

# Bootstrapping Lasso estimators\*

A. Chatterjee<sup>†</sup> and S. N. Lahiri<sup>‡</sup>

Department of Statistics

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 (cf. Chatterjee and Lahiri (2009)). 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. Using the latter result, we formulate a novel data based method for choosing the optimal penalizing parameter 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 two real data examples.

*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

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

---

\*Research partially supported by NSF grant DMS-0707139.

<sup>†</sup>email: cha@stat.tamu.edu

<sup>‡</sup>email: snlahiri@stat.tamu.edu

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

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

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), Fu (1998)). The sparse solutions obtained by using the Lasso automatically leads to model selection. 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)). The Lasso is found to be model consistent when the design matrix satisfies the so-called *irrepresentability* or *incoherence* conditions, which impose suitable restrictions on the design matrix to make the Lasso estimate model consistent. 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 (2008) 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, and this justified the use of Lasso as a variable selection method. Similar results were obtained by Zou (2006) in case of the Adaptive Lasso, which uses a data-dependent regularization. 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 (2009), 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. Tibshirani (1996) suggested an approximation to the covariance matrix of the Lasso estimator, but that lead to the problem of estimated standard error being zero in case the estimated coefficient was zero. Osborne et al. (2000) suggested an alternative that apparently gave better standard error estimates. 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. A related method for variance estimation in penalized regression setup using the SCAD penalty function was suggested by Fan and Li (2001), where they used a local quadratic approximation to derive a *sandwich* formula for estimating the covariance matrix. It was shown by Fan and Peng (2004), that this variance estimator is consistent for non-zero estimated coefficients, but it fails to provide variance estimates for the estimated zero coefficients. 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 balls 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  $\lambda_n/\sqrt{n} \rightarrow \lambda_0$ . In particular, the accuracy of the Lasso estimator  $\hat{\beta}_n$  critically depends on the choice of the regularization parameter  $\lambda_n$ . In Section 4, 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  $\lambda_n$ . This seems to be an important addition to the literature, where no satisfactory alternative method (including other versions of the bootstrap methods) for estimation of the regularization parameter is available. See Section 4 for more details.

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 (2009), in the same set up. In the high-dimensional case, where  $p$  is allowed to grow with  $n$ , work on estimation consistency of the Lasso is limited; most work concentrate on the properties of the Lasso as a variable selection method. Meinshausen and Yu

(2008) investigate the model selection consistency of the Lasso when the stringent *irrepresentable* conditions on the design matrix fail to hold, and they provide slightly weaker conditions under which the Lasso is  $\ell_2$ -convergent and also provide a threshold version of the Lasso estimator, which is shown to be sign-consistent. Zhang and Huang (2008) consider the sparsity and bias of models selected by the Lasso under more general conditions. For high dimensional graphical models, Meinshausen and Bühlmann (2006) study the consistency of neighborhood selection under an  $\ell_1$ -penalty. Huang et al. (2008) considered the asymptotic properties of  $\ell_q$  norm penalized regression estimators ( $0 < q < 1$ ) for high dimensional regression models. Bickel et al. (2008) compare the convergence rates of the Lasso with other penalized model selection methods like the Dantzig selector (Candes and Tao (2007)).

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. The data based method for the selection of the (MSE-optimal) regularization parameter is presented in Section 4. In Section 5, we conduct a simulation study to compare the finite sample performance of the bootstrapped Lasso estimator and its modified version. Two real data examples are presented in Section 6. The proofs of our results are given in Section 7.

