabilities for 90% CIs for different parameters, using CV based and theoretical choices of $\lambda_{2,n}$ in case (a). The optimal CV based $\lambda_{2,n} = 0.049 \approx 0.017 \cdot 60^{1/4}$. For the zero parameter case an additional (theoretical) choice of $\lambda_{2,n} = 0.25 * n^{1/4}$ is compared

		One-sided			Two-sided		
Parameter	**Method**	$\mathbf{R}_n$	$\breve{\mathbf{R}}_n$	**Oracle**	$\mathbf{R}_n$	$\breve{\mathbf{R}}_n$	**Oracle**
$\beta_1 = 4$	CV	0.892	0.894	0.588	0.938	0.890	0.162
	Th.	0.894	0.898	0.668	0.922	0.894	0.158
$\beta_4 = 0.9$	CV	0.882	0.882	0.566	0.924	0.882	0.156
	Th.	0.872	0.944	0.840	0.940	0.864	0.138
$\beta_6 = 0$	CV	0.888	0.886	0.428	0.942	0.902	0
	Th.	0.004	0.004	0.004	0	0	0
	Th.[a]	0.896	0.850	0.180	0.944	0.884	0

[a]At $\lambda_{2,n} = 0.25 * n^{1/4}$.

6.3. *Choice of tuning parameter.* For penalized regression techniques, the cross validation (CV) has been a popular method for choosing the tuning parameters, in both low and high-dimensional cases. We compare the performance of cross validation (CV) based and theoretical choices of tuning parameters. Based on the theoretical rates, we use $\lambda_{2,n} = 2n^{1/4}$ (for the ALASSO stage) and in the $p > n$ case, the tuning parameter $\lambda_{1,n}$, used for the LASSO stage, is set at $\lambda_{1,n} = 0.5n^{1/2}$. When using CV, the initial tuning parameter $\lambda_{1,n}$ is selected by 5-fold CV (only in the $p > n$ case) and kept fixed. Using this fixed value and again using 5-fold CV, the tuning parameter $\lambda_{2,n}$ for the ALASSO stage is selected. When the underlying true parameter is zero, an additional theoretical choice of $\lambda_{2,n} = 0.25 \cdot n^{1/4}$ is used for comparison.

As seen from Table 4, in case (a) (with $p < n$), using the CV-based choice of $\lambda_{2,n}$ leads to very good empirical coverage probabilities for all choices of underlying regression parameters, including zero parameters. The theoretical choice also performs comparably for all parameters, except the zero parameter case, where a smaller value of $\lambda_{2,n}$ performs comparably. The results in Table 5, for case (b) (in the $p > n$ setup), show that there is an overall decrease in the empirical coverage probabilities for both choices. Unlike the results in case (a) (cf. Table 4), the performance is very poor for the zero parameters irrespective of the method used for selecting the tuning parameters.

6.4. *Real data analysis for the low dimensional case.* In this section we apply the bootstrap based methods on a prostrate cancer data-set, available from a clinical study and used in Tibshirani (1996) [originally available from Stamey et al. (1989)]. In this clinical study, a total of $n = 97$ observations were available and

TABLE 5
Comparison of empirical coverage probabilities for 90% CIs for different parameters, using CV based and theoretical choices of $\lambda_{1,n}$ and $\lambda_{2,n}$ in case (b). The optimal CV based choices were $\lambda_{1,n} = 0.124 \approx 0.016 \cdot (60)^{1/2}$ and $\lambda_{2,n} = 0.639 \approx 0.229 \cdot (60)^{1/4}$

		One-sided			Two-sided		
Parameter	**Method**	$\mathbf{R}_n$	$\breve{\mathbf{R}}_n$	**Oracle**	$\mathbf{R}_n$	$\breve{\mathbf{R}}_n$	**Oracle**
$\beta_1 = 4$	CV	0.81	0.838	0.730	0.636	0.506	0.104
	Th.	0.894	0.930	0.740	0.894	0.894	0.154
$\beta_4 = 0.9$	CV	0.798	0.854	0.748	0.656	0.488	0.104
	Th.	0.908	0.944	0.904	0.886	0.942	0.072
$\beta_6 = 0$	CV	0.384	0.398	0.194	0.216	0.116	0.00
	Th.	0.016	0.016	0.016	0	0	0
	Th.[a]	0.348	0.332	0.176	0.224	0.112	0

[a]At $\lambda_{2,n} = 0.25 * n^{1/4}$.

