

# Bootstrapping Lasso estimators

A. Chatterjee\*

S. N. Lahiri†

Stat-Math Unit

Department of Statistics

Indian Statistical Institute, New Delhi

Texas A&M University

## Abstract

In this paper, we consider bootstrapping the Lasso estimator of the regression parameter in a multiple linear regression model. It is known that the standard bootstrap method fails to be consistent. Here, we propose a modified bootstrap method, and show that it provides valid approximation to the distribution of the Lasso estimator, for all possible values of the unknown regression parameter vector, including the case where some of the components are zero. Further, we establish consistency of the modified bootstrap method for estimating the asymptotic bias and variance of the Lasso estimator. We also show that the residual bootstrap can be used to consistently estimate the distribution and variance of the Adaptive Lasso estimator. Using the former result, we formulate a novel data based method for choosing the optimal penalizing parameter for the Lasso using the modified bootstrap. A numerical study is performed to investigate the finite sample performance of the modified bootstrap. The methodology proposed in the paper is illustrated with a real data example.

*AMS (2000) Subject Classification:* Primary: 62J07; Secondary: 62G09, 62E20.

*Keywords and Phrases:* Penalized Regression, bootstrap variance estimation, regularization, shrinkage.

## 1 Introduction

Consider the following regression model

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n, \quad (1.1)$$

---

\*email: cha@isid.ac.in

†email: snlahiri@stat.tamu.edu

where,  $y_i$  is the response,  $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})'$  is a  $p$  dimensional covariate vector,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  is the regression parameter and  $\{\epsilon_i : i = 1, \dots, n\}$  are independent and identically distributed errors. We assume that  $p$  is fixed. The Lasso estimator of  $\boldsymbol{\beta}$  is defined as the minimizer of the  $l_1$ -norm penalized least-square criterion function,

$$\widehat{\boldsymbol{\beta}}_n = \underset{\mathbf{u} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \mathbf{x}_i' \mathbf{u})^2 + \lambda_n \sum_{j=1}^p |u_j|, \quad (1.2)$$

where,  $\lambda_n$  is a regularization parameter. The Lasso estimator was introduced by Tibshirani (1996) as an estimation and variable selection method. The Lasso estimator has two nice properties, namely, (i) the nature of regularization used in the Lasso leads to sparse solutions and (ii) it is also computationally feasible (see Efron et al. (2004), Osborne et al. (2000), Friedman et al. (2007)). The sparse solutions obtained by using the Lasso automatically leads to model selection. In the finite dimensional case, many authors have studied the model-consistency properties of the Lasso and investigated conditions under which the Lasso can recover the true sparsity pattern (see Zhao and Yu (2006), Wainwright (2006), Zou (2006)). Other than the linear model setup like (1.1), Yuan and Lin (2007) have studied the neighborhood selection properties of the Lasso in graphical models. Recently Bach (2009) considered using bootstrap samples in order to improve the model selection accuracy of the Lasso.

An important problem in this context is the estimation consistency of the Lasso. This was first studied by Knight and Fu (2000) for the finite dimensional regression model (1.1). The asymptotic distribution was found and it was shown that the Lasso was weakly consistent. They also showed that if  $\lambda_n$  was sufficiently large, then some components of the Lasso estimate may be exactly zero. It was found that under appropriate regularity conditions, the limiting distribution of the Lasso estimator assigns positive mass at zero for the components where the true regression parameter has zero values. Since the limit distribution of the Lasso estimator is complicated (cf. Knight and Fu (2000)), it is important to have alternative approximations to the distribution of the Lasso estimator that can be used in practice to set confidence regions and to carry out tests on the parameter vector. Knight and Fu (2000) considered using the bootstrap to generate alternative approximations. More specifically, Knight and Fu (2000) considered the residual-based bootstrap method (cf. Freedman (1981)) for the Lasso estimator and sketched out its asymptotic behavior. Recently, it is further investigated rigorously by Chatterjee and Lahiri (2010), who show that the asymptotic distribution of the bootstrapped Lasso estimator is a random measure on  $\mathbb{R}^p$  and that the bootstrap is inconsistent whenever one or more components of the regression parameter is zero. Thus, in situations where the limit distribution of the Lasso estimator is most complicated and alternative approximations are needed the most, the usual bootstrap fails drastically! In this paper, we construct a suitable modification to the residual based bootstrap method and show that under mild regularity conditions, the modified version of the bootstrap is

indeed consistent in estimating the limiting distribution of the Lasso estimator, even when some components of  $\beta$  are zero.

Another important issue that has eluded a satisfactory solution to date is the problem of attaching standard error estimates to the Lasso estimates. Initially, Tibshirani (1996) suggested an approximation that had the drawback of providing zero standard error estimates when the estimated coefficient was zero. Osborne et al. (2000) suggested an improved alternative but as pointed out by Knight and Fu (2000), all these methods suffered from the drawback of considering the Lasso as an approximately linear transformation. Other related methods (like the 'sandwich formula') for variance estimation in penalized regression setup were suggested by Fan and Li (2001) and Fan and Peng (2004) which only provided variance estimates for the underlying non-zero components. One of the main contributions of this paper is to show that the modified bootstrap method, that we propose here, gives a consistent variance estimator for the Lasso, for both zero and nonzero parameter values. In particular, the bootstrap based variance estimate overcomes the drawbacks of some of the earlier variance estimation techniques, like producing zero variance estimates for estimated zero coefficients. As an application to this result, we provide bootstrap based confidence regions for the true parameter vector.

From Knight and Fu (2000)'s work, it is known that the asymptotic distribution of the Lasso estimator depends on the regularization parameter through  $\lambda_0$  where  $\lambda_n n^{-1/2} \rightarrow \lambda_0$ . In particular, the accuracy of the Lasso estimator  $\hat{\beta}_n$  critically depends on the choice of the regularization parameter  $\lambda_n$ . In this context, we formulate a new data-based method for selection of the regularization parameter. Since the modified bootstrap is consistent for the mean squared error (MSE) of the Lasso estimator, we define a criterion function based on the modified bootstrap estimator of the MSE as a rescaled function of  $\lambda_n$ . In this context, recently Hall et al. (2009) suggested using the the m-out-of-n bootstrap, to choose the optimal regularization parameter in the Adaptive Lasso setup.

We also study the properties of the bootstrapped version of the Adaptive Lasso estimator (Zou (2006)). We find that the residual bootstrap based version of the Adaptive Lasso estimator is consistent in estimating the asymptotic distribution and variance of the Adaptive Lasso estimator. Similar to Lasso estimator, the question about the validity of the bootstrap for the Adaptive Lasso has been unresolved till now and our results show that the simple residual bootstrap can consistently estimate the distribution and provide variance estimates for the Adaptive Lasso estimator. This is unlike the case for the Lasso, where simple residual bootstrap fails.

We conclude this section with a brief literature review. Knight and Fu (2000) derived the asymptotic distribution of the Lasso estimator under model (1.1) in the case where the dimension  $p$  of the regression

is fixed. Properties of the standard bootstrap method have been investigated by [Knight and Fu \(2000\)](#) and [Chatterjee and Lahiri \(2010\)](#), in the same set up. In the finite dimensional case, Pötscher and co-authors have interesting results on the impossibility of estimating the distribution function of a Lasso estimator in a uniform sense. The relevance of their results in the context of our work is discussed in [Section 3.1](#). In the high-dimensional case, where  $p$  is allowed to grow with  $n$ , work on estimation consistency of the Lasso is limited; among them [Huang et al. \(2008\)](#) considered the asymptotic properties of  $\ell_q$  norm penalized regression estimators ( $0 < q < 1$ ) for high dimensional regression models. There is a large amount of literature on the variable selection properties of the Lasso and we do not attempt to summarize all the work. We refer to the recent paper by [Zhang et al. \(2008\)](#), who describe a two-step procedure for variable selection using the Lasso in the high dimensional setup, and also provide a clear picture of recent developments on variable selection in the high dimensional setting.

The rest of the paper is organized as follows. In [Section 2.1](#), we briefly review the residual based bootstrap and motivate the intuitive reasons behind its failure. We formulate the modified bootstrap method in [Section 2.2](#). The main results on consistency of modified bootstrap are stated [Section 3](#). In [Section 4](#), the results on the consistency of the residual bootstrap for the Adaptive Lasso estimator are presented. The data-based method for the selection of the (MSE-optimal) regularization parameter is presented in [Section 5](#). The finite sample performance of the proposed modified bootstrapped Lasso estimator and the data-based methodology of choosing the optimal regularization parameter is studied in [Section 6](#) using a simulated data set. A real data example is presented in [Section 7](#). The proofs are provided in [Section 8](#).

## 2 Formulation of the modified bootstrap method

### 2.1 Background and motivation

In a regression setup like [\(1.1\)](#), there are two approaches to bootstrapping depending on whether the  $\mathbf{x}_i$ 's are assumed to be random or not. In the case where  $\mathbf{x}_i$  are random, it is of interest to study the joint distribution of the covariates and the response and hence pairwise bootstrap is a relevant choice. In contrast, here we assume that the  $\mathbf{x}_i$ 's are non-random. In this situation, the standard approach to bootstrapping is the residual bootstrap (cf. [Efron \(1979\)](#), [Freedman \(1981\)](#)), which was considered by [Knight and Fu \(2000\)](#) in the context of the Lasso estimator. To motivate the modified bootstrap method, we first give a brief description of the residual bootstrap. Let  $\widehat{\boldsymbol{\beta}}_n$  denote the Lasso estimator of  $\boldsymbol{\beta}$  given by [\(1.2\)](#). Define the residuals

$$e_i = y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}_n, \quad i = 1, \dots, n.$$

Consider the set of centered residuals  $\{e_i - \bar{e}_n : i = 1, \dots, n\}$ , where  $\bar{e}_n = n^{-1} \sum_{i=1}^n e_i$ . For the residual bootstrap, one selects a with replacement sample of size  $n$ ,  $\{e_i^* : i = 1, \dots, n\}$ , from the set of centered residuals and formulates the bootstrap version of (1.1) as

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

Next, based on the bootstrap dataset  $\{(y_i^*, \mathbf{x}_i) : i = 1, \dots, n\}$ , the bootstrap version of the Lasso estimator is defined as:

$$\boldsymbol{\beta}_n^* = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} \sum_{i=1}^n (y_i^* - \mathbf{x}_i' \mathbf{u})^2 + \lambda_n \sum_{j=1}^p |u_j|, \quad (2.1)$$

The bootstrap version of  $\mathbf{T}_n \equiv n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$  is  $\mathbf{T}_n^* = n^{1/2}(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n)$ . The residual bootstrap estimator of the unknown distribution  $G_n$  (say) of  $\mathbf{T}_n$  is the (conditional) distribution  $\widehat{G}_n(\cdot)$  (say) of  $\mathbf{T}_n^*$  given the observations  $\{y_i : i = 1, \dots, n\}$ , i.e.,

$$\widehat{G}_n(B) = \mathbf{P}_*(\mathbf{T}_n^* \in B), \quad B \in \mathcal{B}(\mathbb{R}^p), \quad (2.2)$$

where  $\mathbf{P}_*$  denotes conditional probability given the error variables  $\{\epsilon_i : i = 1, \dots, n\}$  and  $\mathcal{B}(\mathbb{R}^p)$  denotes the Borel  $\sigma$ -field on  $\mathbb{R}^p$ .

For the bootstrap approximation to be useful, one would expect  $\widehat{G}_n(\cdot)$  to be close to  $G_n(\cdot)$ . However, this is not the case; [Chatterjee and Lahiri \(2010\)](#) show that the residual bootstrap estimator  $\widehat{G}_n(\cdot)$ , instead of converging to the deterministic limit of  $G_n$  given by [Knight and Fu \(2000\)](#), converges weakly to a random probability measure and therefore, it fails to provide a valid approximation to  $G_n(\cdot)$ . To appreciate why the residual bootstrap approximation have a random limit and why it is inconsistent, first observe that the Lasso estimators of the non-zero components of  $\boldsymbol{\beta}$  estimate their signs correctly with high probability but the estimators of the zero-components take both positive and negative values with positive probabilities, thereby erring to capture the target sign value (which is zero for such components) closely. A close examination of the proof of the main result (cf. Theorem 3.1 in [Chatterjee and Lahiri \(2010\)](#)), shows that although the formulation of the residual bootstrap mimics the main features of the regression model closely, it fails to reproduce the sign of the zero-components of  $\boldsymbol{\beta}$  with sufficient accuracy in the formulation of the bootstrap Lasso estimation criterion (2.1), leading to the random limit.

## 2.2 A modified bootstrap method

Based on the discussion of the last paragraph, we now propose a modified version of the bootstrapped Lasso estimator that more closely reproduces the sign-vector corresponding to the unknown parameter  $\beta$ . As seen in Chatterjee and Lahiri (2010), the inconsistency of the standard residual bootstrap arises when some components of  $\beta$  are zero. The key idea behind the modified bootstrap is to force components of the Lasso estimator  $\widehat{\beta}_n$  to be exactly zero whenever they are close to zero. Since the original Lasso estimator is root- $n$  consistent, its fluctuations are of the order  $n^{-1/2}$  around the true value. This suggests a neighborhood of order larger than  $n^{-1/2}$  around the true value will contain the values of the Lasso estimator with high probability. To that end, let  $\{a_n\}$  be a sequence of real numbers such that

$$a_n + (n^{-1/2} \log n) a_n^{-1} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (2.3)$$

For example,  $a_n = cn^{-\delta}$  satisfies (2.3) for all  $c \in (0, \infty)$  and  $\delta \in (0, 2^{-1})$ . We threshold the components of the Lasso estimator  $\widehat{\beta}_n$  at  $a_n$  and define the modified Lasso estimator as

$$\begin{aligned} \widetilde{\beta}_n &= (\widetilde{\beta}_{n,1}, \dots, \widetilde{\beta}_{n,p})', \quad \text{with} \\ \widetilde{\beta}_{n,j} &= \widehat{\beta}_{n,j} \mathbf{1}(|\widehat{\beta}_{n,j}| \geq a_n), \quad j = 1, \dots, p, \end{aligned} \quad (2.4)$$

where  $\widehat{\beta}_n$  is the usual Lasso estimate defined in (1.2) and  $\mathbf{1}(\cdot)$  denotes the indicator function. Note that for a nonzero component  $\beta_j$ ,

$$|\widehat{\beta}_{n,j}| = |\beta_j| + O_p(n^{-1/2}) > \frac{|\beta_j|}{2} \geq a_n$$

for  $n$  large, with high probability and therefore,  $\widetilde{\beta}_{n,j} = \widehat{\beta}_{n,j}$ , for  $n$  large and with probability tending to 1. Thus, this shrinkage does not have any significant effect on the non-zero components. However, for a zero component,  $\beta_j = 0$ ,

$$|\widehat{\beta}_{n,j}| = |\beta_j| + O_p(n^{-1/2}) = O_p(n^{-1/2}) \in [-a_n, a_n],$$

with probability tending to 1 as  $n \rightarrow \infty$ , and thus

$$\widetilde{\beta}_{n,j} = \widehat{\beta}_{n,j} \mathbf{1}(|\widehat{\beta}_{n,j}| > a_n) = 0 \quad \text{for large } n,$$

with probability tending to 1. In particular, the shrinkage by  $a_n$  accomplishes our main objective: namely, to capture the signs of the zero components precisely with probability tending to 1, as the sample size  $n$  goes to infinity.