## 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 this case, 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  $\hat{\boldsymbol{\beta}}_n$  denote the Lasso estimator of  $\boldsymbol{\beta}$  given by (1.2). Define the residuals

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

Consider the set of centered residuals  $\{e_1 - \bar{e}_n, \dots, e_n - \bar{e}_n\}$ , where  $\bar{e}_n = n^{-1} \sum_i e_i$ . For the residual bootstrap, one selects a random sample  $\{e_i^*\}_{i=1}^n$ , of size  $n$  from  $\{e_i - \bar{e}_n : i = 1, \dots, n\}$  with replacement and formulates the bootstrap version of (1.1) as

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

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

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

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

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

where  $\mathbf{P}_*$  denotes conditional probability given the error variables  $\{\epsilon_i : i \geq 1\}$  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  $\hat{G}_n(\cdot)$  to be close to  $G_n(\cdot)$ . However, this is *not* the case; Chatterjee and Lahiri (2009), show that the residual bootstrap estimator  $\hat{G}_n(\cdot)$ , 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 (2009)), 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  $\boldsymbol{\beta}$ . As seen in Chatterjee and Lahiri (2009), the inconsistency of the standard residual bootstrap arises when some components of  $\boldsymbol{\beta}$  are zero. The key idea behind the modified bootstrap is to force components of the Lasso estimator  $\hat{\boldsymbol{\beta}}_n$  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

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

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

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

where  $\widehat{\boldsymbol{\beta}}_n$  is the usual Lasso estimate defined in (1.2). Note that for a nonzero component  $\beta_j$ ,

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

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,  $\beta_j = 0$ ,

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

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

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

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 one, as the sample size  $n$  goes to infinity.

Next, we define the *modified* residuals  $\{\widetilde{e}_i\}_{i=1}^n$  based on this estimator  $\widetilde{\boldsymbol{\beta}}_n$  by

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

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

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

Then, the modified bootstrap Lasso estimator is given by

$$\boldsymbol{\beta}_n^{**} := \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} \sum_{i=1}^n (y_i^{**} - \mathbf{x}_i' \mathbf{u})^2 + \lambda_n \sum_{j=1}^p |u_j|.\tag{2.5}$$

Let

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

and let  $\widetilde{G}_n(\cdot)$  denote the conditional distribution of  $\mathbf{T}_n^{**}$  given the observations (or the error variables  $\{\epsilon_i : i \geq 1\}$ ), i.e.,  $\widetilde{G}_n(B) = \mathbf{P}_*(\mathbf{T}_n^{**} \in B)$ ,  $B \in \mathcal{B}(\mathbb{R}^p)$ . Thus,  $\widetilde{G}_n(\cdot)$  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 given by  $\varphi(\tilde{G}_n)$ . For example, the modified bootstrap estimator of  $\mathbf{E}(\mathbf{T}_n) =$  the scaled bias of  $\hat{\beta}_n$  is given by  $\mathbf{E}_*(\mathbf{T}_n^{**})$  and similarly, that of  $\mathbf{Var}(\mathbf{T}_n)$  is given by  $\mathbf{Var}_*(\mathbf{T}_n^{**})$ , where  $\mathbf{E}_*$  and  $\mathbf{Var}_*$  denote the expectation and variance under  $\mathbf{P}_*$ .

In the next section we show that under some mild conditions, the modified bootstrap estimators of the distribution function of  $\mathbf{T}_n$  and of its bias and variance functionals are consistent.

### 3 Main results

#### 3.1 Consistency of the distributional approximation

The first result shows that the modified bootstrap gives a valid approximation to the distribution of  $\mathbf{T}_n$ :

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

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

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

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

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 conditions (C.1)-(C.3). In contrast, for the standard version of the residual bootstrap, Chatterjee and Lahiri (2009), shows that under the same set of regularity conditions, if  $\lambda_0 > 0$  in (C.2) and if  $\beta$  has at least one zero-component, then