the variable of interest was log(prostrate specific antigen) (`lpsa`) and eight different predictors ($p = 8$) were used to study the behavior of this quantity. The predictors were log(cancer volume) (`lcavol`), log(prostrate weight) (`lweight`), `age`, log(benign prostratic hyperplasia amount) (`lbph`), seminal vesicle invasion (`svi`), log(capsular penetration) (`lcp`), Gleason score (`gleason`) and percentage Gleason scores 4 or 5 (`pgg45`). The columns of the design matrix are centered and scaled to have unit norm. We use the following theoretical choice for the penalty parameter: $\lambda_{2,n} = n^{1/4}$. Table 6 shows CIs for estimated nonzero coefficients. Note that in more than one instance, the estimated values of $\beta_{j,n}$ fall outside the bootstrap CIs. This can be explained by considering that the histograms of the bootstrap replicates which showed that the distributions of $\mathbf{R}_n^*$ and $\breve{\mathbf{R}}_n^*$ are heavily skewed and far from the oracle normal distribution. This is reflected by the endpoints of the corresponding CIs in Table 6.

TABLE 6
Analysis of prostrate cancer data from Tibshirani (1996). The penalty parameter used is $\lambda_2 = n^{1/4}$. ALASSO estimates and resultant 90% two-sided CIs for estimated nonzero components are shown

Predictor (j)	$\hat{\beta}_{j,n}$	$\mathbf{R}_n$	$\breve{\mathbf{R}}_n$	**Oracle**
`lcavol`	0.688	(0.520, 0.822)	(0.616, 0.944)	(0.636, 0.741)
`lweight`	0.112	(0.140, 0.235)	(0.162, 0.395)	(0.067, 0.156)
`svi`	0.167	(0.138, 0.352)	(0.178, 0.487)	(0.115, 0.219)

*Obtained from http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/prostate.data.

TABLE 7
Analysis of microarray data with $n = 30$ and $p = 545$ (after initial screening step). All six predictors with nonzero ALASSO coefficients and corresponding 90% two-sided CIs based on the bootstrap and oracle methods

[a]**Predictor (j)**	$\hat{\beta}_{j,n}$	$\mathbf{R}_n$	$\check{\mathbf{R}}_n$	**Oracle**
G709	−0.066	(−0.146, −0.120)	(−0.490, −0.331)	(−0.127, −0.005)
G2272	0.095	(0.087, 0.207)	(0.376, 0.619)	(0.010, 0.180)
G3655	0.475	(0.250, 0.759)	(0.749, 1.309)	(0.375, 0.575)
G4322	−0.021	(−0.047, −0.041)	(−0.443, −0.432)	(−0.091, 0.048)
G5904	0.240	(0.161, 0.507)	(0.495, 0.900)	(0.168, 0.311)
G6252	0.112	(0.029, 0.241)	(0.414, 0.687)	(0.030, 0.193)

[a]Data available from supplementary material of Hall and Miller (2009).

6.5. *Real data analysis for the high-dimensional case*. The data, available from a microarray experiment was collected from Hall and Miller (2009) and originally used in Segal, Dahlquist and Conklin (2003). The data consisted of observations from $n = 30$ specimens on the Ro1 expression level (y), and genetic expression levels $\mathbf{x} = (x_1, \ldots, x_p)'$ for 6319 genes. The absolute value of the correlation between y and each covariate x_i was used as an initial screening tool and only those covariates with absolute correlation value ≥ 0.5 were selected for further study. This resulted in a smaller set of $p = 545$ covariates. The columns of the design matrix were centered and scaled (by the columnwise standard deviation) and the response vector $\mathbf{y}$ was also transformed by centering and scaling. The selected tuning parameters were $\lambda_1 = 0.5 \cdot n^{1/2}$ and $\lambda_2 = 0.5 \cdot n^{1/4}$. After the initial LASSO step, twenty covariates are selected and after the ALASSO step only six covariates (genes) were selected (shown in Table 7). The residual sum of squares divided by (n-number of nonzero parameters) provides the following: for the initial LASSO estimate 0.1082 (equivalent to a R^2 value of 0.888) and for the ALASSO estimate we obtain 0.092 (equivalent to $R^2 = 0.904$). This suggests that the extra 14 variables, present in the LASSO estimator provide very little information about the response. Note that here also the estimated values of $\beta_{j,n}$'s often fall outside the bootstrap CIs based on the bias corrected pivot $\check{\mathbf{R}}_n$. This suggests that the true values of the nonzero parameters are probably much larger in absolute value than suggested by their ALASSO point estimates.