Next, we define the modified residuals  $\{r_i : i = 1, \dots, n\}$  based on this estimator  $\tilde{\boldsymbol{\beta}}_n$  by

$$r_i = y_i - \mathbf{x}_i' \tilde{\boldsymbol{\beta}}_n, \quad i = 1, \dots, n. \quad (2.5)$$

Let  $\bar{r}_n = n^{-1} \sum_{i=1}^n r_i$ . We select a random sample  $\{e_1^{**}, \dots, e_n^{**}\}$  of size  $n$  with replacement from the centered residuals  $\{r_i - \bar{r}_n : i = 1, \dots, n\}$  and set

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

Then, the modified bootstrap Lasso estimator is

$$\boldsymbol{\beta}_n^{**} := \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} \sum_{i=1}^n (y_i^{**} - \mathbf{x}_i' \mathbf{u})^2 + \lambda_n \sum_{j=1}^p |u_j|. \quad (2.6)$$

Let

$$\mathbf{T}_n^{**} = n^{1/2} \left( \boldsymbol{\beta}_n^{**} - \tilde{\boldsymbol{\beta}}_n \right), \quad n \geq 1,$$

and let  $\tilde{G}_n(\cdot)$  denote the conditional distribution of  $\mathbf{T}_n^{**}$  given the observations (or the error variables  $\{\epsilon_i : i = 1, \dots, n\}$ ), i.e.,  $\tilde{G}_n(B) = \mathbf{P}_*(\mathbf{T}_n^{**} \in B)$ ,  $B \in \mathcal{B}(\mathbb{R}^p)$ . Thus,  $\tilde{G}_n(\cdot)$  is the modified bootstrap approximation to the unknown distribution  $G_n(\cdot)$  of  $\mathbf{T}_n$ . The modified bootstrap estimator of a population parameter  $\theta_n = \varphi(G_n)$ , defined through a functional  $\varphi(\cdot)$  of  $G_n$ , is  $\varphi(\tilde{G}_n)$ . For example, the modified bootstrap estimator of  $\mathbf{E}(\mathbf{T}_n) =$  the scaled bias of  $\hat{\boldsymbol{\beta}}_n$  is  $\mathbf{E}_*(\mathbf{T}_n^{**})$  and similarly, that of  $\mathbf{Var}(\mathbf{T}_n)$  is  $\mathbf{Var}_*(\mathbf{T}_n^{**})$ , where  $\mathbf{E}_*$  and  $\mathbf{Var}_*$  denote the expectation and variance under  $\mathbf{P}_*$ .

**Remark 1:** It should be noted that centering the residuals  $\{r_i : i = 1, \dots, n\}$  is a must for the validity of the residual bootstrap in general, as it ensures that the bootstrap analogue of the model condition  $\mathbf{E}(\epsilon_1) = 0$ . Note that this is a sufficient condition to ensure  $\mathbf{E}_*(\sum_{i=1}^n \mathbf{x}_i \epsilon_i^{**}) = \mathbf{0}$ . For centered  $\mathbf{x}_i$ 's, the 'centering of residuals' step can be bypassed, as the conditional mean of the sum  $\sum_{i=1}^n \mathbf{x}_i \epsilon_i^{**}$  will still be zero. For real data-sets, where the responses and covariates are already known to be centered, the centering of residuals is not a required step.

On the other hand, scaling the residuals is not as critical a condition: the residual bootstrap is known to be consistent for the OLS (and also for the Lasso, as shown here) without any scale adjustments. In the literature, the rescaling factor  $(1 - p/n)^{-1/2}$  is sometimes used (see [Efron \(1982\)](#)) to improve finite sample accuracy, but asymptotically this has negligible effect as  $p$  is fixed in this case.

**Remark 2:** In (2.4), the thresholded version of the Lasso estimator  $\tilde{\boldsymbol{\beta}}_n$  has been constructed by thresholding the usual Lasso estimator  $\hat{\boldsymbol{\beta}}_n$ . It is possible to replace  $\hat{\boldsymbol{\beta}}_n$  by any other  $\sqrt{n}$ -consistent estimator of

$\beta$ , and this fact follows from the proof of Theorem 3.1. For example, the usual least-squares estimator of  $\beta$  can be used. In such a situation,  $\tilde{\beta}_n$  can be similarly defined by thresholding each component of the selected  $\sqrt{n}$ -consistent estimator of  $\beta$ . In the context of computational efficiency, the choice of an alternative estimator, like the usual least-squares estimator, does provide some computational advantage over using the Lasso estimator  $\hat{\beta}_n$ . But, in practical situations, with the existence of extremely fast computational algorithms (Efron et al. (2004), Friedman et al. (2007)), that can be used for computing Lasso estimators, the extra effort in computing the Lasso estimator  $\hat{\beta}_n$  becomes a minor issue.

In the next section we show that under some mild conditions, the modified bootstrap estimators of the distribution function of  $T_n$  and of its bias and variance functionals are consistent.

### 3 Bootstrapping the Lasso estimator

#### 3.1 Consistency of the distributional approximation

The first result shows that the modified bootstrap gives a valid approximation to the distribution of  $T_n$ :

**Theorem 3.1** (CONSISTENCY OF MODIFIED BOOTSTRAP). *Suppose that the following assumptions hold:*

$$(C.1) \quad n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \rightarrow \mathbf{C}, \text{ where } \mathbf{C} \text{ is positive definite. Further, } n^{-1} \sum_{i=1}^n \|\mathbf{x}_i\|^3 = O(1).$$

$$(C.2) \quad \lambda_n n^{-1/2} \rightarrow \lambda_0 \geq 0.$$

$$(C.3) \quad \text{The errors } \{\epsilon_i : i = 1, \dots, n\} \text{ are independent and identically distributed with } \mathbf{E}(\epsilon_1) = 0 \text{ and } \mathbf{Var}(\epsilon_1) = \sigma^2 \in (0, \infty).$$

Then

$$\varrho\left(\tilde{G}_n(\cdot), G_n(\cdot)\right) \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad \text{with probability 1,}$$

where  $\varrho(\cdot, \cdot)$  denotes the Prohorov metric on the set of all probability measures on  $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ .

Theorem 3.1 asserts strong consistency of the modified bootstrap distribution function estimator under Assumptions (C.1) - (C.3). In contrast, for the standard version of the residual bootstrap, Chatterjee and Lahiri (2010), shows that under the same set of regularity assumptions, if  $\lambda_0 > 0$  in Assumption (C.2) and if  $\beta$  has at least one zero-component, then

$$\varrho\left(\hat{G}_n(\cdot), G_n(\cdot)\right) \not\rightarrow 0 \quad \text{in probability, as } n \rightarrow \infty,$$

where  $\widehat{G}_n$  is the residual bootstrap estimate of  $G_n$  (cf. (2.2)). Thus, while the standard residual bootstrap has limited success in presence of zero-components, the modified bootstrap removes the limitation of the residual bootstrap, and provides a valid approximation to the distribution of the centered and scaled Lasso estimator for all values of the regression parameter  $\beta$ .

Next let  $G_\infty(\cdot)$  denote the limit distribution of  $\mathbf{T}_n$  (cf. Knight and Fu (2000)). Then, from Theorem 3.1, it follows that for any Borel set  $B \subset \mathbb{R}^p$  with  $G_\infty(\partial B) = 0$ ,

$$\mathbf{P}_*(\mathbf{T}_n^{**} \in B) - \mathbf{P}(\mathbf{T}_n \in B) \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

with probability 1, where  $\partial B$  is the boundary of  $B$ . As a result, one can use the modified bootstrap method to approximate the distribution of the centered and scaled Lasso estimator  $\mathbf{T}_n$  for any  $\beta \in \mathbb{R}^p$ , even when some of the components of  $\beta$  are zero. Since the limit distribution of  $\mathbf{T}_n$  is rather complicated in such cases, the standard approach of using the quantiles of the limit distribution to construct confidence sets for  $\beta$  and its components is not very easy to apply in practice. In contrast, the modified bootstrap method gives a viable and unified way to construct valid large sample confidence set estimators of  $\beta$  for all values of the unknown regression parameters  $\beta \in \mathbb{R}^p$ , including the cases where one or more components of  $\beta$  are zero. More specifically, let  $\widehat{t}_n(\alpha)$  denote the  $\alpha$  quantile of the bootstrap distribution of  $\|\mathbf{T}_n^{**}\|$ ,  $\alpha \in (0, 1)$ . Then, the set

$$I_{n,\alpha} \equiv \left\{ \mathbf{t} \in \mathbb{R}^p : \|\mathbf{t} - \widehat{\beta}_n\| \leq n^{-1/2} \widehat{t}_n(\alpha) \right\} \quad (3.1)$$

is an approximate confidence set for  $\beta$  of level  $\alpha$ , as shown by the following result. To state it, let  $\mathbf{T}_\infty$  denote the limiting random vector such that  $\mathbf{T}_n \rightarrow \mathbf{T}_\infty$  in distribution (cf. Knight and Fu (2000)), i.e.,  $\mathbf{T}_\infty$  has distribution  $G_\infty$ . Also, let  $t(\alpha)$  denote the  $\alpha$  quantile of  $\|\mathbf{T}_\infty\|$ ,  $\alpha \in (0, 1)$ .

**Corollary 3.2** (MODIFIED BOOTSTRAP CONFIDENCE INTERVAL). *Suppose the assumptions of Theorem 3.1 hold.*

(i) *If  $\alpha \in (0, 1)$  is such that  $\mathbf{P}(\|\mathbf{T}_\infty\| \leq t(\alpha) + \eta) > \alpha$  for all  $\eta > 0$ . Then,*

$$\mathbf{P}(\beta \in I_{n,\alpha}) \rightarrow \alpha \quad \text{as } n \rightarrow \infty, \quad (3.2)$$

for all  $\beta \in \mathbb{R}^p$ .

(ii) *Suppose that  $\{j : \beta_j \neq 0\}$  is nonempty. Then, (3.2) holds for all  $\alpha \in (0, 1)$ .*

Part (i) of Corollary 3.2 requires a mild condition on the  $\alpha$ , as the limiting distribution  $\mathbf{T}_\infty$  is partly

discrete, with a nontrivial mass at zero (cf. [Knight and Fu \(2000\)](#)) for the zero components. It rules out at most a countable set of values of  $\alpha$  when the distribution of  $\|\mathbf{T}_\infty\|$  is partly discrete. Part (ii) removes this under the condition that at least one component of  $\boldsymbol{\beta}$  is nonzero. The latter condition is satisfied in most applications, and is tantamount to justifying the use of the regression model (1.1). The main implication of Corollary 3.2 is that under some mild regularity conditions, the modified bootstrap method can be used to construct valid large sample confidence region for  $\boldsymbol{\beta}$ , including in the cases where one or more components of  $\boldsymbol{\beta}$  are zero. By exploiting the relationship between confidence regions and tests, it can also be used to test the null hypothesis  $H_0 : \beta_j = 0$  for all  $j \in J$  for a given  $J \subset \{1, \dots, p\}$ , which plays an important role in model selection.

**Remark 3:** In context to the consistency result presented in Theorem 3.1, it is important to mention the interesting work done by Pötscher and co-authors ([Leeb and Pötscher \(2006\)](#), [Leeb and Pötscher \(2008\)](#) and [Pötscher and Schneider \(2009\)](#)), who have studied the (uniform) consistency of estimates of distributions functions of the Lasso, Adaptive-Lasso and Hard-thresholding based estimators. Particularly, in context of the Lasso, they have shown that it is impossible to consistently estimate the distribution function of the Lasso estimator in a uniform sense: uniformity with respect to the underlying regression parameter  $\boldsymbol{\beta}$  in a shrinking neighborhood of the origin. And in particular, any resampling based estimator of the distribution function would also fail to be uniformly consistent. These results are of great interest, but uniform consistency of distribution function estimators is more relevant when many of the coefficients of the true regression parameter are very close to zero, but not exactly zero. And, such scenarios commonly arise in cases where the underlying regression parameter is allowed to depend on the sample size  $n$ , and changes with  $n$ .

On the other hand, in a classical linear regression set up, which is of interest in many applications (for example, see [Draper and Smith \(1998\)](#)), the underlying regression parameter is considered to be a *fixed* parameter, and the nonzero coefficients lie outside any given collection of shrinking neighborhoods of the origin, for large  $n$ . In such situations, estimating the distribution function of the Lasso estimator is still an important issue. In fact, the presence of some zero components in the true  $\boldsymbol{\beta}$  makes the usual bootstrap based estimator useless, because the presence of any zero component in the underlying  $\boldsymbol{\beta}$  will make the usual bootstrap estimator of  $\boldsymbol{\beta}$  (with appropriate scaling and centering) converge weakly to a random probability measure (cf. Theorem 3.1 in [Chatterjee and Lahiri \(2010\)](#)), instead of the actual target of interest (which is the asymptotic distribution of the Lasso estimator), thereby making the usual bootstrap inconsistent. In this context, we wish to emphasize that the issue of considering the underlying regression parameter  $\boldsymbol{\beta}$  as a *fixed* quantity is subjective and depends on a given situation, and in some cases, modeling the underlying

regression parameter as functions of the sample size may be a smarter choice. Nevertheless, the result in Theorem 3.1 provides a method to obtain a consistent estimator in case the underlying regression parameter  $\beta$  is fixed. This makes the modified bootstrap based estimator perfectly applicable in any linear regression set up, where the underlying true parameter can be considered as a *fixed* parameter. We stress that in our results, we do not claim to prove uniform consistency of the modified bootstrap based distribution function estimator, and hence our results are not in contradiction to the work of Pötscher and co-authors. Also, as noted by Andrews and Guggenberger (2009), the existence of a uniform consistent estimator of the sampling distribution is not necessary to achieve the goal of producing a uniformly valid confidence interval. Corollary 3.2 asserts that the modified bootstrap method is able to control the asymptotic size of a confidence interval. At this point, it is not clear if the resulting confidence intervals are valid uniformly in the parameter values.

**Remark 4:** The  $m$ -out-of- $n$  bootstrap technique has been recently used by Hall et al. (2009) in context of estimating the optimal penalty parameter for the Adaptive Lasso. We expect the  $m$ -out-of- $n$  bootstrap with  $m \ll n$  to be a valid procedure in the context of our results. However, as it is well documented in the literature (cf. Bickel et al. (1997)), the accuracy of  $m$ -out-of- $n$  bootstrap is typically much less than the standard bootstrap method (with  $m = n$ ). This is also supported in our case by the simulation results on variance estimation reported in Section 6.3.

### 3.2 Bootstrap bias and variance estimation

In this section, we show that not only does the modified bootstrap method give a valid distributional approximation with probability one, it also produces strongly consistent estimators of the asymptotic bias and variance of  $\mathbf{T}_n$ . From the work of Knight and Fu (2000), it follows that the limit distribution of  $\mathbf{T}_n$  has a nontrivial asymptotic bias when the the penalty parameter  $\lambda_n$  satisfies Assumption (C.2) with a  $\lambda_0 \neq 0$ . Also, as pointed out earlier, existing methods of estimating the variance matrix of the Lasso estimator have limitations when one or more components of the regression parameter  $\beta$  are zero. However, as the following result shows, the modified bootstrap method produces a consistent estimator of the bias and the variance matrix for all values of  $\beta$ .

**Theorem 3.3** (BIAS AND VARIANCE CONSISTENCY). *Under the assumptions of Theorem 3.1,*

$$\begin{aligned} \mathbf{E}_*(\mathbf{T}_n^{**}) &\rightarrow \mathbf{E}(\mathbf{T}_\infty), \quad \text{and} \\ \left(\mathbf{Var}_*(\mathbf{T}_n^{**})\right)_{p \times p} &\rightarrow \left(\mathbf{Var}(\mathbf{T}_\infty)\right)_{p \times p}, \end{aligned} \tag{3.3}$$

*with probability 1.*

Note that for  $\lambda_0 = 0$  in Assumption (C.2), the centered and scaled Lasso estimator has the same limit distribution as the centered and scaled least squares estimator, and therefore, in this case, the Lasso estimator is asymptotically unbiased. However, for  $\lambda_0 \neq 0$  in Assumption (C.2),  $\mathbf{T}_n$  is no longer guaranteed to be asymptotically unbiased. Thus, estimation of the asymptotic bias is an important problem in the context of penalized regression. Since the modified bootstrap produces consistent estimators of the bias and variance of  $\mathbf{T}_n$ , Theorem 3.3 allows one to attach a mean squared error estimate to the Lasso estimate and quantify the associated uncertainty, for all possible values of  $\boldsymbol{\beta}$ , and thereby removes the limitations of the existing methods of mean squared error estimation.