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

where  $\hat{G}_n$  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 one, 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  $\boldsymbol{\beta} \in \mathbb{R}^p$ , even when some of the components of  $\boldsymbol{\beta}$  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  $\boldsymbol{\beta}$  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  $\boldsymbol{\beta}$  for all values of the unknown regression parameters  $\boldsymbol{\beta} \in \mathbb{R}^p$ , including the cases where one or more components of  $\boldsymbol{\beta}$  are zero. More specifically, let  $\hat{t}_{n,\alpha}$  denote the  $\alpha$  quantile of the bootstrap distribution of  $\|\mathbf{T}_n^{**}\|$ ,  $\alpha \in (0, 1)$ . Then, the set

$$I_{n,\alpha} \equiv \left\{ \mathbf{t} \in \mathbb{R}^p : \|\mathbf{t} - \hat{\boldsymbol{\beta}}_n\| \leq n^{-1/2} \hat{t}_{n,\alpha} \right\}$$

is an approximate confidence set for  $\boldsymbol{\beta}$  of level  $\alpha$ , as shown by the following result. To state it, let  $\mathbf{T}_\infty$  denote the limiting random vector such that  $\mathbf{T}_n \xrightarrow{d} \mathbf{T}_\infty$  (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 conditions 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,*

$$\mathbf{P}(\boldsymbol{\beta} \in I_{n,\alpha}) \rightarrow \alpha \quad \text{as } n \rightarrow \infty, \tag{3.1}$$

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

(ii) *Suppose that  $\{j : \beta_j \neq 0\}$  is nonempty. Then, (3.1) holds for all  $\alpha \in (0, 1)$ .*

Part (i) of Corollary 3.2 requires a mild condition on the  $\alpha$ , as the limit distribution of  $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$  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.

### 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 Condition (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,*

$$\begin{aligned} \mathbf{E}_*(\mathbf{T}_n^{**}) &\rightarrow \mathbf{E}(\mathbf{T}_\infty), \quad \text{and} \\ \left(\mathbf{Var}_*(\mathbf{T}_n^{**})\right)_{p \times p} &\rightarrow \left(\mathbf{Var}(\mathbf{T}_\infty)\right)_{p \times p}, \end{aligned} \tag{3.2}$$

with probability 1.

Note that for  $\lambda_0 = 0$  in (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 (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 an MSE estimate to the Lasso estimate and quantify the associated uncertainty, for all possible values of  $\beta$ , and thereby removes the limitations of the existing methods of MSE estimation.

## 4 Data based choice of the regularization parameter

In this section, we consider the problem of choosing the optimal penalty parameter  $\lambda_0$  in a data-based manner. In Section 4.1, we formalize the notion of the optimal parameter through a natural reparametrization and in Section 4.2, we describe a data based method for choosing the optimal regularization parameter based on the modified bootstrap method.

### 4.1 The optimal regularization parameter

Let

$$V(\mathbf{u}) = -2\mathbf{u}'\mathbf{W} + \mathbf{u}'\mathbf{C}\mathbf{u} + \lambda_0 \sum_{j=1}^p \left[ u_j \text{sgn}(\beta_j) \mathbf{1}(\beta_j \neq 0) + |u_j| \mathbf{1}(\beta_j = 0) \right], \tag{4.1}$$

where  $\mathbf{W} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{C})$ ,  $\mathbf{C} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$  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 conditions (C.1)-(C.3), the work of Knight and Fu (2000) implies that

$$\mathbf{T}_n \xrightarrow{d} \mathbf{T}_\infty,$$

where  $\mathbf{T}_\infty = \underset{\mathbf{u} \in \mathbb{R}^p}{\operatorname{argmin}} V(\mathbf{u})$ . 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  $\lambda_n = \lambda_0 \sqrt{n}$ ,  $\lambda_0 \in [0, \infty)$ . Note that the mean squared error (MSE) of  $\hat{\boldsymbol{\beta}}_n$  for estimating can be expressed as  $\operatorname{MSE}(\hat{\boldsymbol{\beta}}_n) \equiv \mathbf{E} \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|^2 = n^{-1} \mathbf{E} \|\mathbf{T}_n\|^2$ . 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  $n \operatorname{MSE}(\hat{\boldsymbol{\beta}}_n)$  converges to the MSE of the limiting random variable  $\mathbf{T}_\infty$ , we define the *optimal penalization parameter*  $\lambda_0^{opt}$  as

$$\lambda_0^{opt} \equiv \operatorname{argmin} \Gamma(\lambda_0), \quad (4.2)$$

where  $\Gamma(\lambda_0) = \mathbf{E} \|\mathbf{T}_\infty\|^2$ , the MSE of the limit distribution of  $\mathbf{T}_n$  with  $\lambda_n = \lambda_0 \sqrt{n}$ ,  $\lambda_0 \in [0, \infty)$ . 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  $\lambda_0^{opt}$ .