7. Proofs.

7.1. *Notation*. For notational simplicity, we shall set $p_n = p$, $p_{0,n} = p_0$. Let $\mathbb{Z}_+ = \{0, 1, \ldots\}$. Let $K, K(\cdot) \in (0, \infty)$ denote generic constants not depending on their arguments (if any), but not on n. Also, in the proofs below, let $n_0 \geq 1$ denotes a generic (large) integer. For $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r) \in \mathbb{Z}_+^r$, let $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_p$, $\boldsymbol{\alpha}! =$

$\alpha_1! \cdots \alpha_r!$ and let $D^{\boldsymbol{\alpha}}$ denote the differential operator $\frac{\partial^{|\boldsymbol{\alpha}|}}{\partial x_1^{\alpha_1} \cdots \partial x_r^{\alpha_r}}$ on $\mathbb{R}^r$, where $r \geq 1$ is an integer. Let $\mathbf{W}_n = n^{-1/2} \sum_{i=1}^n \mathbf{x}_i' \varepsilon_i$. Partition $\mathbf{W}_n$ as $\mathbf{W}_n = (\mathbf{W}_n^{(1)'}, \mathbf{W}_n^{(2)'})'$, where $\mathbf{W}_n^{(1)}$ is $p_0 \times 1$. Also, set $\mathbf{W}_n^{(0)} = \mathbf{W}_n$, $p^{(0)} = p$, $p^{(1)} = p_0$ and $p^{(2)} = p - p_0$. Let $\mathbf{b}_n = \mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} \mathbf{s}_n^{(1)} \cdot \lambda_n n^{-1/2}$, $\Upsilon_n = n^{-1} \sum_{i=1}^n \boldsymbol{\xi}_i^0 (\boldsymbol{\xi}_i^0)'$ and $\breve{\Upsilon}_n = n^{-1} \sum_{i=1}^n (\boldsymbol{\xi}_i^0 + \boldsymbol{\eta}_i^{(0)})(\boldsymbol{\xi}_i^0 + \boldsymbol{\eta}_i^{(0)})'$, where $\boldsymbol{\xi}_i^{(0)} = \mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} \mathbf{x}_i^{(1)}$, $\boldsymbol{\eta}_i^{(0)} = \mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} \boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_i = (\xi_{i,1}, \ldots, \xi_{i,p_0})$ with $\xi_{i,j} = -\frac{\lambda_n}{n^{1/2}} \cdot \tilde{x}_{i,j} \cdot \operatorname{sgn}(\beta_{j,n}) \gamma |\beta_{j,n}|^{-(\gamma+1)}$, $1 \leq j \leq p_0$. Next note that by conditions (C.2), (C.3) and (C.6),

$$\|\mathbf{b}_n\| \leq \|\mathbf{D}^{(1)} \mathbf{C}_{11,n}^{-1/2}\| \cdot \|\mathbf{C}_{11,n}^{-1/2}\| \cdot \|\mathbf{s}_n^{(1)}\| \cdot \frac{\lambda_n}{\sqrt{n}} = O(n^{-\delta}).$$

Let $r_1 = \min\{r \geq 1 : \|\mathbf{b}_n\|^{r+1} = o(n^{-1/2})\}$. Define the Lebesgue density of the EE for $\mathbf{T}_n$ by

$$\psi_n(\mathbf{x}) = \phi(\mathbf{x}, \sigma^2 \breve{\Upsilon}_n) \Bigg[1 + \sum_{|\boldsymbol{\alpha}|=1}^{r_1} \mathbf{b}_n^{\boldsymbol{\alpha}} \chi_{\boldsymbol{\alpha}}(\mathbf{x}; \sigma^2 \breve{\Upsilon}_n) + \frac{\mu_3}{6\sqrt{n}} \sum_{|\boldsymbol{\alpha}|=3} \bar{\boldsymbol{\xi}}_n^{(0)}(\boldsymbol{\alpha}) \chi_{\boldsymbol{\alpha}}(\mathbf{x}; \sigma^2 \breve{\Upsilon}_n) \Bigg], \qquad \mathbf{x} \in \mathbb{R}^q,$$

where $\bar{\boldsymbol{\xi}}_n^{(0)}(\boldsymbol{\alpha}) = n^{-1} \sum_{i=1}^n (\boldsymbol{\xi}_i^{(0)})^{\boldsymbol{\alpha}}$, $\phi(\mathbf{x}, \Upsilon)$ denotes the density of the $N(\mathbf{0}, \Upsilon)$ distribution on $\mathbb{R}^q$ and where $\chi_{\boldsymbol{\alpha}}(\mathbf{x}; \Upsilon)$ is defined by the identity

$$\chi_{\boldsymbol{\alpha}}(\mathbf{x}; \Upsilon) \phi(\mathbf{x}; \Upsilon) = (-D)^{\boldsymbol{\alpha}} \phi(\mathbf{x}; \Upsilon), \qquad \boldsymbol{\alpha} \in \mathbb{Z}_+^q.$$