## 4 Bootstrapping the Adaptive Lasso estimator

Zou (2006) introduced the Adaptive Lasso method in the literature on penalized regression for simultaneous variable selection and estimation of the non-zero parameters in the regression model (1.1). Let  $\bar{\boldsymbol{\beta}}_n$  denote a generic preliminary estimator of  $\boldsymbol{\beta}$ , such as the ordinary least-squares (OLS) estimator of  $\boldsymbol{\beta}$ , given by

$$\bar{\boldsymbol{\beta}}_n = \left[ \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \sum_{i=1}^n \mathbf{x}_i y_i.$$

Then the Adaptive Lasso estimator (or ALASSO) estimator of  $\boldsymbol{\beta}$  is defined as

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

where  $\bar{\boldsymbol{\beta}}_n = (\bar{\beta}_{1,n}, \dots, \bar{\beta}_{p,n})$ ,  $\lambda_n \geq 0$  is the penalty and  $\gamma > 0$ . Let  $\check{\beta}_{j,n}$  denote the  $j$ th component of  $\check{\boldsymbol{\beta}}_n$  and let  $B_n = \{j : 1 \leq j \leq p, \check{\beta}_{j,n} = 0\}$  and assume w.l.g  $\{j : \beta_j = 0\} = \{p_0 + 1, \dots, p\}$ . Under some mild regularity conditions, Zou (2006) established the *oracle property* of the ALASSO method:

$$\begin{aligned} \mathbf{P}(B_n = \{p_0 + 1, \dots, p\}) &\rightarrow 1, \quad \text{as } n \rightarrow \infty, \\ \sqrt{n} \left( \check{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}^{(1)} \right) &\xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{C}_{11}); \end{aligned}$$

where  $\mathbf{C}_{11}$  is a  $p_0 \times p_0$ -submatrix of  $\mathbf{C}$  corresponding to the first  $p_0$  rows and  $p_0$  columns and where  $\check{\boldsymbol{\beta}}_n^{(1)}$  and  $\boldsymbol{\beta}^{(1)}$  is the vector of the first  $p_0$ -components of  $\check{\boldsymbol{\beta}}_n$  and  $\boldsymbol{\beta}$  respectively. Thus, the ALASSO estimator correctly identifies the zero-components of  $\boldsymbol{\beta}$  with probability tending to 1 and has the same limit distribution as the OLS estimator of the non-zero components of  $\boldsymbol{\beta}$ . We now study properties of a suitable residual bootstrap method for the ALASSO estimator.

## 4.1 A residual bootstrap method for the Adaptive Lasso

Let  $\check{\beta}_n$  denote the ALASSO estimator defined by (4.1). Define the ALASSO 'residuals' and their centered versions as

$$\check{\epsilon}_{1,i} = y_i - \mathbf{x}'_i \check{\beta}_n, \quad \text{and} \quad \check{\epsilon}_i = \check{\epsilon}_{1,i} - n^{-1} \sum_{j=1}^n \check{\epsilon}_{1,j}, \quad (4.2)$$

for all  $1 \leq i \leq n$ . Let  $\{\epsilon_1^+, \dots, \epsilon_n^+\}$  be a random sample of size  $n$  drawn with replacement from  $\{\check{\epsilon}_1, \dots, \check{\epsilon}_n\}$ . Define the ALASSO-based residual bootstrap (ARB) variables

$$y_i^+ = \mathbf{x}'_i \check{\beta}_n + \epsilon_i^+, \quad 1 \leq i \leq n.$$

The ARB version of the ALASSO-estimator is now defined as

$$\check{\beta}_n^+ = \underset{\mathbf{u} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n (y_i^+ - \mathbf{x}'_i \mathbf{u})^2 + \lambda_n \sum_{j=1}^p \frac{|u_j|}{|\check{\beta}_{j,n}^+|^\gamma}, \quad (4.3)$$

where  $\bar{\beta}_n^+ = (\bar{\beta}_{1,n}^+, \dots, \bar{\beta}_{j,n}^+)$  and  $\bar{\beta}_n^+$  is defined by replacing  $\{y_1, \dots, y_n\}$  in the definition of  $\bar{\beta}_n$  with  $\{y_1^+, \dots, y_n^+\}$ .

**Remark 5:** In contrast to the LASSO, the ARB does not required an additional truncation (*i.e.* hard thresholding) step. This is due to the form of the penalty term in the definition of the ALASSO that already incorporates a built-in soft-thresholding for the zero components. Thus, using the hard-thresholding step for the ALASSO does not change its asymptotic validity, but it may affect the finite sample performance.

## 4.2 Main results

Next we study consistency of the ARB for the ALASSO. For concreteness, we shall suppose that  $\bar{\beta}_n$  is the OLS of  $\beta$  and therefore

$$\bar{\beta}_n^+ = \left[ \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right]^{-1} \sum_{i=1}^n \mathbf{x}_i y_i^+.$$

Let  $\widehat{H}_n(\cdot)$  denote the conditional cdf of the ARB version  $\mathbf{T}_n^+ \equiv \sqrt{n} (\check{\beta}_n^+ - \check{\beta}_n)$  of the centered and scaled ALASSO estimator  $\check{\mathbf{T}}_n \equiv \sqrt{n} (\check{\beta}_n - \beta)$ . Then we have the following result:

**Theorem 4.1.** *Suppose that Assumptions (C.1) and (C.3) hold. Also suppose that*

$$\frac{\lambda_n}{\sqrt{n}} \rightarrow 0, \quad \text{and} \quad \lambda_n n^{(\gamma-1)/2} \rightarrow \infty. \quad (4.4)$$

Then,

$$\varrho\left(\widehat{H}_n, H_n\right) \xrightarrow{P} 0, \quad \text{as } n \rightarrow \infty, \quad (4.5)$$

where  $H_n(x) = \mathbf{P}\left(\check{\mathbf{T}}_n \leq x\right)$ ,  $x \in \mathbb{R}$  and  $\varrho$  is as in Theorem 3.1.

A proof of Theorem 4.1 is given in Section 8. Thus, it follows that under the conditions of Theorem 3.1, the ARB provides a valid approximation to the distribution of the ALASSO estimator. The in-probability convergence in (4.5) can be strengthened to almost-sure convergence under a stronger version of (4.4), which we do not pursue here. Indeed, (4.5) is adequate for constructing valid confidence intervals for  $\boldsymbol{\beta}$  based on the ALASSO estimator  $\check{\boldsymbol{\beta}}_n$ , as shown below. Let  $\check{t}_n(\alpha)$  denote the  $\alpha$ -quantile of the bootstrap distribution of  $\|\check{\mathbf{T}}_n^+\|$ ,  $\alpha \in (0, 1)$ . Define

$$\check{I}_{n,\alpha} \equiv \{\mathbf{t} \in \mathbb{R}^p : \|\mathbf{t} - \check{\boldsymbol{\beta}}_n\| \leq n^{-1/2}\check{t}_n(\alpha)\},$$

the level- $\alpha$  ARB confidence set for  $\boldsymbol{\beta}$ . Then, we have the following result:

**Corollary 4.2** (ARB CONFIDENCE SET). *Suppose that the assumptions of Theorem 4.1 hold. Let  $\alpha \in (0, 1)$  be such that  $\mathbf{P}\left(\|\check{\mathbf{T}}_\infty\| \leq \check{t}(\alpha) + \eta\right) > \alpha$  for all  $\eta > 0$ . Then,*

$$\mathbf{P}\left(\boldsymbol{\beta} \in \check{I}_{n,\alpha}\right) \rightarrow \alpha, \quad \text{as } n \rightarrow \infty. \quad (4.6)$$

Further, (4.6) holds for all  $\alpha \in (0, 1)$ , if  $p_0 \geq 1$ .

Thus, the ARB confidence sets attain the nominal coverage probability, asymptotically. A similar argument shows that under the assumptions of Theorem 4.1, for any linear combination  $\mathbf{c}'\boldsymbol{\beta}$ ,  $\mathbf{c} = (c_1, \dots, c_p)'$ ,  $\mathbf{c} \neq \mathbf{0}$ , the ARB confidence interval for  $\theta = \mathbf{c}'\boldsymbol{\beta}$ , is given by

$$I_{n,\alpha}(\theta) \equiv \left\{ \mathbf{x} \in \mathbb{R} : |\mathbf{x} - \mathbf{c}'\check{\boldsymbol{\beta}}_n| \leq n^{-1/2}\check{s}_n(\alpha) \right\}, \quad (4.7)$$

attains the target level  $\alpha$  for any  $\alpha \in (0, 1)$ , provided  $p_0 \geq 1$  and  $c_j \neq 0$  for some  $1 \leq j \leq p_0$ . Here,  $\check{s}_n(\alpha)$  is the  $\alpha$ -quantile of the bootstrap variable  $|\mathbf{c}'\check{\mathbf{T}}_n^+|$ . Note that the ARB confidence interval  $I_{n,\alpha}(\theta)$  in (4.7) does not require the explicit estimation of the standard error of the ALASSO estimator  $\mathbf{c}'\check{\boldsymbol{\beta}}_n$ .

We briefly investigate the accuracy of ARB confidence intervals in finite samples in Section 6.4. As in the case of the LASSO, estimation of the mean squared error of the ALASSO estimator is a nontrivial problem, particularly for the estimators of the zero components (cf. Zou (2006)). The next result shows that the ARB provides a consistent estimator of the mean squared error matrix of the scaled ALASSO estimator  $\check{\boldsymbol{\beta}}_n$ , given

by  $\text{MSE}(\check{\mathbf{T}}_n) \equiv n\mathbf{E}(\check{\boldsymbol{\beta}}_n - \boldsymbol{\beta})(\check{\boldsymbol{\beta}}_n - \boldsymbol{\beta})'$ .

**Corollary 4.3.** *Suppose that the assumptions of Theorem 4.1 hold. Then*

$$\text{MSE}_*(\check{\mathbf{T}}_n^+) - \text{MSE}(\check{\mathbf{T}}_n) \xrightarrow{P} 0, \quad \text{as } n \rightarrow \infty.$$

## 5 Data-based choice of the regularization parameter for the Lasso estimator

### 5.1 The optimal regularization parameter

Here, we consider the problem of choosing the optimal penalty parameter  $\lambda_0$  for the Lasso estimator, in a data-dependent manner. Firstly, we formalize the notion of the optimal parameter through a natural reparametrization and subsequently, we describe a data-based method for choosing the optimal regularization parameter based on the modified bootstrap method. Let

$$V(\mathbf{u}) = -2\mathbf{u}'\mathbf{W} + \mathbf{u}'\mathbf{C}\mathbf{u} + \lambda_0 \sum_{j=1}^p \left\{ u_j \text{sgn}(\beta_j) \mathbf{1}(\beta_j \neq 0) + |u_j| \mathbf{1}(\beta_j = 0) \right\}, \quad (5.1)$$

where  $\mathbf{W} \sim N(\mathbf{0}, \sigma^2 \mathbf{C})$ ,  $\mathbf{C} = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$  and where  $\lambda_0 \in [0, \infty)$ . Here and in the following, we write  $\text{sgn}(x)$  to denote the sign of  $x \in \mathbb{R}$  and  $\mathbf{1}(\cdot)$  to denote the indicator function. Under Assumptions (C.1) - (C.3), the work of Knight and Fu (2000) implies that  $\mathbf{T}_n \rightarrow \mathbf{T}_\infty$  in distribution, where  $\mathbf{T}_\infty = \underset{\mathbf{u}}{\text{argmin}} V(\mathbf{u})$ . Thus, the limit distribution of  $\mathbf{T}_n$  depends on  $\lambda_n$  only through  $\lambda_0$ . We now reparametrize  $\lambda_n \in [0, \infty)$  and write it as  $\lambda_n = \lambda_0 n^{1/2}$ ,  $\lambda_0 \in [0, \infty)$ . Note that MSE of  $\hat{\boldsymbol{\beta}}_n$  for estimating can be expressed as  $\text{MSE}(\hat{\boldsymbol{\beta}}_n) \equiv \mathbf{E} \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|^2 = n^{-1} \mathbf{E} \|\mathbf{T}_n\|^2$ . Since the effect of the penalization by  $\lambda_n$  on the overall accuracy of  $\hat{\boldsymbol{\beta}}_n$  is reflected by its MSE-function and since  $n \times \text{MSE}(\hat{\boldsymbol{\beta}}_n)$  converges to the MSE of the limiting random variable  $\mathbf{T}_\infty$ , we define the optimal penalization parameter  $\lambda_0^{opt}$  as

$$\lambda_0^{opt} \equiv \text{argmin } \phi(\lambda_0), \quad (5.2)$$

where  $\phi(\lambda_0) = \mathbf{E} \|\mathbf{T}_\infty\|^2$ , the MSE of the limit distribution of  $\mathbf{T}_n$  with  $\lambda_n = \lambda_0 n^{1/2}$ ,  $\lambda_0 \in [0, \infty)$ . Thus, choosing  $\lambda_0 = \lambda_0^{opt}$  in the reparametrization yields a Lasso estimator that minimizes the MSE in large samples. Our goal is to estimate the target parameter  $\lambda_0^{opt}$ .

## 5.2 Data-based selection of the optimal regularization parameter

We now describe a data-based method for estimating  $\lambda_0^{opt}$  of (5.2) using the modified bootstrap. For any choice of penalization parameter  $\lambda_0$  and thresholding value  $a$ , rewrite  $\mathbf{T}_n^{**} \equiv \mathbf{T}_n^{**}(\lambda_0, a)$ . The modified bootstrap based estimator of  $\phi(\lambda_0)$  is

$$\widehat{\phi}_n(\lambda_0, a) \equiv \mathbf{E}_* \|\mathbf{T}_n^{**}(\lambda_0, a)\|^2, \quad \lambda_0 \in [0, \infty), \quad a \in (0, \infty), \quad (5.3)$$

where  $a$  denotes the thresholding value used in defining (2.4). Note that by Theorem 3.1,  $\widehat{\phi}_n(\lambda_0, a_n)$  is a strongly consistent estimator of  $\phi(\lambda_0)$ , for an appropriate choice of thresholding values  $\{a_n\}$  satisfying (2.3). Therefore, we replace  $\phi(\lambda_0)$  in (5.2) by its bootstrap estimator  $\widehat{\phi}_n(\lambda_0, a)$  and define the bootstrap estimator of the target penalization parameter  $\lambda_0^{opt}$  by

$$\lambda_{0,n}^* = \operatorname{argmin}_{\lambda_0, a} \widehat{\phi}_n(\lambda_0, a). \quad (5.4)$$

If we fix the value of the thresholding parameter  $a$ , we can minimize over  $\lambda_0$  and define the optimal penalty parameter at a fixed choice of  $a$ :

$$\lambda_{0,n}^*(a) = \operatorname{argmin}_{\lambda_0} \widehat{\phi}_n(\lambda_0, a). \quad (5.5)$$

Similarly, we can define

$$a_n^* = \operatorname{argmin}_{\lambda_0, a} \widehat{\phi}_n(\lambda_0, a), \quad (5.6)$$

which provides the value of  $a$  where  $\widehat{\phi}_n$  is minimized. Since the modified bootstrap provides a consistent approximation to the MSE of the Lasso estimator (cf. Theorem 3.3), (5.4) defines an accurate estimator of the optimal penalization parameter  $\lambda_0^{opt}$  in large samples. However, the performance of this estimator in finite samples depends on various factors, including the number of zero components of the true parameter value  $\beta$  and the sizes of its non-zero components and quite importantly, on the choice of the thresholding value  $a_n$  in (2.4). We now suggest a data-based method of selecting a thresholding value.