## 4.2 Data based selection of the optimal regularization parameter

We now describe a data based method for estimating  $\lambda_0^{opt}$  of (4.2) using the modified bootstrap. To that end, first we define the bootstrap estimator of the function  $\Gamma(\lambda_0)$ ,  $\lambda_0 \in [0, \infty)$ , as

$$\hat{\Gamma}_n(\lambda_0) \equiv \mathbf{E}_* \|\mathbf{T}_n^{**}\|^2, \quad \lambda_0 \in [0, \infty).$$

Note that by Theorem 3.1,  $\hat{\Gamma}_n(\lambda_0)$  is a strongly consistent estimator of  $\Gamma(\lambda_0)$ . Therefore, we replace  $\Gamma(\lambda_0)$  in (4.2) by its bootstrap estimator  $\hat{\Gamma}_n(\lambda_0)$  and define the bootstrap estimator of the target penalization parameter  $\lambda_0^{opt}$  by

$$\hat{\lambda}_0^{opt} = \operatorname{argmin} \hat{\Gamma}_n(\lambda_0). \quad (4.3)$$

Since the modified bootstrap provides a consistent approximation to the MSE of the Lasso estimator (cf. Theorem 3.3), (4.3) 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  $\boldsymbol{\beta}$  and the sizes of its non-zero components, the choice of the truncation parameter  $a_n$  in (2.4), etc. In the next section, we describe the computational aspects of  $\hat{\lambda}_0^{opt}$  and explore its finite sample properties.

## 5 Numerical results

In this numerical study, we intend to compare the performances of the modified and usual bootstrap procedures. We will compare the coverage accuracy of confidence sets for  $\boldsymbol{\beta}$  obtained by the two procedures in finite samples. We will also study the accuracy of the variance estimation procedure by comparing the bias and MSE's of the bootstrap based variance-covariance estimates. For clarity of exposition, we first explore the data based method of choosing the *optimal*  $\lambda_0$  in Section 5.1 below.

## 5.1 Modified bootstrap based choice of optimal penalization

Consider the Lasso estimator  $\widehat{\beta}_n$  in (1.2), which can be written as  $\widehat{\beta}_n(\lambda_0)$ , to emphasize its dependence on the penalizing parameter  $\lambda_0 \equiv \lambda_n/\sqrt{n}$ , and similarly, write  $\mathbf{T}_n(\lambda_0)$  for  $\mathbf{T}_n$ ,  $n \in \mathbb{N}$ . For practical purposes, it is important to know an optimal choice of the penalizing parameter  $\lambda_0$ . We now explore the feasibility of using such a modified bootstrap based estimator through our numerical study, .

We first provide the details of our simulation setup. For a given  $\mathbf{X}_n$ ,  $\beta$ , error distribution  $F$  and any integer  $M$ , we construct the  $m^{\text{th}}$  dataset of size  $n$  :  $\mathcal{S}_m = \left\{ \left( y_i^{(m)}, \mathbf{x}_i \right) : 1 \leq i \leq n \right\}$  for all  $1 \leq m \leq M$  (where  $M$  denotes the number of simulation runs), using the regression model (1.1). We consider a grid of  $\lambda_0$  values, denoted as  $\{\lambda_{0,k} : 1 \leq k \leq K\}$  and define  $\lambda_{n,k} = \sqrt{n}\lambda_{0,k}$ . For the  $m^{\text{th}}$  data set  $\mathcal{S}_m$  and a given value  $\lambda_{0,k}$ , the usual Lasso and the thresholded Lasso estimates of  $\beta$  are denoted as  $\widehat{\beta}_n(k, m)$  and  $\widetilde{\beta}_n(k, m)$ , respectively. In all of our simulation study, we set the truncation parameter  $a_n$  in (2.4) as  $cn^{-1/4}$  for various choices of  $c$ .