Next define the density of the EE for $\mathbf{R}_n$ by

$$\begin{aligned} \pi_n(\mathbf{x}) = \phi(\mathbf{x}, \breve{\Upsilon}_n) \Bigg[1 &+ \sum_{k=1}^{r_1} \frac{1}{k!} \Bigg\{ \sum_{|\boldsymbol{\alpha}|=k} (-\mathbf{b}_n)^{\boldsymbol{\alpha}} \chi_{\boldsymbol{\alpha}}(\mathbf{x} : \breve{\Upsilon}_n) \Bigg\} \\ &+ \frac{1}{\sqrt{n}} \cdot \frac{\mu_3}{6\sigma^3} \Bigg\{ \sum_{|\boldsymbol{\alpha}|=1} \sum_{|\boldsymbol{\gamma}|=2} [\bar{\boldsymbol{\xi}}_n^{(0)}(\boldsymbol{\alpha} + \boldsymbol{\gamma}) - 3 \bar{\boldsymbol{\xi}}_n^{(0)}(\boldsymbol{\alpha}) \bar{\boldsymbol{\xi}}_n^{(0)}(\boldsymbol{\gamma})] \\ &\times \chi_{\boldsymbol{\alpha}+\boldsymbol{\gamma}}(\mathbf{x}; \breve{\Upsilon}_n) - 3 \sum_{|\boldsymbol{\alpha}|=1} \bar{\boldsymbol{\xi}}_n^{(0)}(\boldsymbol{\alpha}) \chi_{\boldsymbol{\alpha}}(\mathbf{x}; \breve{\Upsilon}_n) \Bigg\} \Bigg], \\ &\mathbf{x} \in \mathbb{R}^q. \end{aligned}$$

7.2. *Auxiliary results.*

LEMMA 7.1. *Under* (C.2) *and* (C.4):

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

PROOF. See the supplementary material Chatterjee and Lahiri (2013) (hereafter referred to as [CL]). □

The key step in the proofs of Theorems 3.1–5.1 is EEs for the ALASSO estimator and its studentized version which are given below.

THEOREM 7.2. (a) *If conditions* (C.1)–(C.6) *hold with* $r = 4$, *then*

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

(b) *If conditions* (C.1)′–(C.6)′ *hold with* $r = 6$, *then*

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

PROOF. See [CL]. □

7.3. *Proof of the main results*.

PROOF OF THEOREM 3.1. We only give an outline of the proof here. For the details of the steps, see [CL]. Let $\boldsymbol{\Lambda}_n^{(1)}$ be a $p_0 \times p_0$ diagonal matrix with jth diagonal entry given by $\operatorname{sgn}(\beta_{j,n})|\beta_{j,n}|^{-(\gamma+1)}$, $1 \le j \le p_0$. Then it can be shown that

$$n^{-1}\sum_{i=1}^{n} \boldsymbol{\xi}_i^{(0)}\boldsymbol{\eta}_i^{(0)\prime} = -\frac{\lambda_n\gamma}{n}\mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}\boldsymbol{\Lambda}_n^{(1)}\mathbf{C}_{11,n}^{-1}\mathbf{D}_n^{(1)\prime}. \tag{7.1}$$

Using Theorem 7.2(a), one gets

$$\begin{aligned}
\Delta_n &\equiv \sup_{B\in\mathcal{C}_q}\left|\mathbf{P}(\mathbf{T}_n \in B) - \int_B \phi(\mathbf{x};\sigma^2\Upsilon_n)\,d\mathbf{x}\right| \\
&= \sup_{B\in\mathcal{C}_q}\left|\int_B [\phi(\mathbf{x};\sigma^2\breve{\Upsilon}_n) - \phi(\mathbf{x};\sigma^2\Upsilon_n)]\,d\mathbf{x}\right. \\
&\qquad + \sum_{|\boldsymbol{\alpha}|=1}\mathbf{b}_n^{\boldsymbol{\alpha}}\int_B \chi_{\boldsymbol{\alpha}}(\mathbf{x};\sigma^2\breve{\Upsilon}_n)\phi(\mathbf{x};\sigma^2\breve{\Upsilon}_n)\,d\mathbf{x} \\
&\qquad \left. + \frac{\mu_3}{6\sqrt{n}}\sum_{|\boldsymbol{\alpha}|=3}\bar{\boldsymbol{\xi}}_n^{(0)}(\boldsymbol{\alpha})\int_B \chi_{\boldsymbol{\alpha}}(\mathbf{x};\sigma^2\breve{\Upsilon}_n)\phi(\mathbf{x};\sigma^2\breve{\Upsilon}_n)\,d\mathbf{x}\right|
\end{aligned} \tag{7.2}$$

$$+ o(n^{-1/2} + \|\mathbf{b}_n\|)$$

$$\equiv \sup_{B \in \mathcal{C}_q} |I_{1,n}(B) + I_{2,n}(B) + I_{3,n}(B)| + o(n^{-1/2} + \|\mathbf{b}_n\|).$$