## 5.3 Jackknife-after-bootstrap based choice of thresholding parameter

We suggest using the Jackknife-after-bootstrap (JAB) method (see Efron (1992) for details on the JAB method) for selecting an appropriate thresholding value. The main idea is to construct an estimate of the error of the bootstrap estimator  $\widehat{\phi}_n(\lambda_0, a)$  using the JAB method and select a thresholding value  $a$  that minimizes this error. For any fixed choice of the pair  $(\lambda_0, a)$ , once we have obtained the bootstrap samples

$\{T_n^{**}(b : \lambda_0, a) : b = 1, \dots, B\}$ , we can construct delete-1 JAB replicates of  $\widehat{\phi}_n(\lambda_0, a)$ :

$$\widehat{\phi}_{n,j}(\lambda_0, a) = \frac{1}{|I_j|} \sum_{b \in I_j} \|T_n^{**}(b : \lambda_0, a)\|^2, \quad j = 1, \dots, n,$$

where

$$I_j = \{b : 1 \leq b \leq B, \{e_{1,b}^{**}, \dots, e_{n,b}^{**}\} \text{ does not contain the residual } r_j - \bar{r}_n\},$$

with residuals  $r_j$  as defined in (2.5) and  $|A|$  denoting the cardinality of a set  $A$ . The EMSE of the bootstrap estimator  $\widehat{\phi}_n$ , based on the JAB replicates will be

$$\begin{aligned} \text{EMSE}(\widehat{\phi}_n(\lambda_0, a)) &= \left[ \frac{1}{n-1} \sum_{j=1}^n \{\widehat{\phi}_{n,j}(\lambda_0, a) - \widehat{\phi}_n(\lambda_0, a)\} \right]^2 \\ &\quad + \frac{1}{n(n-1)} \sum_{j=1}^n \{\widehat{\phi}_{n,j}(\lambda_0, a) - \widehat{\phi}_n(\lambda_0, a)\}^2. \end{aligned} \quad (5.7)$$

The best choice of thresholding value is

$$\widehat{a} = \underset{\lambda_0, a}{\text{argmin}} \text{EMSE}(\widehat{\phi}_n(\lambda_0, a)). \quad (5.8)$$

It is to be noted that the above best choice of  $a$  is evaluated over all possible choices of  $\lambda_0$  and  $a$  values, and aims to minimize a completely different optimization criterion compared to the previously defined bootstrap based optimal values  $\lambda_{0,n}^*$  and  $a_n^*$  (cf. (5.4) and (5.6)). We can also assess the performance of  $\widehat{a}$  in a simulation setup in terms of how closely it approximates the overall minimum value of  $\widehat{\phi}_n$ , realized at the coordinates  $(\lambda_{0,n}^*, a_n^*)$ . Thus, an accuracy measure (AM) for  $\widehat{a}$  can be defined as

$$\text{AM}(\widehat{a}) \equiv \frac{|\widehat{\phi}_n(\lambda_{0,n}^*, a_n^*) - \widehat{\phi}_n(\lambda_{0,n}^*, \widehat{a})|}{\phi(\lambda_0^{opt})}, \quad (5.9)$$

with scaling done by the value at the true minimum. Such a measure of accuracy is obviously not unique, and other accuracy criterion can be constructed for this purpose. Other alternatives can be the (average) distance  $|a_n^* - \widehat{a}|$  or its square.

**Remark 6:** The JAB based choice of the optimum thresholding value,  $\widehat{a}$  defined in (5.8) aims to optimize the MSE of the bootstrap estimator  $\widehat{\phi}_n$ . On the other hand,  $a_n^*$ , defined in (5.6) aims to optimize the value of the bootstrap estimator  $\widehat{\phi}_n$  as a function of  $a$ . These two choices:  $\widehat{a}$  and  $a_n^*$ , offer independent ways to choose a thresholding parameter for any specific data-set.

## 6 Numerical results

### 6.1 Modified bootstrap based choice of optimal penalization and JAB based choice of thresholding value

In this numerical study we explore the performance of the modified bootstrap based method of choosing the optimal  $\lambda_0$  and also the JAB based method of choosing a thresholding value, as described in the Section 5.

For our simulation study we considered a fixed design matrix with elements of the design matrix selected independently from the standard normal distribution. The errors were independent standard normal. Three different cases, with different choices of the true  $\beta$  (cf. (1.1)) were considered:

(i)  $\beta = (2, 5, 0, -1, 6, 0, 0, 0, -3, 10)'$ . This corresponds to the case where there are no 'small' non-zero coefficients.

(ii)  $\beta = (4, -0.25, 0, 0.35, 1, 0, 0, 0, -2, 0.65)'$ . This case represents a situation where there are some small non-zero coefficients.

(iii)  $\beta = (n^{-1/2}, -3n^{-1/2}, 0, 0.75, 4, 0, 0, 0, 5n^{-1/2}, -1.5)'$ . This represents a case where the underlying regression coefficient is no longer 'fixed' and depends on the sample size  $n$ .

In all cases the dimension of the regression coefficient  $p$  ( $= 10$ ), is kept fixed. In all cases the sample size was fixed at  $n = 250$ . The errors were selected independently from the standard normal distribution.

For any choice of a design matrix, true regression parameter and error distribution, the true optimal  $\lambda_0$  defined in (5.2) is intractable and hence is replaced by a Monte-Carlo based approximation. We decided to compare the performances of the modified bootstrap based choice of optimal  $\lambda_0$  (cf. (5.4)), an analogous estimate based on the usual (naive) residual bootstrap and also an estimate based on the cross-validation (CV) technique. It should be noted that, a CV based choice of optimal  $\lambda_0$  tries to minimize the prediction error, while the criterion function used in (5.2) minimizes the MSE. It can be shown that, except under very special situations, the optimal  $\lambda_0$  values under these two different criterion are incomparable. We also looked at the behavior of  $\lambda_{0,n}^*(a)$  (cf. (5.5)), which is the modified bootstrap based optimum choice of  $\lambda_0$ , at a fixed value of the thresholding parameter  $a$ .

Figures 1-2 show the relative behavior of the modified bootstrap based optimal  $\lambda_0$ , the corresponding naive residual bootstrap based optimal  $\lambda_0$  and the CV based optimal  $\lambda_0$  for the models in Cases (i) and (ii) respectively. The vertical solid line in each figure shows the true optimal  $\lambda_0$  value (cf. (5.2)). In all these cases, the CV based choices are nowhere near the true optimal  $\lambda_0$ , as is to be expected. In Case (i), when  $\beta$  does not have any 'small' components, the modified bootstrap based choice of optimal  $\lambda_0$  is far better than

the naive bootstrap. On the other hand, in Case (ii) the naive bootstrap apparently performs better than modified bootstrap. This is despite the fact that the naive bootstrap based estimator of  $\beta$  (appropriately scaled and centered) fails to converge to any fixed distribution. If we investigate the behavior of the modified bootstrap based estimate  $\lambda_{0,n}^*(a)$  (cf. (5.5)) at certain fixed choices of  $a$ , we find that a different picture emerges. At certain choices of  $a$ , the modified bootstrap performs far better than what is suggested by Figure 2. In fact, for the models in Cases (ii) and (iii), and as shown in Figures 3 and 4, at certain choices of  $a$ , the estimator  $\lambda_{0,n}^*(a)$  actually approximates the true underlying optimal  $\lambda_0$  value very accurately, better than naive bootstrap and also better than  $\lambda_{0,n}^*$ . This apparent anomaly arises because of the choice of  $a$  values, over which the optimum  $\lambda_{0,n}^*$  value (cf. (5.4)) is computed, specifically the larger values of  $a$  contributing to the failure of  $\lambda_{0,n}^*$ . Specifically for the choices of  $\beta$  in Cases (ii) and (iii), we find that using  $a = 0.25$  and  $a = 0.005$ , provide us a much better approximation to the true optimal  $\lambda_0$ .

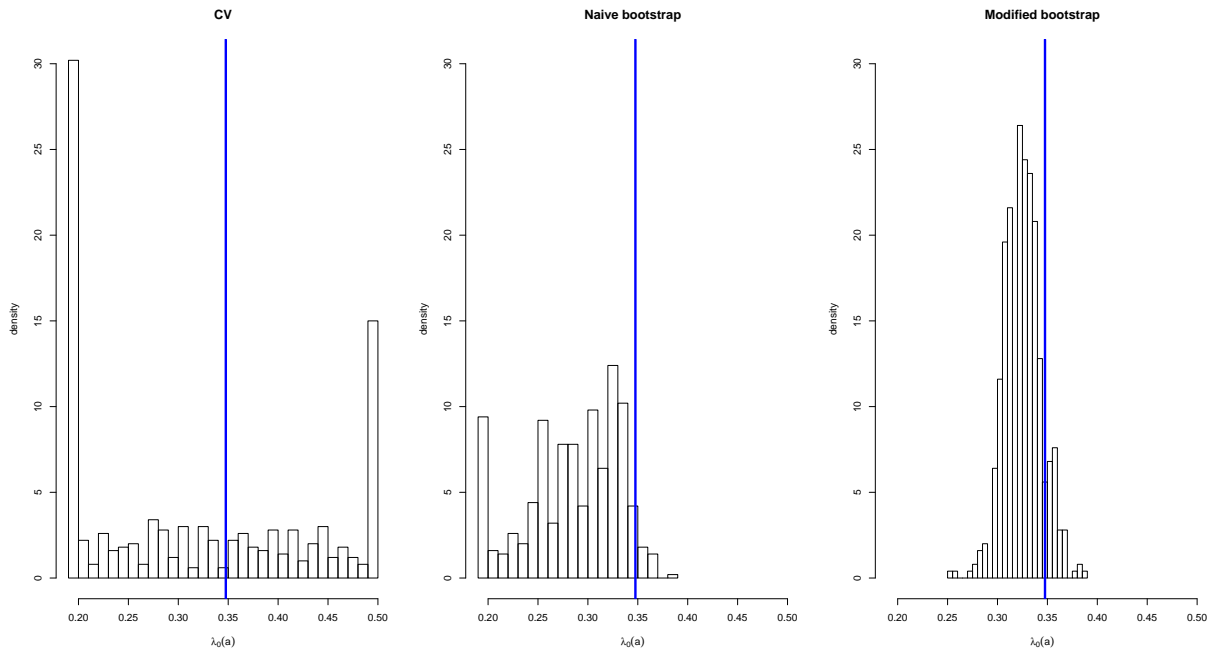


Figure 1: Histogram showing the performance of CV based (at left), naive bootstrap (at center) and modified bootstrap (at right) based choices of optimal  $\lambda_0$ . The vertical dotted line shows the true optimal value  $\lambda_0^{opt}$ . The true  $\beta$  corresponds to Case (i) and  $n = 250$ .

Next consider the impact of choosing the thresholding parameter  $a$  in finite samples. In all the three cases, we had selected a grid of six  $a$  values, over which the optimum values  $a^*$  and  $\hat{a}$  were computed. In all three cases, over a set of 500 simulations, we found the average absolute and squared distances between these two values. The results are presented in Table 1:

The results in Table 1 show that the accuracy of the JAB based estimate of  $a$  is very good for the model

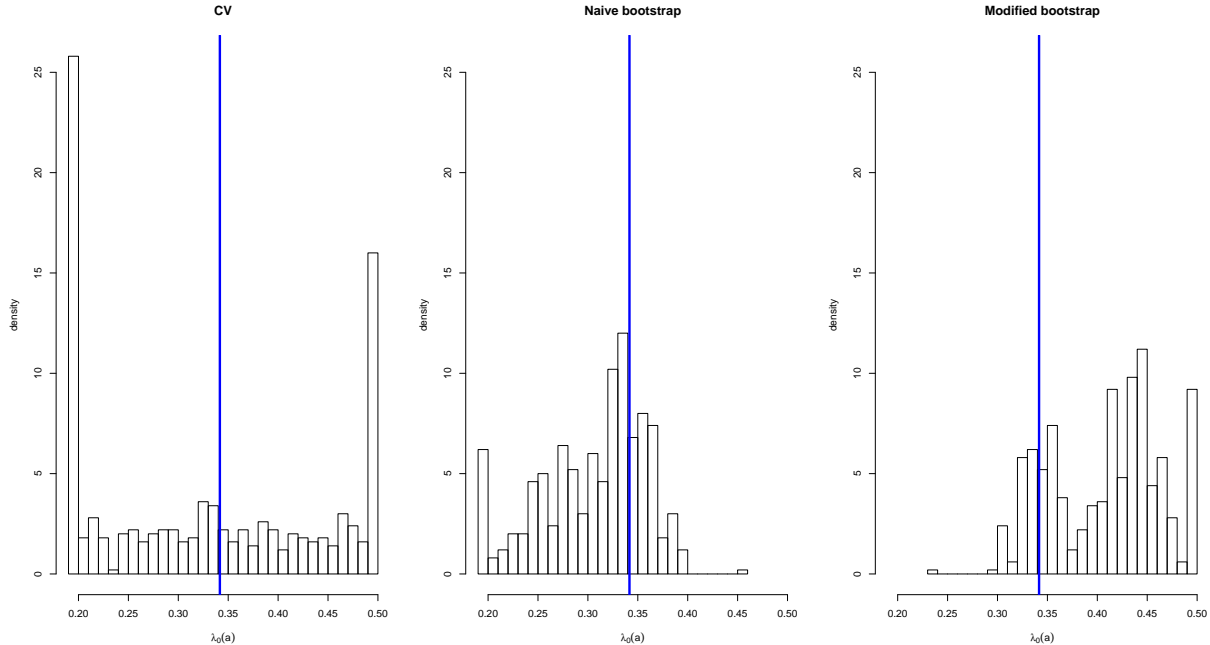


Figure 2: Histogram showing the performance of CV based (at left), naive bootstrap (at center) and modified bootstrap (at right) based choices of optimal  $\lambda_0$ . The vertical dotted line shows the true optimal value  $\lambda_0^{opt}$ . The true  $\beta$  corresponds to Case (ii) and  $n = 250$ .

Table 1: Accuracy of the JAB based choice  $\hat{a}$

Case	<sup>d</sup> Average squared distance $(\hat{a} - a_n^*)^2$	<sup>d</sup> Average absolute distance $ \hat{a} - a_n^* $
(i)	0.0034	0.0975
(ii)	0.0371	0.1539
(iii)	0.1052	0.2376

<sup>d</sup> over 500 simulations

used in Case (i) and deteriorates marginally for the other models. In finite samples, the choice of the grid of thresholding values over which the optimum  $a$  is computed is very important. The simulation results in Table 1 suggest that the JAB method is usable, at least in a situation where the underlying  $\beta$  has distinct (and large) non-zero and zero coefficients.