Let  $\mathbf{T}_n(k, m) = \sqrt{n} \left( \widehat{\beta}_n(k, m) - \beta \right)$ . For each  $\mathcal{S}_m$  and  $\lambda_{0,k}$ , we construct  $B$  bootstrap replicates for the usual and modified bootstrap versions, from which we obtain  $\mathbf{T}_n^*(b, k, m)$  and  $\mathbf{T}_n^{**}(b, k, m)$ ,  $1 \leq b \leq B$ . A simulation based estimate of  $\lambda_0^{\text{opt}}$  (cf. (4.2)) can be found using the  $M$  data sets  $\{\mathcal{S}_m : 1 \leq m \leq M\}$ ,

$$\lambda_0^{\text{opt}}(M, K) = \underset{k}{\operatorname{argmin}} \sum_{j=1}^p \left( \frac{1}{M} \sum_{m=1}^M T_{n,j}^2(k, m) \right), \quad (5.1)$$

where the minimum is taken over the grid of values  $\{\lambda_{0,k} : 1 \leq k \leq K\}$ . A bootstrap estimate of  $\lambda_0^{\text{opt}}$ , given by (4.3) can be similarly approximated by using the  $B$  bootstrap replicates within each fixed sample  $\mathcal{S}_m$ , as

$$\widehat{\lambda}_0^{\text{opt}}(m, B, K) = \underset{k}{\operatorname{argmin}} \sum_{j=1}^p \left( \frac{1}{B} \sum_{b=1}^B \left( T_{n,j}^{**}(b, k, m) \right)^2 \right), \quad m = 1, \dots, M.$$

The minimal  $\lambda_0$  can be found on the grid  $\{\lambda_{0,k} : 1 \leq k \leq K\}$  or over a smaller set of points centered around the estimated true value  $\lambda_0^{\text{opt}}(M, K)$ . As stated earlier, due to the consistency of  $\mathbf{T}_n^{**}$ , the modified bootstrap based estimate in (4.3) is expected to be very close to the true optimal value  $\lambda_0^{\text{opt}}$ . The MSE of  $\widehat{\lambda}_0^{\text{opt}}$  can be approximated using the  $M$  data-sets, as

$$\operatorname{MSE}(\widehat{\lambda}_0^{\text{opt}}) = \frac{1}{M} \sum_{m=1}^M \left( \widehat{\lambda}_0^{\text{opt}}(m, B, K) - \lambda_0^{\text{opt}}(M, K) \right)^2,$$

with a similar definition for the bias. In our numerical study, we consider different values of  $n$  with  $\beta$  fixed and errors  $\epsilon \sim N(0, 1)$ . The columns of  $\mathbf{X}_n$  were generated as *iid* samples from a  $N(0, 1)$  distribution at each value of  $n$ , and was held fixed for all  $M$  simulation runs. For this simulation, we found  $\lambda_0^{\text{opt}}(M, K) (\approx \lambda_0^{\text{opt}})$  separately, using 5000 simulated datasets  $\mathcal{S}_m$  and over 200 grid points of

$\lambda_0$  values. The MSE and bias were estimated using the above Monte-Carlo approximation. For the bootstrap part, we used  $M = 500$ ,  $B = 500$  and a search grid of 120 points around the estimated value  $\lambda_0^{opt}(M, K)$ .