Also, by conditions (C.2)–(C.6),

$$\|\breve{\Upsilon}_n - \Upsilon_n\| = \left\| 2n^{-1} \sum_{i=1}^{n} \boldsymbol{\xi}_i^{(0)} \boldsymbol{\eta}_i^{(0)\prime} + n^{-1} \sum_{i=1}^{n} \boldsymbol{\eta}_i^{(0)} \boldsymbol{\eta}_i^{(0)\prime} \right\|$$
$$\leq K(q, \gamma) \cdot \frac{\lambda_n}{n} \cdot n^{a+b(\gamma+1)}. \tag{7.3}$$

The proof of Theorem 3.1 now follows from (7.1)–(7.3); See [CL]. □

PROOF OF THEOREM 3.2. Since $\mathrm{tr}(\boldsymbol{\Gamma}_n) \geq \delta q n^{a+b(\gamma+1)}$ for some $\delta \in (0,1)$ and $\boldsymbol{\Gamma}_n$ is $q \times q$, for each $n \geq 1$, there exist a $j_n \in \{1, \ldots, q\}$ such that $(\boldsymbol{\Gamma}_n)_{j,j} \geq \delta n^{a+b(\gamma+1)}$. Write $\mathcal{C}_{q,n} = \{\{\mathbf{x} \in \mathbb{R}^q : x_{j_n} \in (-a, a)\} : a \in \mathbb{R}\}$. Also, let $\breve{\tau}_n^2 = \sigma^2 \cdot (\breve{\Upsilon}_n)_{j_n, j_n}$ and $\tau_n^2 = \sigma^2 \cdot (\Upsilon_n)_{j_n, j_n}$. Then, $I_{k,n} = 0$, for all $B \in \mathcal{C}_{q,n}$ for $k = 2, 3$, (7.2) and by (7.1)–(7.3),

$$\Delta_n \geq \sup_{B \in \mathcal{C}_n} |I_{1,n}(B)| + o(n^{-1/2} + \|\mathbf{b}_n\|)$$
$$= \sup\left\{ \left| \int_{-a}^{a} [\phi(x, \breve{\tau}) - \phi(x, \tau)]\, dx \right| : a \in \mathbb{R} \right\} + o(n^{-1/2} + \|\mathbf{b}_n\|)$$
$$\geq K|\breve{\tau}_n^2 - \tau_n^2| + o(n^{-1/2} + \|\mathbf{b}_n\|) \tag{7.4}$$
$$\geq K \cdot \delta\gamma \cdot \frac{\lambda_n}{n} \cdot n^{a+b(\gamma+1)} + o(n^{-1/2} + \|\mathbf{b}_n\|).$$

This proves part (b) in the case where $n^{-1/2} + \frac{\lambda_n}{\sqrt{n}} \cdot n^{b\gamma} = O(\lambda_n \cdot n^{-1+a+b(\gamma+1)})$. A subsequence argument proves part (b) when this condition fails. See [CL] for more details. □

LEMMA 7.3. *Suppose that conditions* (C.1)′–(C.6)′ *holds with* $r = 5$, *and let* $n^{-1} \sum_{i=1}^{n} \|\mathbf{C}_{11,n}^{-1/2} \mathbf{x}_i^{(1)}\|^5 = O(1)$. *Then, for any* $\delta > 0$ *and* $K \in (0, \infty)$, *there exists* $\delta_0 \in (0, 1)$ *such that*

$$\sup\{|\widehat{\omega}_n(t_1, t_2)| : \delta^2 \leq t_1^2 + t_2^2 \leq n^K\} = 1 - \delta_0 + o_p(1),$$

where

$$\widehat{\omega}_n(t_1, t_2) = \mathbf{E}_* \exp(\iota t_1 \varepsilon_1^* + \iota t_2 (\varepsilon_1^*)^2),$$
$$\omega(t_1, t_2) = \mathbf{E} \exp(\iota t_1 \varepsilon_1 + \iota t_2 (\varepsilon_1)^2), \qquad t_1, t_2 \in \mathbb{R}.$$