## 6.2 Coverage accuracy of confidence regions

In this numerical study, we compare the finite sample performance of the confidence regions for  $\beta$  obtained by using the modified bootstrap procedure and the naive (usual) residual bootstrap procedure. Table 2

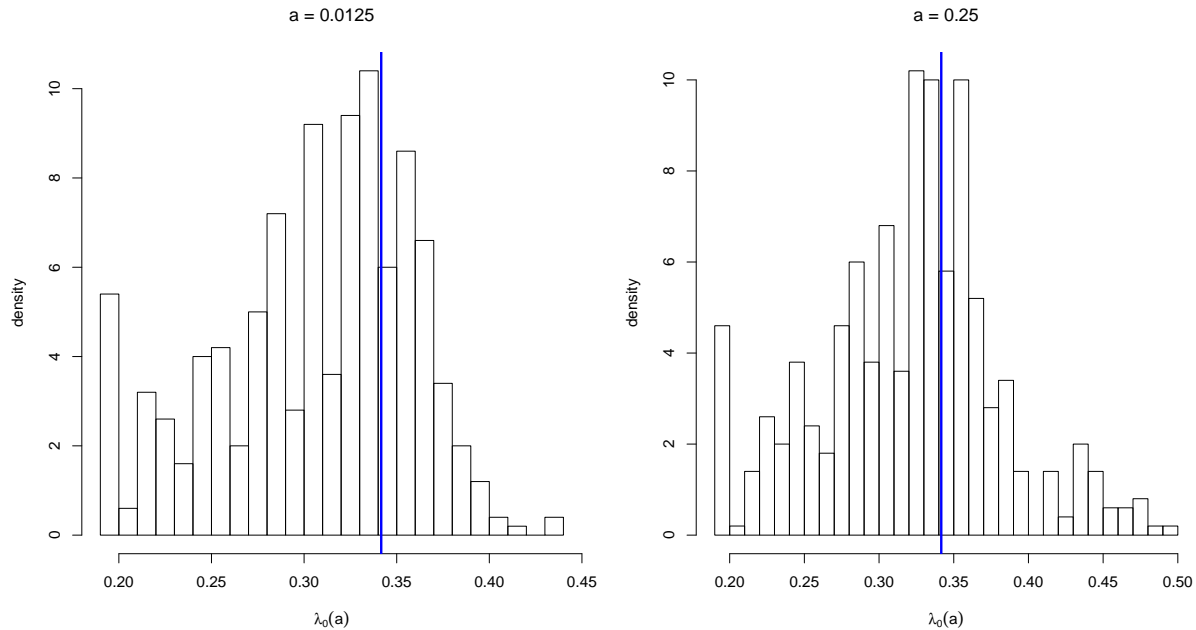


Figure 3: Histogram showing the performance of modified bootstrap based choices of optimal  $\lambda_0$  at two fixed thresholding values:  $a = 0.0125$  and  $0.25$ . The vertical dotted line shows the true optimal value  $\lambda_0^{opt}$ . The true  $\beta$  corresponds to Case (ii) and  $n = 250$ .

shows the empirical coverage probabilities for the modified bootstrap based confidence region  $I_{n,\alpha}$ , defined in (3.1) for  $\alpha = 0.9$  and also for an analogous naive bootstrap based confidence set (constructed similarly, but without the thresholding step).

Tables 2-4 shows that both the modified bootstrap and naive bootstrap methods performs exceedingly well in terms of achieving the desired nominal coverage rate, for all the three models in Cases (i) - (iii). Although, as the threshold value  $a$  increases, the modified bootstrap confidence regions show much higher than nominal coverage (except for Case (iii)), suggesting that these choices of  $a$  are not suitable. Apparently, the inconsistency of the naive bootstrap does not have any effect on the coverage accuracy of the confidence regions. The results in Tables 2-4 represent the pointwise coverage accuracy of the confidence intervals (3.1), and there is no assurance that there will be a similar level of accuracy in an uniform sense (see Remark 3). Although the naive bootstrap works well in all the above cases, the inherent inconsistency property of the naive residual bootstrap based Lasso estimator makes the use of naive bootstrap based confidence intervals a very doubtful choice.

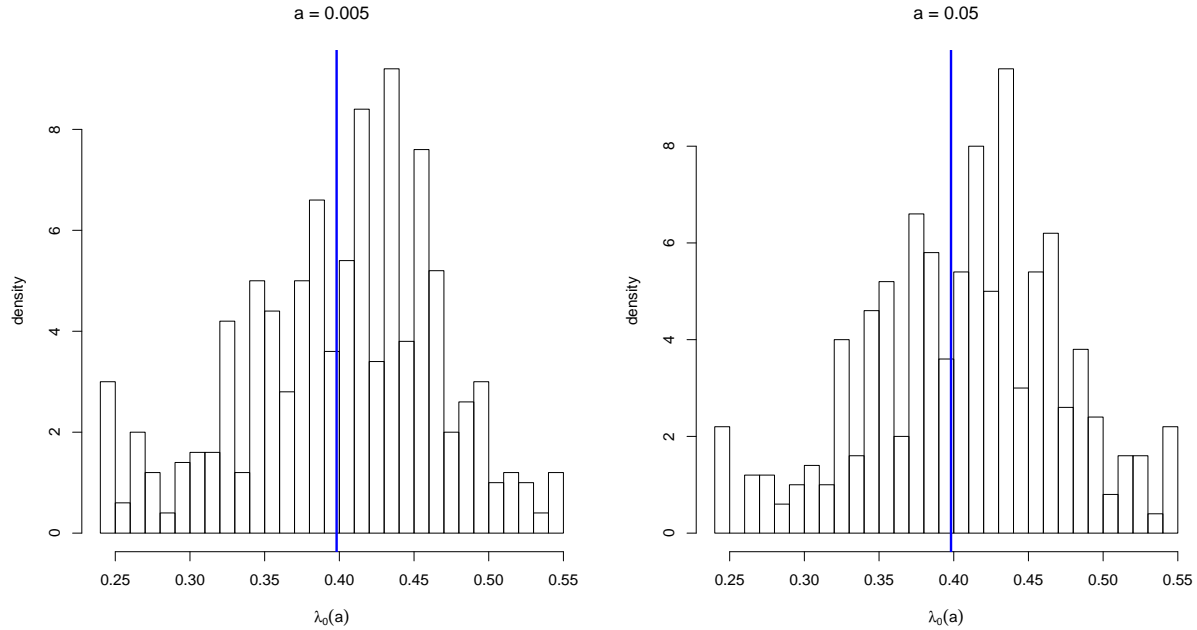


Figure 4: Histogram showing the performance of modified bootstrap based choices of optimal  $\lambda_0$  at two fixed thresholding values:  $a = 0.005$  and  $0.05$ . The vertical dotted line shows the true optimal value  $\lambda_0^{opt}$ . The true  $\beta$  corresponds to Case (iii) and  $n = 250$ .

Table 2: Empirical coverage probabilities for 90% confidence regions based on the modified and naive bootstrap procedures for the model in Case (i), with  $n = 250$ .

Empirical coverage probabilities for 90% confidence regions					
$\lambda_0$	Naive Bootstrap	Modified Bootstrap			
		$a$			
		0.125	0.25	0.75	1.25
0.1975	0.880	0.880	0.877	0.877	0.997
0.2975	0.880	0.870	0.870	0.870	0.997
0.3475*	0.877	0.867	0.867	0.867	0.996
0.3975	0.877	0.867	0.863	0.863	0.996
0.4975	0.883	0.870	0.870	0.870	0.997

\* true optimal  $\lambda_0$

Table 3: Empirical coverage probabilities for 90% confidence regions based on the modified and naive bootstrap procedures for the model in Case (ii), with  $n = 250$ .

Empirical coverage probabilities for 90% confidence regions					
$\lambda_0$	Naive Bootstrap	Modified Bootstrap			
		$a$			
		0.0125	0.05	0.25	0.75
0.1916	0.860	0.860	0.860	0.867	0.977
0.2916	0.867	0.860	0.860	0.863	0.980
0.3416*	0.863	0.863	0.863	0.860	0.977
0.3916	0.870	0.870	0.867	0.860	0.970
0.4916	0.880	0.873	0.873	0.860	0.967

\* true optimal  $\lambda_0$

Table 4: Empirical coverage probabilities for 90% confidence regions based on the modified and naive bootstrap procedures for the model in Case (iii), with  $n = 250$ .

Empirical coverage probabilities for 90% confidence regions					
$\lambda_0$	Naive Bootstrap	Modified Bootstrap			
		$a$			
		0.005	0.0125	0.05	0.25
0.2481	0.890	0.887	0.887	0.887	0.890
0.3481	0.893	0.893	0.893	0.893	0.877
0.3981*	0.890	0.893	0.893	0.890	0.873
0.4481	0.887	0.893	0.893	0.887	0.867
0.5481	0.893	0.893	0.893	0.890	0.850

\* true optimal  $\lambda_0$

### 6.3 Variance estimation

The next numerical study focuses on the aspect of computing variance estimates for Lasso estimators. We study three approaches, the naive (residual) bootstrap, the modified bootstrap and also the  $m$ -out-of- $n$  bootstrap (see [Hall et al. \(2009\)](#)). For each of three cases under consideration, we picked particular sets of covariate pairs and compared the accuracy of covariance estimates obtained by these three methods. The true covariances for every selected covariate pair was approximated by Monte-Carlo simulations. For simplicity of presentation, we only present the covariance estimates and their MSE's only at a single value of  $\lambda_0$ .

Tables 5-7 show the results: for each selected choice of  $\beta$ , the covariate pairs were carefully chosen to represent cases where the underlying  $\beta_j$  is non-zero and 'large', non-zero and 'small', exactly zero, and combinations of such pairs. The Monte-Carlo estimate of the true covariance between these covariate pairs is shown in the second column of these tables. We used four different choices of the thresholding parameter  $a$  for the Modified bootstrap (MB) and three different choices of the bootstrap sample size  $m$ , for the  $m$ -out-of- $n$  bootstrap sample size. The sample size used is  $n = 250$ .

From Table 5 we see that for the choice of  $\beta$  in Case (i), the Naive bootstrap (NB) and Modified bootstrap (MB) have very comparable performances, except for the choice of  $a = 1.25$ , where the performance of the MB deteriorates. The  $m$ -out-of- $n$  bootstrap method has very poor performance when  $m$  is small, and gradually improves as  $m$  increases, although remains worse than the other two methods. Though NB and MB have comparable performances, the MB clearly outperforms the NB for the pair (3, 3), where  $\beta_3 = 0$ . It shows that the MB has some advantage over the NB, when the true underlying value of a component is zero. For the  $m$ -out-of- $n$  bootstrap case, the sample sizes  $m$  were chosen arbitrarily and there is always the option of selecting the optimum sample size for fine-tuning the performance.

We also computed a sandwich formula (cf. [Zou \(2006\)](#), [Fan and Peng \(2004\)](#)) based variance estimate for the Lasso estimate of  $\beta_2$ . We found that the (mean) variance estimate is 5.150 and the estimated MSE is 349.83. Although this suggests that the sandwich formula performs very badly, but a closer inspection revealed that due to the presence of very few extreme values, the MSE became inflated. In fact, when three extremely large values (out of a set of 500) are removed, the (mean) variance estimates becomes 4.05 and the estimated MSE is 0.044, which is much better than all of the other bootstrap based variance estimates as shown in Table 5.

Table 6 corresponds to the  $\beta$  in Case (ii). Once again, the NB method performs as well as the MB method, except for the last choice of the thresholding value  $a$ , where the performance of the MB deteriorates. The  $m$ -out-of- $n$  bootstrap method is worse than the other two methods, with marginally better performance with

Table 5: Comparison of variance estimates and the MSE's of the estimates for the model in Case (i) with  $\boldsymbol{\beta} = (2, 5, 0, -1, 6, 0, 0, 0, -3, 10)'$ : obtained by the naive (residual) bootstrap (NB), the modified bootstrap (MB) and the  $m$ -out-of- $n$  bootstrap techniques, at  $\lambda_0 = \lambda_0^{opt}$  and  $n = 250$ .

Variance estimates and their MSE's for different covariate pairs									
Pair: $(i, j)$	True value	NB		MB			$m$ -out-of- $n$		
		Estimate	MSE	$a^*$	Estimate	MSE	$m^{**}$	Estimate	MSE
(1, 5)	0.0238	-0.0394	0.0261	0.125	-0.0369	0.0247	50	-1.350	2.460
				0.250	-0.0372	0.0247	125	-0.207	0.148
				0.750	-0.0372	0.0247	200	0.195	0.0635
				1.250	-0.0579	0.0940			
(2, 2) <sup>a</sup>	4.1800	3.8600	0.280		3.91	0.239		20.40	267.00
					3.92	0.238		7.55	12.00
					3.92	0.238		4.62	0.446
					8.11	16.00			
(3, 3)	2.4800	2.6200	0.3770		2.39	0.219		5.25	10.20
					2.35	0.140		4.42	5.05
					2.35	0.140		3.22	1.13
					5.94	12.50			
(3, 7)	0.0109	0.0364	0.0091		0.0269	0.00825		-0.00276	0.0319
					0.0256	0.00797		-0.06650	0.0321
					0.0256	0.00797		0.04240	0.0145
					0.1030	0.05100			
(8, 10)	0.4140	0.3080	0.0280		0.297	0.0281		1.710	2.20
					0.295	0.0276		0.305	0.0622
					0.295	0.0276		0.310	0.0350
					0.720	0.1600			

\* same set of  $a$  values are used for each covariate pair.

\*\* same set of  $m$  values are used for each covariate pair.

<sup>a</sup> variance estimate based on the sandwich formula has been obtained for this pair

Table 6: Comparison of variance estimates and the MSE's of the estimates for the model in Case (ii) with  $\beta = (4, -0.25, 0, 0.35, 1, 0, 0, 0, -2, 0.65)'$ : obtained by the naive (residual) bootstrap (NB), the modified bootstrap (MB) and the  $m$ -out-of- $n$  bootstrap techniques, at  $\lambda_0 = \lambda_0^{opt}$  and  $n = 250$ .

Variance estimates and their MSE's for different covariate pairs									
Pair: $(i, j)$	True value	NB		$a^*$	MB		$m$ -out-of- $n$		
		Estimate	MSE		Estimate	MSE	$m^{**}$	Estimate	MSE
(1, 1) <sup>a</sup>	4.33	4.11	0.248	0.0125	4.11	0.236	50	25.80	470.0
				0.05	4.13	0.232	125	9.46	27.4
				0.25	4.13	0.225	200	5.33	1.3
				0.75	6.82	6.730			
(2, 2)	4.10	3.84	0.246		3.84	0.252		14.20	112.0
					3.85	0.248		7.22	10.6
					2.92	2.120		4.58	0.506
					4.08	0.0284			
(3, 3)	2.28	2.67	0.488		2.68	0.510		5.66	14.50
					2.62	0.492		4.55	6.51
					2.50	0.178		3.21	1.67
					4.48	5.180			
(2, 3)	-0.136	-0.139	0.0126		-0.127	0.0151		0.0924	0.162
					-0.124	0.0148		0.1530	0.130
					-0.078	0.0148		-0.220	0.0293
					-0.133	0.0246			
(6, 10)	0.00951	-0.00894	0.0130		-0.00446	0.0112		-0.0096	0.316
					-0.00418	0.0110		0.0407	0.0449
					-0.00565	0.0115		-0.0717	0.0255
					-0.00851	0.0212			

\* same set of  $a$  values are used for each covariate pair.

\*\* same set of  $m$  values are used for each covariate pair.

<sup>a</sup> variance estimate based on the sandwich formula has been obtained for this pair

increasing bootstrap sample size  $m$ . The interesting result aspect for this case arises for the pair  $(2, 2)$ , where  $\beta_2 = -0.25$ . In this case, with  $a = 0.25$ , the MB method fails drastically compared to other choices of  $a$  and also the NB. For the pair  $(6, 10)$ , the true covariance is very small, and the estimated values fail to have the correct sign, but the MB based estimates have relatively lower bias than the NB estimates.

As earlier, we computed the sandwich formula based variance estimate for the Lasso estimate of  $\beta_1$ . The (mean) variance estimate based on the sandwich method is 4.30 and the MSE is 0.090. Once again the sandwich estimate for the variance performs exceedingly well, compared to any of the bootstrap variance estimates for the the pair  $(1, 1)$ .

Table 7 shows the result for the choice of  $\beta$  in Case (iii). There is very little to choose from among the NB and MB methods, although the MB method shows slightly better MSE's for some covariate pairs. The  $m$ -out-of- $n$  bootstrap method still ranks behind the other two.

In this case, we looked at  $\beta_5$  and the sandwich formula based variance estimator. The (mean) variance estimate is 3.85 and the MSE is 0.26, which is comparable to the NB and MB bootstrap estimates. In this case only, the sandwich formula based variance estimate is unable to outperform the bootstrap based competitors. Although the sandwich formula has very good performance, it does not provide us any answers for the case when  $\beta_j = 0$ .