In Table 1 we compare the performance of the modified bootstrap based optimal value  $\widehat{\lambda}_0^{opt}$  in terms of its bias and MSE at different thresholding constants  $c$  and sample sizes  $n$ . Figure 1 shows the boxplots of the deviation of the  $\widehat{\lambda}_0^{opt}$  from the true optimal value at the same value of  $c$  and  $n = 600$ . From Table 1, it can be seen that for a fixed sample size  $n$ , the modified bootstrap has better performance with increasing values of the thresholding constant  $c$ . The magnitude of bias apparently decreases as  $n$  increases but less noticeably than the MSE. The thresholding value  $c$  has asymptotically negligible effect, but in the finite sample case it creates substantial differences, with higher values of  $c$  showing better results. The results suggest that modified bootstrap based optimal  $\lambda_0$  can be used to select the penalizing constant  $\lambda_0^{opt}$ , which in turn will minimize the criterion function in (4.2).

## 5.2 Coverage accuracy and variance estimation

Here we consider the performance of the modified bootstrap procedures in terms of coverage accuracy of confidence regions. For any  $\mathcal{S}_m$  and  $\lambda_{0,k}$ , an  $\alpha$ -level ( $0 < \alpha < 1$ ) confidence ball for  $\beta$  based on the modified bootstrap procedure is given by

$$I_{n,\alpha}(k, m) = \left\{ \mathbf{t} : \|\mathbf{t} - \widetilde{\beta}_n(k, m)\| \leq n^{-1/2} \widehat{t}_{n,\alpha}(k, m) \right\},$$

where  $\widehat{t}_{n,\alpha}(k, m)$  is the  $\alpha$ -quantile of  $\|\mathbf{T}_n^{**}(k, m)\|$ . The  $\alpha$ -quantiles can be estimated empirically using the  $B$  bootstrap replicates. For comparison, we simulated the coverage probabilities of the confidence set  $I_{n,\alpha}(k, m)$  at the following *triplet* of  $\lambda_0$  values:  $\{\lambda_0^{opt} - 0.2, \lambda_0^{opt}, \lambda_0^{opt} + 0.2\}$ , *i.e.* at the true value and at two neighboring points. Table 2 provides the empirical coverage probabilities at different sample sizes and  $\alpha = 0.9$ . In all cases the modified bootstrap procedure performs better than the usual bootstrap, though both of them have empirical coverage higher than the nominal value ( $\alpha = 0.9$ ). The simulated results suggest that the modified bootstrap based confidence region has empirical coverage slightly more than the nominal value, with better performance as the  $\lambda_0$  value increases.

We now consider the accuracy of the moment estimates obtained by the competing bootstrap methods. Let  $\sigma_n(i, j : \lambda_0) \equiv \mathbf{E}(T_{n,i}(\lambda_0) T_{n,j}(\lambda_0))$ . As in the earlier case, we only study the accuracy at the previously described *triplet* of  $\lambda_0$  values. For a fixed  $\mathcal{S}_m$  and  $\lambda_{0,k}$ , the modified bootstrap estimate of the  $(i, j)$ -th

product-moment can be found by using the  $B$ -bootstrap replicates,

$$\widehat{\sigma}_{2,n,B}(i, j : k, m) = \frac{1}{B} \sum_{b=1}^B T_{n,i}^{**}(b, k, m) T_{n,j}^{**}(b, k, m), \quad 1 \leq i, j \leq p.$$

The Monte-Carlo approximation to  $\mathbf{E}\left(T_{n,i}(k) T_{n,j}(k)\right)$  is

$$\check{\sigma}_{2,n,M}(i, j : k) = \frac{1}{M} \sum_{m=1}^M T_{n,i}(k, m) T_{n,j}(k, m), \quad 1 \leq i, j \leq p.$$

Using the above, the MSE of the  $(i, j)$ -th bootstrap product-moment estimate can be approximated by

$$\text{MSE}\left(\widehat{\sigma}_{2,n}(i, j : k)\right) = \frac{1}{M} \sum_{m=1}^M \left(\widehat{\sigma}_{2,n,B}(i, j : k, m) - \check{\sigma}_{2,n,M}(i, j : k)\right)^2, \quad (5.2)$$

and similarly the bias estimate can be found. Analogously all of the above can be defined for the usual bootstrap case by using the subscript 1.