PROOF. See [CL]. □

PROOF OF THEOREM 4.1. Restricting attention to a suitable set $A_{3,n}$ with $\mathbf{P}(A_{3,n}) \to 1$ and retracing the steps in the proof of Theorem 7.2, one can show (cf. [CL]) that

$$\sup_{B\in\mathcal{C}_q}\left|\mathbf{P}_*(\mathbf{T}_n^* \in B) - \int_B \widehat{\psi}_n(\mathbf{x})\,d\mathbf{x}\right| = o(n^{-1/2});$$
(7.5)
$$\sup_{B\in\mathcal{C}_q}\left|\mathbf{P}_*(\mathbf{R}_n^* \in B) - \int_B \widehat{\pi}_n(\mathbf{x})\,d\mathbf{x}\right| = o(n^{-1/2}),$$

where $\widehat{\psi}_n$ and $\widehat{\pi}_n$ are obtained from ψ_n and π_n, respectively, by replacing $(\sigma^2, \mu_3, \mathbf{b}_n')$ by $(\widehat{\sigma}_n^2, \widehat{\mu}_{3,n}, \widehat{\mathbf{b}}_n')$, where

$$\widehat{\sigma}_n^2 = \mathrm{Var}_*(\varepsilon_1^*), \qquad \widehat{\mu}_{3,n} = \mathbf{E}_*(\varepsilon_1^* - \mathbf{E}_*\varepsilon_1^*)^3, \qquad \widehat{\mathbf{b}}_n = \mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}\widehat{\mathbf{s}}_n^{(1)},$$

and the jth element of $\widehat{\mathbf{s}}_n^{(1)}$ is given by $\mathrm{sgn}(\widehat{\beta}_{j,n})\lambda_n \cdot n^{-1/2} \cdot |\widehat{\beta}_{j,n}|^{-\gamma}$, $1 \le j \le p_0$. For part (a), we have, for $n \ge n_0$,

$$\begin{aligned}
&\mathbf{P}\Big(\sup_{B\in\mathcal{C}_q}\big|\mathbf{P}_*(\mathbf{T}_n^* \in B) - \mathbf{P}(\mathbf{T}_n \in B)\big| > Kn^{-1/2}\Big)\\
&\quad\le \mathbf{P}\Big(\Big\{\sup_{B\in\mathcal{C}_q}\big|\widehat{\Psi}_n(B) - \Psi_n(B)\big| > Kn^{-1/2}\Big\} \cap A_{3,n}\Big) + \mathbf{P}(A_{3,n}^c)\\
&\quad\le \mathbf{P}\Big(\int \big|\phi(\mathbf{x}; \widehat{\sigma}^2\breve{\Upsilon}_n) - \phi(\mathbf{x}; \sigma_n^2\breve{\Upsilon}_n)\big|\,d\mathbf{x} > Kn^{-1/2}\Big) + \mathbf{P}(A_{3,n}^c)\\
&\quad\le \mathbf{P}\big(|\hat{\sigma}_n^2 - \sigma^2| > Kn^{-1/2}\big) + o(1),
\end{aligned}$$

which can be made arbitrarily small by choosing $K \in (0, \infty)$ large. Hence, part (a) follows. The proof of part (b) is similar; see [CL] for more details. □

PROOF OF THEOREM 4.3. From the proof of Theorem 7.2 in [CL], there exists a set $A_{1,n}$ with $P(A_{1,n}^c) = o(n^{-1})$, such that on $A_{1,n}^c$ and for $n \ge n_0$,

$$\begin{aligned}
&\widehat{I}_n = I_n \quad \text{and}\\
&\breve{\mathbf{R}}_n \equiv \frac{\sqrt{n}\mathbf{D}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) + \breve{\mathbf{b}}_n}{\breve{\sigma}_n}\\
&\quad = \Big[\Big\{\mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}\mathbf{W}_n^{(1)} - \frac{\lambda_n}{\sqrt{n}}\mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}\tilde{\mathbf{s}}_n^{(1)}\Big\} + \frac{\lambda_n}{\sqrt{n}}\mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}\mathbf{s}_n^{\dagger(1)}\Big] \cdot \frac{1}{\breve{\sigma}_n}\\
&\quad \equiv \mathbf{D}_n^{(1)}\mathbf{C}_{11,n}^{-1}\mathbf{W}_n^{(1)} \cdot \frac{1}{\breve{\sigma}_n} + \mathbf{Q}_{3,n} \qquad \text{(say)},
\end{aligned}$$

where, $\mathbf{Q}_{3,n} = \frac{\lambda_n}{\sqrt{n}} \cdot \mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} (\mathbf{s}_n^{\dagger(1)} - \widetilde{\mathbf{s}}_n^{(1)})$, and the jth element of $\mathbf{s}_n^{\dagger(1)}$ is given by $s_{j,n}^{\dagger} = \operatorname{sgn}(\widehat{\beta}_{j,n})|\widetilde{\beta}_{j,n}|^{-\gamma}$, $1 \le j \le p_0$. Note that

$$\begin{aligned}
&\mathbf{P}(\|\mathbf{Q}_{3,n}\| \neq 0) \\
&\quad \le \mathbf{P}(\{\mathbf{s}_n^{\dagger(1)} \neq \widetilde{\mathbf{s}}_n^{(1)}\} \cap A_{1,n}) + \mathbf{P}(A_{1,n}^c) \\
&\quad \le \mathbf{P}(\{\operatorname{sgn}(\widehat{\beta}_{j,n}) \neq \operatorname{sgn}(\beta_{j,n}), \text{for some } 1 \le j \le p_0\} \cap A_n) + \mathbf{P}(A_{1,n}^c) \\
&\quad = 0 + \mathbf{P}(A_{1,n}^c) \qquad \text{for } n \ge n_0 \\
&\quad = o(n^{-1}).
\end{aligned}$$