## 6.4 Numerical results with the Bootstrapped Adaptive Lasso estimator

In this section we conduct a small simulation to study the performance of the Adaptive Lasso residual bootstrap estimator (ARB) in context of variance estimation and interval estimation. We only consider the  $\beta$  described in Case (i). We study the accuracy of ARB based variance estimates of ALASSO components and also ARB based confidence intervals for some underlying non-zero  $\beta_j$ 's. From Table 8 we find that the ARB based variance estimator performs exceedingly well in estimating the true variance of the ALASSO estimator. The coverage accuracy of ARB based intervals is very close to the nominal coverage probability.

## 7 Prostrate cancer data example

In this section we study the performance of the modified Lasso estimator on the prostrate cancer data originally from [Stamey et al. \(1989\)](#), which has been used in [Tibshirani \(1996\)](#) and is also available from [this web page](#). In this clinical study, the variable of interest was log(prostrate specific antigen) (`lpsa`) and eight different predictors 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`),

Table 7: Comparison of variance estimates and the MSE's of the estimates for the model in Case (iii) with  $\beta = (n^{-1/2}, -3n^{-1/2}, 0, 0.75, 4, 0, 0, 0, 5n^{-1/2}, -1.5)'$ : obtained by the naive (residual) bootstrap (NB), the modified bootstrap (MB) and the  $m$ -out-of- $n$  bootstrap techniques, at  $\lambda_0 = \lambda_0^{opt}$  and  $n = 250$ .

Variance estimates and their MSE's for different covariate pairs									
Pair: $(i, j)$	True value	NB		MB			$m$ -out-of- $n$		
		Estimate	MSE	$a^*$	Estimate	MSE	$m^{**}$	Estimate	MSE
(1, 1)	2.87	2.60	0.558	0.005	2.61	0.559	50	5.14	9.95
				0.0125	2.61	0.560	125	3.97	2.92
				0.05	2.55	0.640	200	3.040	0.827
				0.25	2.21	0.555			
(2, 4)	0.164	0.295	0.0388		0.310	0.0417		1.45	2.21
					0.310	0.0417		0.76	0.44
					0.311	0.0419		0.478	0.13
					0.198	0.0168			
(2, 6)	-0.107	-0.0918	0.0107		-0.0880	0.0110		-0.320	0.129
					-0.0879	0.0110		-0.381	0.120
					-0.0856	0.0108		-0.172	0.0207
					-0.0446	0.0104			
(3, 3)	2.20	2.43	0.420		2.43	0.395		3.77	4.08
					2.43	0.395		3.79	3.71
					2.38	0.391		2.94	1.15
					2.31	0.127			
(5, 5) <sup>a</sup>	3.980	3.71	0.220		3.72	0.213		16.10	150.0
					3.72	0.213		7.73	14.8
					3.74	0.205		4.55	0.541
					3.97	0.173			

\* same set of  $a$  values are used for each covariate pair.

\*\* same set of  $m$  values are used for each covariate pair.

<sup>a</sup> variance estimate based on the sandwich formula has been obtained for this pair

Table 8: Variance estimates and MSE's based on the ARB and coverage accuracy of ARB based (90%) confidence intervals for some non-zero components of the underlying  $\beta$ . The results are for the model in Case (i) with  $\beta = (2, 5, 0, -1, 6, 0, 0, 0, -3, 10)'$ . Here  $n = 250$  and  $\lambda_n = n^{1/3}$  and  $\gamma = 1$  (cf. (4.1)).

Variance estimation and interval estimation based on the ARB				
Component ( $j$ )	Variance estimation			Empirical coverage
	$\mathbf{Var}(\check{\beta}_{n,j}) \times 10^3$	$\mathbf{Var}_*(\check{\beta}_{n,j}^+) \times 10^3$	MSE ( $\times 10^{-7}$ )	
$\beta_1$	4.38	4.15	2.69	0.86
$\beta_2$	3.02	3.89	9.41	0.93

seminal vesicle invasion (**svi**), log(capsular penetration) (**lcp**), Gleason score (**gleason**) and percentage Gleason scores 4 or 5 (**pgg45**). Numerical calculations suggested that we apply a standard normal quantile-transformation on the set of values for each predictor. For the  $j$ th predictor with observations  $\{x_{1,j}, \dots, x_{n,j}\}$ , we computed the empirical distribution function  $F_n(\cdot)$ . For any distinct predictor value,  $x_{i,j}$ , the corresponding observation from the standard normal distribution will be  $\Phi^{-1}(F_n(x_{i,j}))$ . Once  $F_n$  is known, we can easily revert back to the original values from the corresponding standard normal observations. In addition to this, we scaled the response to have mean zero and unit norm. Further, the transformed predictors were also standardized. The whole data set comprising of  $n = 97$  observations were used to fit a linear model to the responses. Initially we applied the modified bootstrap method to select an optimum  $\lambda_0$  value (cf. (5.4)). Figure 5 shows the  $\hat{\phi}_n(\lambda_0, a)$  values for the Prostrate cancer data-set, at different choices of  $a$ . We used  $a \in \{0.016, 0.064, 0.16, 0.255, 0.32, 0.38, 0.50\}$ . The plots for the last three choices were omitted as they were identical with the plot for  $a = 0.255$ . It can be seen from Fig. 5 that the minimum value of  $\hat{\phi}_n(\lambda_0, a)$  occurs at  $\lambda_{0,n}^* = 0.0081$  and at  $a_n^* = 0.255$ . Although it was found that  $\hat{\phi}_n(\lambda_0, a)$  is very insensitive to changes in values of  $a$ , Fig. 5 helps us clearly determine the optimum  $\lambda_0$  for this data-set.

Figure 6 shows the scaled EMSE( $\hat{\phi}_n(\lambda_0, a)$ ) values (cf. (5.7)). As the figure shows, there is little difference in the curves as  $a$  changes, and it can be said that the EMSE is largely insensitive to changes in  $a$ , with exactly same behavior at  $a = 0.255, 0.32, 0.38$  and  $0.50$ . But, all choices of  $a$  show a clear point of minima. Among them, the overall minimum is obtained for  $a = 0.016$  and is shown by the vertical line in the figure. Hence, the best choice of thresholding value is  $\hat{a} = 0.016$ . Smaller values of  $a$  do not make any improvements in the value of the EMSE. Figures 5 and 6 show the applicability of methodology described in Sections 5.2 and 5.3 to real data-sets.

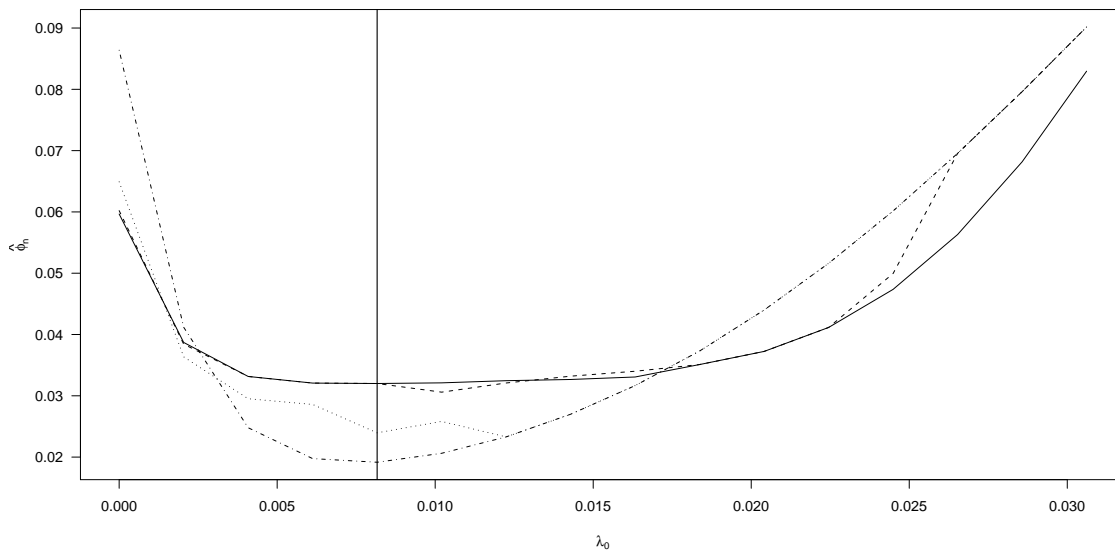


Figure 5: Plot showing  $\hat{\phi}_n(\lambda_0, a)$  values (cf. (5.3), scaled by  $n^{-1}$ ) for the Prostrate cancer data, at four different choices of  $a$ :  $a = 0.016$  (solid line),  $a = 0.064$  (small dashes),  $a = 0.16$  (dots) and  $a = 0.255$  (dash-dot-dash). The vertical solid line indicates the position of  $\lambda_{0,n}^*$ .

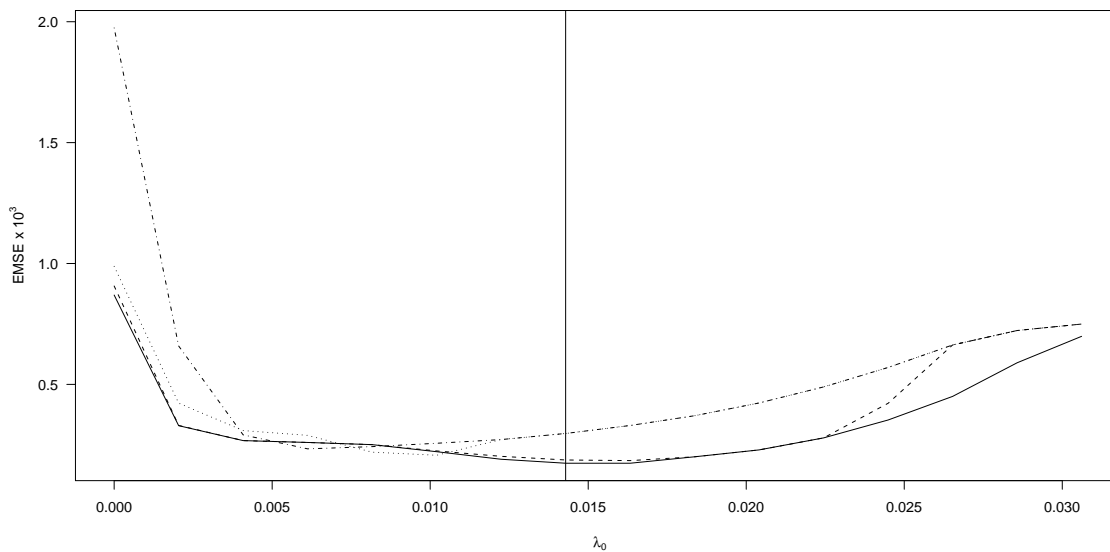


Figure 6: Plot showing  $EMSE(\hat{\phi}_n(\lambda_0, a))$  values (cf. (5.7), scaled by  $10^3$ ) for the Prostrate cancer data, at four different choices of  $a$ :  $a = 0.016$  (solid line),  $a = 0.064$  (small dashes),  $a = 0.16$  (dots) and  $a = 0.255$  (dash-dot-dash). The vertical solid line indicates the position of minima.

At these selected values of  $\lambda_{0,n}^*$  and  $\hat{a}$ , we construct modified bootstrap based confidence intervals of the form

$$I_{n,j,\alpha} \equiv \left\{ u : |u - \hat{\beta}_{n,j}| \leq n^{-1/2} \hat{t}_{n,j}(\alpha) \right\}, \quad j = 1, \dots, p,$$

for each  $\beta_j$ , with  $\hat{t}_{n,j}(\alpha) = \alpha$ -quantile of  $|\mathbf{T}_{n,j}^{**}|$  and  $\hat{\beta}_{n,j}$  being the Lasso estimate for the  $j$ th component. Using the equivalence between confidence sets and tests, we also conduct a bootstrap test for the hypothesis  $H_0 : \beta_j = 0$ , separately for the  $j$ th predictor.

Table 9: Analysis of prostate cancer data from Tibshirani (1996). Table showing componentwise Lasso estimates, modified bootstrap based variance estimates, 90% confidence intervals and tests for  $H_0 : \beta_j = 0$ .  $n = 97$  observations were used with  $\lambda_0 = \lambda_{0,n}^*$  and  $a = \hat{a}$ .

Analysis of Prostrate cancer data *					
predictor	$\hat{\beta}_{n,j}$	Variance estimate ( $\times 10^2$ )	90% Confidence Interval		$H_0 : \beta_j = 0$
			lower	upper	
lcavol	0.493	0.590	0.352	0.635	Reject
lweight	0.182	0.409	0.014	0.351	Reject
age	0	0.061	-0.039	0.039	Accept
lbph	0	0.034	-0.315	0.315	Accept
svi	0.1564	0.560	0	0.313	Reject
lcp	0	0.253	-0.097	0.097	Accept
gleason	0.071	0.201	0	0.142	Reject
pgg45	0	0.098	-0.064	0.064	Accept

\* obtained from <http://www-stat.stanford.edu/tibs/ElemStatLearn/datasets/prostate.data>

From the results in Table 9, it appears that the covariates `lcavol`, `lweight`, `svi` and `gleason` have nontrivial effect on the response variable `lpsa`, the rest of the variables were judged insignificant at level  $1 - \alpha = 0.1$ . The results in Table 9 also show that indeed non-zero variance estimates can be obtained using the modified bootstrap method, even for predictors with  $\hat{\beta}_{n,j} = 0$ .

**Remark 7:** The lower confidence interval values for the covariates `gleason` and `svi` are both zero, and are something of interest. It was found that for the covariate `svi`, the 90% quantile of  $|\mathbf{T}_{n,5}^{**}|$  is approximately same as  $\hat{\beta}_{n,5}$ , and this occurs particularly for  $\alpha = 0.9$ . For other confidence levels, we would have different lower endpoint. For the covariate `svi`, the situation is slightly different. It was found that the distribution of

$|\mathbf{T}_{n,7}^{**}|/\sqrt{n}$  has a large point mass (with probability close to 0.5) at the value 0.071, and that is also precisely same as  $\widehat{\beta}_{n,7}$ , which leads to a lower confidence endpoint value of zero. It should be noted that the bootstrap distributions of  $|\mathbf{T}_{n,j}^{**}|$  are not smooth, and there are values where there are large jumps in the empirical distributions, which can lead to same interval endpoint values for a large range of values of  $\alpha$ .

Overall, we find the the modified bootstrap based method can be applied on real-data sets and the optimum choices of  $\lambda_0$  and thresholding value  $a$  can be obtained in practice and provide satisfactory performance.

## Acknowledgement

The research was supported partially by National Science Foundation grant DMS 0707139. We are also thankful to the Editor and two anonymous referees for their helpful suggestions and comments.

## 8 Proofs

Let  $C, C(\cdot)$  denote generic positive constants that depend on their arguments, but not on  $n$ . Also, recall that we write  $\mathbf{1}(\cdot)$  to denote the indicator function and

$$\text{sgn}(x) = \begin{cases} 1, & \text{if } x > 0, \\ -1, & \text{if } x < 0, \\ 0, & \text{o.w.} \end{cases}$$

Let  $(\Omega, \mathcal{F}, \mathbf{P})$  denote the underlying probability space and let  $\mathcal{E} = \sigma(\epsilon_i : i \geq 1)$  denote the sub- $\sigma$ -field of  $\mathcal{F}$  generated by  $\{\epsilon_i : i \geq 1\}$ . For a random vector  $\mathbf{Z}$  and a  $\sigma$ -field  $\mathcal{C}$ , write  $\mathcal{L}(\mathbf{Z})$  and  $\mathcal{L}(\mathbf{Z}|\mathcal{C})$  to denote the probability distribution of  $\mathbf{Z}$  and the conditional distribution of  $\mathbf{Z}$  given  $\mathcal{C}$ , respectively. For any random vector  $\mathbf{Y}$ , set  $\mathcal{L}(\mathbf{Z}|\sigma(\mathbf{Y})) = \mathcal{L}(\mathbf{Z}|\mathbf{Y})$ , for notational simplicity. Write  $\mathbf{X}_n$  for the  $n \times p$  matrix with rows  $\mathbf{x}'_i$ ,  $i = 1, \dots, n$ , and let  $\mathbf{C}_n = n^{-1}\mathbf{X}'_n\mathbf{X}_n$ . Unless otherwise indicated, limits in the order symbols are taken by letting  $n \rightarrow \infty$ . Recall that  $\mathbf{P}_*$  denotes conditional probability given  $\mathcal{E}$  and  $\mathbf{E}_* = \mathbf{E}(\cdot | \mathcal{E})$ .