Next, using Taylor's expansion, one can write

$$\begin{aligned}
\breve{\mathbf{R}}_n &= \mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} \mathbf{W}_n^{(1)} \left[\sigma^{-1} - \frac{1}{2\sigma^3}(\breve{\sigma}_n^2 - \sigma^2) + \frac{3}{4\sigma^5} \frac{(\breve{\sigma}_n^2 - \sigma^2)^2}{2!}\right] + \mathbf{Q}_{4,n} \\
&\equiv \breve{\mathbf{R}}_{1,n} + \mathbf{Q}_{4,n} \qquad \text{(say)},
\end{aligned}$$

where $\mathbf{P}(\|\mathbf{Q}_{4,n}\| > K n^{-3/2}(\log n)^2) = o(n^{-1})$. As a consequence, EEs for $\breve{\mathbf{R}}_n$ and $\breve{\mathbf{R}}_{1,n}$ coincide upto order n^{-1}. Now using arguments in the proof of Theorem 7.2(b), combined with the arguments in Götze (1987) and Lahiri (1994), and then using the transformation technique of Bhattacharya and Ghosh (1978), one can show (see [CL] for details) that

$$\sup_{B \in \mathcal{C}_q} \left| \mathbf{P}(\breve{\mathbf{R}}_n \in B) - \int_B \pi_{1,n}(\mathbf{x})\, d\mathbf{x} \right| = o(n^{-1}), \tag{7.6}$$

where

$$\pi_{1,n}(\mathbf{x}) = \phi(\mathbf{x} : \Upsilon_n)[1 + n^{-1/2} p_{1,n}(\mathbf{x}; \sigma^2, \mu_3) + n^{-1} p_{2,n}(\mathbf{x}; \sigma^2, \mu_3, \mu_4)],$$

with $\mu_4 = \mathbf{E}\varepsilon_1^4$ and where $p_{1,n}(\cdot)$ and $p_{2,n}(\cdot)$ are polynomials of degree 3 and 6, respectively, with coefficients that are rational functions of the respective sets of parameters such that the denominators depend only on σ^2 [as in the definition of $\pi_n(\cdot)$].

Next, using Lemma 7.3 and similar arguments, one can show that

$$\sup_{B \in \mathcal{C}_q} \left| \mathbf{P}_*(\breve{\mathbf{R}}_n^* \in B) - \int_B \widehat{\pi}_{1,n}(\mathbf{x})\, d\mathbf{x} \right| = o_p(n^{-1}), \tag{7.7}$$

where

$$\widehat{\pi}_{1,n}(x) = \phi(x; \Upsilon_n)[1 + n^{-1/2} p_{1,n}(x; \widehat{\sigma}_n^2, \widehat{\mu}_{3,n}) + n^{-1} p_{2,n}(x; \widehat{\sigma}_n^2, \widehat{\mu}_{3,n}, \widehat{\mu}_{4,n})],$$

with $\widehat{\sigma}_n^2 = \mathbf{E}_*(\varepsilon_1^*)^2$, $\widehat{\mu}_{k,n} = \mathbf{E}_*(\varepsilon_1^*)^k$, $k = 3, 4$. Theorem 4.3 now follows from (7.6) and (7.7). □

PROOF OF THEOREM 5.1. Using the arguments similar to the proof of Theorem 7.2, one can show that

$$\mathbf{T}_n = \mathbf{D}_n^{(1)} \mathbf{C}_{11,n}^{-1} \mathbf{W}_n^{(1)} - \mathbf{b}_n + \Delta_{1,n} \equiv \mathbf{T}_{1,n}^{\dagger} + \Delta_{1,n} \qquad \text{(say)}, \tag{7.8}$$

where

$$\mathbf{P}\big(\|\Delta_{1,n}\| > K\lambda_n \sqrt{p_0 \log n}/n\big) = o\big(n^{-1/2}\big). \tag{7.9}$$