**Lemma 8.1.** *Let  $s_n^2 = n^{-1} \sum_{j=1}^n (r_j - \bar{r}_n)^2$  and  $m_{3,n} = n^{-1} \sum_{j=1}^n |r_j - \bar{r}_n|^3$ . Assume that*

$$\frac{1}{n} \sum_{j=1}^n \|\mathbf{x}_j\|^2 = O(1)$$

and Assumption (C.3) holds. Then

$$|s_n^2 - \sigma^2| + n^{-1/2}m_{3,n} \rightarrow 0, \quad \text{with probability 1,}$$

where recall that  $\sigma^2 = \mathbf{Var}(\epsilon_1)$ .

*Proof of Lemma 8.1.* First consider  $|s_n^2 - \sigma^2|$ . Define  $\sigma_n^2 = n^{-1} \sum_{j=1}^n (\epsilon_j - \bar{\epsilon}_n)^2$ , where  $\bar{\epsilon}_n = n^{-1} \sum_{j=1}^n \epsilon_j$ . By Lemma 4.2 of Chatterjee and Lahiri (2010),

$$\|\mathbf{T}_n\| = O(\log n), \quad \text{with probability 1.} \quad (8.1)$$

Hence, using (8.1) and the definition of  $\tilde{\boldsymbol{\beta}}_n$ , we have

$$\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| = O\left(n^{-1/2} \log n\right) \quad \text{with probability 1.} \quad (8.2)$$

By (8.2)

$$\begin{aligned} (s_n - \sigma_n)^2 &\leq \frac{1}{n} \sum_{j=1}^n \left\{ (r_j - \bar{r}_n) - (\epsilon_j - \bar{\epsilon}_n) \right\}^2 \\ &\leq \frac{1}{n} \sum_{j=1}^n (r_j - \epsilon_j)^2 \\ &\leq \left( n^{-1} \sum_{j=1}^n \|\mathbf{x}_j\|^2 \right) \|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|^2 \\ &= o(1), \quad \text{with probability 1.} \end{aligned}$$

Since  $\sigma_n^2 \rightarrow \sigma^2$  with probability 1, it follows that  $|s_n^2 - \sigma^2| = o(1)$  with probability 1. Next consider  $m_{3,n}$ .

Using the assumptions on the  $\mathbf{x}_i$ 's, we get

$$\max_{1 \leq i \leq n} \|\mathbf{x}_i\| \leq \left( \sum_{i=1}^n \|\mathbf{x}_i\|^2 \right)^{1/2} = O\left(n^{1/2}\right). \quad (8.3)$$

Hence, by the Marcinkiewicz-Zygmund Strong Law of Large Numbers, (8.2), and (8.3), we have

$$\begin{aligned}
|n^{-1/2}m_{3,n}| &\leq 8n^{-3/2} \sum_{i=1}^n \left| \epsilon_i - \mathbf{x}_i(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \right|^3 \\
&\leq 32 \left\{ n^{-3/2} \sum_{i=1}^n |\epsilon_i|^3 + \left( n^{-3/2} \sum_{i=1}^n \|\mathbf{x}_i\|^3 \right) \|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|^3 \right\} \\
&= o(1), \quad \text{with probability 1.}
\end{aligned}$$

This completes the proof of the lemma.  $\square$

**Lemma 8.2.** *Suppose that Assumptions (C.1) and (C.3) hold. Then*

(i)

$$\mathcal{L} \left( n^{-1/2} \sum_{i=1}^n \mathbf{x}_i e_i^{**} \mid \mathcal{E} \right) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{C}) \quad \text{with probability 1.}$$

(ii)

$$\mathcal{L} \left( n^{-1/2} \sum_{i=1}^n \mathbf{x}_i \epsilon_i^+ \mid \mathcal{E} \right) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{C}), \quad \text{as } n \rightarrow \infty, \text{ in probability.} \quad (8.4)$$

*Proof of Lemma 8.2.* By Lemma 8.1, there exists a set  $A \in \mathcal{F}$  be such that  $\mathbf{P}(A) = 1$  and for every  $\omega \in A$ ,

$$|\mathbf{E}_*(e_i^{**})^2 - \sigma^2| + n^{-1/2} \mathbf{E}_* |e_i^{**}|^3 \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Fix  $\omega \in A$ . For this  $\omega$ , we will use the Cramer-Wold device to prove the result. Accordingly, consider a  $\mathbf{a} = (a_1, \dots, a_p)' \in \mathbb{R}^p$ ,  $\mathbf{a} \neq \mathbf{0}$ . Let  $s_n^2(\mathbf{a}) = \mathbf{Var}_* \left( n^{-1/2} \sum_{i=1}^n \mathbf{a}' \mathbf{x}_i e_i^{**} \right)$ . Note that by the definition of  $A$ ,  $s_n^2(\mathbf{a}) \rightarrow \mathbf{a}' \mathbf{C} \mathbf{a} \sigma^2 \in (0, \infty)$ . By the Berry-Esseen Theorem for independent (but possibly non-identically distributed random variables) (cf. Chapter 12, [Bhattacharya and Ranga Rao \(1986\)](#)),

$$\begin{aligned}
\sup_{x \in \mathbb{R}} \left| \mathbf{P}_* \left( n^{-1/2} \sum_{i=1}^n \mathbf{a}' \mathbf{x}_i e_i^{**} \leq x \right) - \Phi \left( x s_n^{-1}(\mathbf{a}) \right) \right| &\leq (2.75) \frac{\sum_{i=1}^n \mathbf{E}_* \left| n^{-1/2} \mathbf{a}' \mathbf{x}_i e_i^{**} \right|^3}{\left( \sum_{i=1}^n \mathbf{E}_* \left| n^{-1/2} \mathbf{a}' \mathbf{x}_i e_i^{**} \right|^2 \right)^{3/2}} \\
&\leq C \frac{\|\mathbf{a}\|^3 n^{-3/2} \sum_{i=1}^n \|\mathbf{x}_i\|^3 \mathbf{E}_* |e_i^{**}|^3}{s_n^3(\mathbf{a})} \\
&= o(1).
\end{aligned}$$

This completes the proof of part (i).

For part (ii): by a subsequence argument and by retracing the arguments in the proof of part (i), it is

enough to show that

$$n^{-1} \sum_{i=1}^n \check{\epsilon}_{1,i} \xrightarrow{P} 0; \quad n^{-1} \sum_{i=1}^n \check{\epsilon}_{1,i}^2 \xrightarrow{P} \sigma^2, \quad \text{and} \quad n^{-3/2} \sum_{i=1}^n |\check{\epsilon}_{1,i}|^3 \xrightarrow{P} 0.$$

Since,

$$\max\{|\check{\epsilon}_{1,i} - \epsilon_i| : 1 \leq i \leq n\} \leq \max\{\|\mathbf{x}_i\| : 1 \leq i \leq n\} \cdot \|\check{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| = O_p(n^{-1/2}),$$

the above follows the Marcinkiewz-Zygmund SLLN.  $\square$

*Proof of Theorem 3.1.* Let  $\mathbf{u} = (u_1, \dots, u_p)' \in \mathbb{R}^p$ . Define

$$U_n^{**}(\mathbf{u}) = \sum_{i=1}^n \left\{ y_i^{**} - \mathbf{x}_i' \left( \tilde{\boldsymbol{\beta}}_n + \mathbf{u} n^{-1/2} \right) \right\}^2 + \lambda_n \sum_{j=1}^p \left| \tilde{\beta}_{n,j} + u_j n^{-1/2} \right|.$$

Also define  $V_n^{**}(\mathbf{u}) = U_n^{**}(\mathbf{u}) - U_n^{**}(\mathbf{0})$ . Note that  $\mathbf{T}_n^{**} = \text{argmin}(V_n^{**})$ . Define

$$\mathbf{W}_n^{**} = n^{-1/2} \sum_{i=1}^n \mathbf{x}_i e_i^{**}.$$

Then we can write  $V_n^{**}$  as

$$V_n^{**}(\mathbf{u}) = \mathbf{u}' \mathbf{C}_n \mathbf{u} - 2\mathbf{u}' \mathbf{W}_n^{**} + \lambda_n \sum_{j=1}^p \left( \left| \tilde{\beta}_{n,j} + u_j n^{-1/2} \right| - \left| \tilde{\beta}_{n,j} \right| \right). \quad (8.5)$$

Let  $A \in \mathcal{F}$  be such that  $\mathbf{P}(A) = 1$  and for every  $\omega \in A$ , (8.1), (8.2) hold and

$$\mathcal{L}\left(\mathbf{W}_n^{**} \mid \mathcal{E}\right)(\omega) \longrightarrow N(\mathbf{0}, \sigma^2 \mathbf{C}), \quad \text{in distribution.} \quad (8.6)$$

Now fix  $\omega \in A$ . Then, there exists  $N = N(\omega) \in [1, \infty)$  such that for all  $n \geq N$ ,

$$\begin{aligned} \text{sgn}(\hat{\beta}_{n,j}) &= \text{sgn}(\beta_j) \quad \text{and} \quad \tilde{\beta}_{n,j} = \hat{\beta}_{n,j} \quad \text{for all } 1 \leq j \leq p_0, \\ \text{and } \tilde{\beta}_{n,j} &= 0 \quad \text{for all } p_0 + 1 \leq j \leq p. \end{aligned}$$

Hence, for all  $n \geq N$ ,

$$V_n^{**}(\mathbf{u}) = \mathbf{u}' \mathbf{C}_n \mathbf{u} - 2\mathbf{u}' \mathbf{W}_n^{**} + \lambda_n n^{-1/2} \left( \sum_{j=1}^{p_0} \text{sgn}(\tilde{\beta}_{n,j}) u_j + \sum_{j=p_0+1}^p |u_j| \right),$$

for all  $\mathbf{u} \in \mathbb{R}^p$  with  $|u_j| \leq n^{1/2} \hat{\beta}_{n,j}$  for  $j = 1, \dots, p_0$ . Now using (8.6) and the arguments as in the

unbootstrapped case (cf. Knight and Fu (2000)), one can establish the weak convergence

$$\mathcal{L}\left(\mathbf{V}_n^{**}(\cdot) \mid \mathcal{E}\right)(\omega) \longrightarrow \mathcal{L}(\mathbf{V}(\cdot)), \quad \text{in distribution,}$$

on the space of all functions on  $\mathbb{R}^p$  that are uniformly bounded on compact subsets of  $\mathbb{R}^p$ . This, in turn, implies that  $\mathcal{L}(\mathbf{T}_n^{**} \mid \mathcal{E})(\omega) \longrightarrow \mathcal{L}(\mathbf{T}_\infty)$  in distribution, as random vectors. Since this is true for all  $\omega \in A$  and  $\mathbf{P}(A) = 1$ , the theorem is proved.  $\square$

*Proof of Corollary 3.2.* Part (a) follows from Theorem 3.1, by using the fact that if  $F_n \longrightarrow F$  weakly and  $F(\cdot)$  is strictly increasing to the right of its  $\alpha$ -quantile  $F^{-1}(\alpha)$ , then,  $F_n^{-1}(\alpha) \rightarrow F^{-1}(\alpha)$  as  $n \rightarrow \infty$ . For the second, use the fact that  $\|\mathbf{T}_\infty\|$  has a continuous distribution on  $\mathbb{R}$  when  $\beta_j \neq 0$  for at least one  $j$ .  $\square$

*Proof of Theorem 3.3.* Firstly we show that  $\left\{\|\mathbf{T}_n^{**}\|^2 : n \geq 1\right\}$  is uniformly integrable with probability 1, i.e.,

$$\limsup_{\alpha \rightarrow \infty} \sup_{n \geq 1} \mathbf{E}_* \|\mathbf{T}_n^{**}\|^2 \mathbf{1}(\|\mathbf{T}_n^{**}\| \geq \alpha) = 0, \quad \text{with probability 1.}$$

Consider any fixed  $w_0 \in (0, \infty)$  such that  $\|\mathbf{W}_n^{**}\| < w_0$ . Now, we can write

$$\begin{aligned} V_n^{**}(\mathbf{u}) &= \mathbf{u}' \mathbf{C}_n \mathbf{u} - 2\mathbf{u}' \mathbf{W}_n^{**} + \lambda_n \sum_{j=1}^p \left( \left| \tilde{\beta}_{n,j} + u_j n^{-1/2} \right| - |\tilde{\beta}_{n,j}| \right) \\ &\geq \|\mathbf{u}\| \left\{ \eta_{1,n} \|\mathbf{u}\| - 2\|\mathbf{W}_n^{**}\| - \lambda_n (n^{-1}p)^{1/2} \right\} \\ &\geq \|\mathbf{u}\| \left[ \eta_{1,n} \|\mathbf{u}\| - \left\{ 2w_0 + \lambda_n (n^{-1}p)^{1/2} \right\} \right] \\ &\geq 2\eta_{1,n}^{-1} \left\{ 2w_0 + \lambda_n (n^{-1}p)^{1/2} \right\}^2 \\ &> 8w_0^2 \eta_{1,n}^{-1} (> 0), \end{aligned}$$

for all  $\|\mathbf{u}\| > 2\eta_{1,n}^{-1} \left\{ 2w_0 + \lambda_n (n^{-1}p)^{1/2} \right\}$ . But note that  $\lim_{\|\mathbf{u}\| \rightarrow 0} V_n^{**}(\mathbf{u}) = 0$ , and that implies

$$\mathbf{T}_n^{**} \in \left[ \mathbf{u} : \|\mathbf{u}\| \leq 2\eta_{1,n}^{-1} \left\{ 2w_0 + \lambda_n (n^{-1}p)^{1/2} \right\} \right].$$

This leads to the observation that

$$\left\{ \|\mathbf{W}_n^{**}\| < w_0 \right\} \Rightarrow \left[ \|\mathbf{T}_n^{**}\| \leq 2\eta_{1,n}^{-1} \left\{ 2w_0 + \lambda_n (n^{-1}p)^{1/2} \right\} \right].$$

For any fixed  $t_0 = 2\eta_{1,n}^{-1} \left\{ 2w_0 + \lambda_n(n^{-1}p)^{1/2} \right\}$ , we obtain

$$w_0 = \frac{1}{2} \left\{ \frac{t_0 \eta_{1,n}}{2} - \lambda_n(n^{-1}p)^{1/2} \right\}.$$

And using the above relation we can write

$$\left\{ \|\mathbf{T}_n^{**}\| > t_0 \right\} \Rightarrow \left[ \|\mathbf{W}_n^{**}\| \geq \frac{1}{2} \left\{ \frac{t_0 \eta_{1,n}}{2} - \lambda_n(n^{-1}p)^{1/2} \right\} \right]. \quad (8.7)$$