Note that by (C.6), $\lambda_n n^{-1}\sqrt{p_0 \log n} = o(n^{-1/2})$, when $b = 0$. Now using the arguments in the proof of Theorem 7.2 (with $\eta_i^{(0)} = 0$ for all $i = 1, \dots, n$), one can conclude (cf. [CL]) that

$$\sup_{B \in \mathcal{C}_q} \left| \mathbf{P}(\mathbf{R}_n \in B) - \int_B \pi_n^{\dagger}(\mathbf{x})\, d\mathbf{x} \right| = o\big(n^{-1/2}\big), \tag{7.10}$$

and that

$$\sup_{B \in \mathcal{C}_q} \left| \mathbf{P}_*\big(\mathbf{R}_n^* \in B\big) - \int_B \big(\pi^{\dagger}\big)^*(\mathbf{x})\, d\mathbf{x} \right| = o_p\big(n^{-1/2}\big), \tag{7.11}$$

where $\pi_n^{\dagger}(\cdot)$ is defined by setting $\eta_i^{(0)} = 0$ for $1 \le i \le n$ in the definition of $\pi_n(\cdot)$, and where $(\pi^{\dagger})^*(\cdot)$ is obtained from $\pi^{\dagger}(\cdot)$ by replacing $\mathbf{b}_n$, σ^2 and μ_3 with $\widehat{\mathbf{b}}_n$, $\widehat{\sigma}^2$ and $\widehat{\mu}_{3,n}$, as in (7.5). Using (7.10) and (7.11), one can conclude that

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

The proof for $\breve{\mathbf{R}}_n$ is similar. We omit the routine details to save space. □

Acknowledgments. We thank three anonymous referees, the Associate Editor and the Co-Editor, Professor Tony Cai, for a number of constructive comments that, in particular, led to the addition of Section 5 on the $p > n$ case and, also the real data example in Section 6.5.

The first author acknowledges the help from the staff, excellent infrastructure and atmosphere and financial support from the Statistical and Applied Mathematical Sciences Institute (SAMSI), Raleigh, NC, and the Department of Statistics at North Carolina State University, Raleigh, NC, where part of this work was completed.

SUPPLEMENTARY MATERIAL

Supplement to "Rates of convergence of the Adaptive LASSO estimators to the Oracle distribution and higher order refinements by the bootstrap" (DOI: 10.1214/13-AOS1106SUPP; .pdf). Detailed proofs of all results.

REFERENCES

BACH, F. (2009). Model-consistent sparse estimation through the bootstrap. Preprint. Available at http://arxiv.org/abs/0901.3202.

BERK, R. A., BROWN, L. D., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post selection inference. *Ann. Statist.* **41** 802–837.

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

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

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. (2011a). Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.* **106** 608–625. MR2847974

CHATTERJEE, A. and LAHIRI, S. N. (2011b). Strong consistency of Lasso estimators. *Sankhyā A* **73** 55–78. MR2887087

CHATTERJEE, A. and LAHIRI, S. N. (2013). Supplement to "Rates of convergence of the adaptive LASSO estimators to the Oracle distribution and higher order refinements by the bootstrap." DOI:10.1214/13-AOS1106SUPP.

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

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

FREEDMAN, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218–1228. MR0630104

GÖTZE, F. (1987). Approximations for multivariate U-statistics. *J. Multivariate Anal.* **22** 212–229. MR0899659

GUPTA, S. (2012). A note on the asymptotic distribution of LASSO estimator for correlated data. *Sankhyā A* **74** 10–28. MR3010290

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

HALL, P. and MILLER, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.* **18** 533–550. MR2751640

HUANG, J., HOROWITZ, J. L. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613. MR2396808

HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18** 1603–1618. MR2469326

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

LAHIRI, S. N. (1994). On two-term Edgeworth expansions and bootstrap approximations for Studentized multivariate M-estimators. *Sankhyā A* **56** 201–226. MR1664912

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

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

PÖTSCHER, B. M. and SCHNEIDER, U. (2009). On the distribution of the adaptive LASSO estimator. *J. Statist. Plann. Inference* **139** 2775–2790. MR2523666

SEGAL, M., DAHLQUIST, K. and CONKLIN, B. (2003). Regression approaches for microarray data analysis. *J. Comput. Biol.* **10** 961–980.

STAMEY, T. A., KABALIN, J. N., MCNEAL, J. E., JOHNSTONE, I. M., FREIHA, F., REDWINE, E. A. and YANG, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *J. Urol.* **141** 1076–1083.

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

WAINWRIGHT, M. J. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. Technical report, Dept. of Statistics, Univ. California, Berkeley. Available at http://arxiv.org/abs/math/0605740.

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

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

STATISTICS AND MATHEMATICS UNIT
INDIAN STATISTICAL INSTITUTE
NEW DELHI 110067
INDIA
E-MAIL: cha@isid.ac.in

DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NORTH CAROLINA 27695
USA
E-MAIL: snlahiri@ncsu.edu