Using the Assumptions (C.1) and (C.2) we can say,  $\eta_{1,n} \rightarrow \eta_1 (> 0)$  and  $\lambda_n(n^{-1}p)^{1/2} \rightarrow \lambda_0 p^{1/2}$ , which is finite. Using (8.7), we can write

$$\begin{aligned} \mathbf{E}_* \|\mathbf{T}_n^{**}\|^2 \mathbf{1}(\|\mathbf{T}_n^{**}\| \geq \alpha) &\leq \left\{ \alpha^2 \mathbf{P}_* (\|\mathbf{T}_n^{**}\| > \alpha) + \int_{\alpha}^{\infty} t \mathbf{P}_* (\|\mathbf{T}_n^{**}\| > t) dt \right\} \\ &\leq 4 \int_{\alpha/2}^{\infty} t \mathbf{P}_* (\|\mathbf{T}_n^{**}\| > t) dt \\ &\leq 4 \int_{\alpha/2}^{\infty} t \mathbf{P}_* \left\{ \|\mathbf{W}_n^{**}\| \geq \frac{1}{2} \left( \frac{t \eta_{1,n}}{2} - \lambda_n n^{-1/2} p^{1/2} \right) \right\} dt \\ &\leq 4 \int_{\alpha/2}^{\infty} t \mathbf{P}_* \left( \|\mathbf{W}_n^{**}\| \geq \frac{t \eta_1}{4} \right) dt, \end{aligned} \quad (8.8)$$

for  $n$  and  $\alpha$  large enough. Let  $\mathbf{W}_{\infty}$  be a random vector following the  $N(\mathbf{0}, \sigma^2 \mathbf{C})$  distribution. Note that using (8.6) and the continuous mapping theorem, we can say that for any fixed  $c \in (0, \infty) \setminus D$ , where  $D$  is a countable set,

$$\|\mathbf{W}_n^{**}\|^2 \mathbf{1}(\|\mathbf{W}_n^{**}\| \geq c) \rightarrow \|\mathbf{W}_{\infty}\|^2 \mathbf{1}(\|\mathbf{W}_{\infty}\| \geq c), \quad \text{in distribution,}$$

with probability 1 and using Lemma 8.1 we have

$$\mathbf{E}_* \|\mathbf{W}_n^{**}\|^2 = \tilde{s}_n^2 \text{tr}(\mathbf{C}_n) \rightarrow \mathbf{E} \|\mathbf{W}_{\infty}\|^2 = \text{tr}(\sigma^2 \mathbf{C}), \quad \text{with probability 1.}$$

Combining the above two results along with the Dominated Convergence Theorem, we obtain

$$\mathbf{E}_* \|\mathbf{W}_n^{**}\|^2 \mathbf{1}(\|\mathbf{W}_n^{**}\| \geq c) \rightarrow \mathbf{E} \|\mathbf{W}_{\infty}\|^2 \mathbf{1}(\|\mathbf{W}_{\infty}\| \geq c) \quad \text{with probability 1,} \quad (8.9)$$

for every  $c \in (0, \infty) \setminus D$ . Further, the right side of (8.9) is finite for any  $c > 0$  and goes to zero as  $c \rightarrow \infty$ . Hence  $\{\|\mathbf{W}_n^{**}\|^2 : n \geq 1\}$  is uniformly integrable with probability 1. This implies that the integral on the right side of (8.8) is finite as  $\alpha > 0$  and it tends to zero as  $\alpha \uparrow \infty$ . Hence,  $\{\|\mathbf{T}_n^{**}\|^2 : n \geq 1\}$  is uniformly integrable with probability 1. Since  $\|\mathbf{x}\|^2$  is continuous function on  $\mathbb{R}^p$ , by Theorem 3.1 and the uniform

integrability of  $\{\|\mathbf{T}_n^{**}\|^2 : n \geq 1\}$ ,

$$\mathbf{E}_* \|\mathbf{T}_n^{**}\|^2 \rightarrow \mathbf{E} \|\mathbf{T}_\infty\|^2, \quad \text{with probability 1.}$$

Now note that for any fixed  $j \in \{1, \dots, p\}$  and any  $\alpha > 0$ , we have

$$\mathbf{E}_* |\mathbf{T}_{n,j}^{**}|^2 \mathbf{1}(|\mathbf{T}_{n,j}^{**}| \geq \alpha) \leq \mathbf{E}_* \|\mathbf{T}_n^{**}\|^2 \mathbf{1}(\|\mathbf{T}_n^{**}\| \geq \alpha).$$

Since  $\{\|\mathbf{T}_n^{**}\|^2 : n \geq 1\}$  is uniformly integrable, this implies  $\{|\mathbf{T}_{n,j}^{**}|^2 : n \geq 1\}$  is uniformly integrable, with probability 1 for all  $j \in \{1, \dots, p\}$ . Also note that the projection mapping  $g_j : \mathbf{x} \mapsto x_j$  is continuous on  $\mathbb{R}^p$  and  $\mathbf{P}(\mathbf{T}_\infty \in \mathbb{R}^p) = 1$ . This implies  $\mathbf{T}_{n,j}^{**} \rightarrow \mathbf{T}_{\infty,j}$  in distribution, with probability 1. Thus

$$\mathbf{E}_* |\mathbf{T}_{n,j}^{**}|^2 \rightarrow \mathbf{E} |\mathbf{T}_{\infty,j}|^2, \quad \text{with probability 1.} \quad (8.10)$$

Since  $L_2$  convergence implies  $L_1$  convergence,  $\mathbf{E}_*(\mathbf{T}_{n,j}^{**}) \rightarrow \mathbf{E}(\mathbf{T}_{\infty,j})$ , with probability 1. Hence, for all  $j \in \{1, \dots, p\}$ ,

$$\mathbf{Var}_*(\mathbf{T}_{n,j}^{**}) \rightarrow \mathbf{Var}(\mathbf{T}_{\infty,j}) \quad \text{with probability 1.} \quad (8.11)$$

For the off-diagonal elements, using similar arguments, we can write

$$\mathbf{T}_{n,j}^{**} \mathbf{T}_{n,k}^{**} \rightarrow \mathbf{T}_{\infty,j} \mathbf{T}_{\infty,k}, \quad \text{in distribution,}$$

with probability 1 for any  $j \neq k$ . Also note that

$$|\mathbf{T}_{n,j}^{**} \mathbf{T}_{n,k}^{**}| \leq \frac{(|\mathbf{T}_{n,j}^{**}|^2 + |\mathbf{T}_{n,k}^{**}|^2)}{2} \leq \|\mathbf{T}_n^{**}\|^2,$$

and  $\{\|\mathbf{T}_n^{**}\|^2 : n \geq 1\}$  is uniformly integrable. Hence, for all  $j, k \in \{1, \dots, p\}$ , ( $j \neq k$ ),  $\{\mathbf{T}_{n,j}^{**} \mathbf{T}_{n,k}^{**} : n \geq 1\}$  is uniformly integrable and

$$\mathbf{E}_*(\mathbf{T}_{n,j}^{**} \mathbf{T}_{n,k}^{**}) \rightarrow \mathbf{E}(\mathbf{T}_{\infty,j} \mathbf{T}_{\infty,k}), \quad \text{with probability 1.} \quad (8.12)$$

Combining (8.11) and (8.12), we have the proof for the strong consistency of the modified bootstrap variance matrix estimator. The proof for the bias part is similar, in view of (8.10).  $\square$

*Proof of Theorem 4.1.* Let  $A_n$  denote the event that the ALASSO correctly identifies the set of all zero

components of  $\beta$ , i.e.,  $A_n$  is the set of all  $\omega \in \Omega$  such that

$$\{j : 1 \leq j \leq p, \check{\beta}_{n,j}(\omega) = 0\} = \{p_0 + 1, \dots, p\}.$$

Also, let

$$V_n^+(\mathbf{u}) = \mathbf{u}'\mathbf{C}_n\mathbf{u} - 2\mathbf{u}'\mathbf{W}_n^+ + \lambda_n \sum_{j=1}^p |\check{\beta}_{n,j}^+|^{-\gamma} \left( |\check{\beta}_{n,j} + \frac{u_j}{\sqrt{n}}| - |\check{\beta}_{n,j}| \right),$$

where  $\mathbf{W}_n^+ = n^{-1/2} \sum_{i=1}^n \mathbf{x}_i \epsilon_i^+$ ,  $n \geq 1$ . Then it is easy to check that

$$\check{\beta}_n^+ = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} V_n^+(\mathbf{u}).$$

Since  $\mathbf{P}(A_n) \rightarrow 1$ , there exists a subsequence  $\{n_k\}$  such that

$$\mathbf{P}(A_{n_k}^c \text{ i.o.}) = 0. \quad (8.13)$$

Let  $\Omega_0^c \in \mathcal{F}$  be the union of the set  $\limsup_{k \rightarrow \infty} A_{n_k}^c$  and the set where (8.4) fails to hold. Then  $\mathbf{P}(\Omega_0) = 1$ . Fix,  $\omega \in \Omega_0$ . Then, there exists  $n_\omega \geq 1$  such that for all  $n \geq n_\omega$ ,  $B_n = \{p_0 + 1, \dots, p\}$ . Using arguments similar to those in the proof of Theorem 3.1 and Zou (2006) it is easy to show that, on  $\Omega_0$ ,

$$\mathcal{L}(V_{n_k}^+(\cdot) | \mathcal{E}) \rightarrow \mathcal{L}(\check{V}_\infty(\cdot)), \quad \text{in distribution}$$

on the space of all functions on  $\mathbb{R}^p$  that are uniformly bounded on compact subsets of  $\mathbb{R}^p$ , where

$$\check{V}_\infty(\mathbf{u}) = \begin{cases} \mathbf{u}'\mathbf{C}\mathbf{u} - 2\mathbf{u}'\mathbf{W} & \text{if } u_{p_0+1} = \dots = u_p = 0. \\ 0 & \text{o.w.} \end{cases}$$

On  $\Omega_0$ , weak convergence of  $\mathcal{L}(\sqrt{n}(\check{\beta}_{n_k}^+ - \check{\beta}_n) | \mathcal{E})$  to  $\mathcal{L}(\check{\mathbf{T}}_\infty)$  now follows from the argmax theorem (cf. Knight and Fu (2000)). This, together with Theorem 2 of Zou (2006) implies the assertion of Theorem 4.1.  $\square$

*Proof of Corollary 4.2.* The proof is similar to the proof of Corollary 3.2. Routine details are omitted.  $\square$

*Proof of Corollary 4.3.* As in the proof of Theorem 4.1, it is enough to show that along a subsequence  $\{n_k\}$ ,

$$\lim_{\alpha \rightarrow \infty} \sup_{k \geq 1} \mathbf{E}_* \|\mathbf{T}_{n_k}^+\|^2 \mathbf{1}(\|\mathbf{T}_{n_k}^+\| \geq \alpha) = 0, \quad \text{with probability 1.} \quad (8.14)$$

Using arguments similar to those leading to (8.7), one can show that for any  $n \geq 1$ , on the set

$$\left\{ \check{\beta}_{n,j} = 0, \text{ for } j = p_0 + 1, \dots, p \right\},$$

for any  $t_0 > 0$ ,

$$\left\{ \|\check{\mathbf{T}}_n^+\| > t_0 \right\} \subset \left\{ \|\mathbf{W}_n^+\| > \frac{1}{2} \left[ \frac{t_0 \eta_{1,n}}{2} - \frac{\sqrt{p_0} \lambda_n}{\sqrt{n}} \cdot \frac{1}{\min \left\{ |\check{\beta}_{j,n}|^{\gamma/2} : 1 \leq j \leq p_0 \right\}} \right] \right\}, \quad n \geq 1.$$

Next, (i) restricting attention to the set  $\Omega_0$  (from the proof of Theorem 4.1), (ii) using the facts that

$$\min \left\{ |\beta_{j,n_k}|^{\gamma/2} : 1 \leq j \leq p_0 \right\} = O(1), \quad \text{on } \Omega_0,$$

and  $n^{-1/2} \lambda_n \rightarrow 0$  as  $n \rightarrow \infty$ , and (iii) using arguments similar to (8.8) and (8.9), one can prove (8.14). This proves Corollary 4.3.  $\square$

## References

- Andrews, D. W. K. and Guggenberger, P. (2009), “Incorrect asymptotic size of subsampling procedures based on post-consistent model selection estimators,” *Journal of Econometrics*, 152, 19–27.
- Bach, F. (2009), “Model-Consistent Sparse Estimation through the Bootstrap.” Preprint available at <http://arxiv.org/abs/0901.3202>.
- Bhattacharya, R. N. and Ranga Rao, R. (1986), *Normal Approximation and Asymptotic Expansions*, Melbourne, FL: Robert E. Krieger Publishing Co. Inc., reprint of the 1976 original.
- Bickel, P. J., Götze, F., and van Zwet, W. R. (1997), “Resampling fewer than  $n$  observations: gains, losses, and remedies for losses,” *Statistica Sinica*, 7, 1–31.
- Chatterjee, A. and Lahiri, S. N. (2010), “Asymptotic properties of the residual bootstrap for Lasso estimators,” *Proceedings of the American Mathematical Society*, 138, 4497–4509.
- Draper, N. R. and Smith, H. (1998), *Applied regression analysis*, Wiley Series in Probability and Statistics: Texts and References Section, New York: John Wiley & Sons Inc., 3rd ed.
- Efron, B. (1979), “Bootstrap methods: another look at the jackknife,” *The Annals of Statistics*, 7, 1–26.

- (1982), *The jackknife, the bootstrap and other resampling plans*, vol. 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Philadelphia, Pa.: Society for Industrial and Applied Mathematics (SIAM).
- (1992), “Jackknife-after-bootstrap standard errors and influence functions,” *Journal of the Royal Statistical Society. Series B. Methodological*, 54, 83–127.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), “Least angle regression,” *The Annals of Statistics*, 32, 407–499, with discussion, and a rejoinder by the authors.
- Fan, J. and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96, 1348–1360.
- Fan, J. and Peng, H. (2004), “Nonconcave penalized likelihood with a diverging number of parameters,” *The Annals of Statistics*, 32, 928–961.
- Freedman, D. A. (1981), “Bootstrapping regression models,” *The Annals of Statistics*, 9, 1218–1228.
- Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), “Pathwise coordinate optimization,” *The Annals of Applied Statistics*, 1, 302–332.
- Hall, P., Lee, E. R., and Park, B. U. (2009), “Bootstrap-based penalty choice for the lasso, achieving oracle performance,” *Statistica Sinica*, 19, 449–471.
- Huang, J., Horowitz, J. L., and Ma, S. (2008), “Asymptotic properties of bridge estimators in sparse high-dimensional regression models,” *The Annals of Statistics*, 36, 587–613.
- Knight, K. and Fu, W. (2000), “Asymptotics for lasso-type estimators,” *The Annals of Statistics*, 28, 1356–1378.
- Leeb, H. and Pötscher, B. M. (2006), “Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results,” *Econometric Theory*, 22, 69–97.
- (2008), “Sparse estimators and the oracle property, or the return of Hodges’ estimator,” *Journal of Econometrics*, 142, 201–211.
- Osborne, M. R., Presnell, B., and Turlach, B. A. (2000), “On the LASSO and its dual,” *Journal of Computational and Graphical Statistics*, 9, 319–337.
- Pötscher, B. M. and Schneider, U. (2009), “On the distribution of the adaptive LASSO estimator,” *Journal of Statistical Planning and Inference*, 139, 2775–2790.

- 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.” *The Journal of Urology*, 141, 1076–1083.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B. Methodological*, 58, 267–288.
- Wainwright, M. J. (2006), “Sharp thresholds for high-dimensional and noisy recovery of sparsity,” Tech. rep., Dept. of Statistics, UC Berkeley., preprint 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.
- Zhang, J., Jeng, X. J., and Liu, H. (2008), “Some Two-Step Procedures for Variable Selection in High-Dimensional Linear Regression.” Preprint available at <http://arxiv.org/abs/0810.1644>.
- Zhao, P. and Yu, B. (2006), “On model selection consistency of Lasso,” *Journal of Machine Learning Research (JMLR)*, 7, 2541–2563.
- Zou, H. (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101, 1418–1429.