As a measure of overall accuracy of the bootstrap product moment estimates, we can consider the sum of MSE's over all the  $p(p+1)/2$  upper diagonal elements of the matrix of MSE estimates of the form (5.2). For any  $\lambda_{0,k}$  define,

$$\tau_M(a, \lambda_{0,k}) = \sum_{j \geq i} \text{MSE}\left(\widehat{\sigma}_{a,n}(i, j : k)\right), \quad (5.3)$$

with  $a = 1$  or  $2$ , for the usual and modified case respectively. Accuracy of component wise product-moment estimates can be judged from the corresponding bias and MSE values. Table 3 compares the above overall accuracy criterion for the usual and modified cases, at different sample sizes and at the *triplet* of  $\lambda_0$  values. The overall performance of the modified bootstrap is found to be superior than the usual bootstrap. With increasing sample size, the magnitude of error decreases for both bootstrap procedures, but the relative performance of the modified version is better. There is some evidence that at different sample sizes, the best overall performance of the modified bootstrap is attained at the *optimal*  $\lambda_0$  value, with slightly worse performance at the neighboring points.

We also study the accuracy of some of the component wise covariance estimates. We will be interested in pairs of covariates  $(i, j)$ , (with  $i = j$  and  $i \neq j$ ) for which the corresponding  $\beta_i$  and  $\beta_j$  are both non-zero, both zero or one zero and the other non-zero. We will choose the following five specific covariate pairs  $(i, j)$ :  $\{(1, 1), (4, 4), (1, 3), (2, 4), (3, 6)\}$  corresponding to the regression parameter  $\boldsymbol{\beta} = (5, 0, -12, 0, 4, 0)'$ . Table 4 studies the MSE's of the bootstrap based estimates of  $\sigma_n(i, j : \lambda_0)$  for these selected pairs of  $(i, j)$ 's at the same *triplet* of  $\lambda_0$  values over different sample sizes.

For the covariate pairs  $(1, 1)$  and  $(1, 3)$ , which consist of non-zero components of  $\boldsymbol{\beta}$ , the usual and modified bootstrap have very similar performances over all sample sizes and over all  $\lambda_0$  values, with the usual bootstrap being slightly better in some cases. This suggests that modified bootstrap does not have any effect when the underlying  $\beta_j$  is far from zero. The difference between the two bootstrap methods is evident when we see the other pairs  $\{(4, 4), (2, 4), (3, 6)\}$  which consists of the zero components of  $\boldsymbol{\beta}$ . The modified bootstrap performs remarkably well in estimating the variance and covariances when the underlying regression coefficient is zero.

## 6 Real data examples

In this section we apply the modified bootstrap method on two different real data sets. The first data is the Prostrate cancer data set, which has been used in his original LASSO paper by Tibshirani (1996). The second is the Iowa wheat data set of Draper and Smith (1998).

### 6.1 Prostrate cancer data

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 his webpage. 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`), seminal vesicle invasion (`svi`), log(capsular penetration) (`lcp`), Gleason score (`gleason`) and percentage Gleason scores 4 or 5 (`pgg45`). The predictors were standardized and a linear model was fitted against the `lpsa` values. The whole data set comprising of  $n = 97$  observations have been used in our numerical computations. The penalization parameter was chosen as  $\lambda_n \approx 0.69$ . The thresholding value used is  $a_n = cn^{-1/4}$  with  $c = 1.25$ , which corresponds to a thresholding value  $\approx 0.4$ . We compute the usual and thresholded Lasso estimates and modified bootstrap based variance estimates for each predictor. We also construct symmetric 90% modified bootstrap 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  $|T_{n,j}^{**}|$  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  $j^{\text{th}}$  predictor.

The upper and lower confidence limits and the results of the componentwise bootstrap (at level  $\alpha = 0.1$ ) are summarized in Table 5. It appears that the covariates `lcavol`, `lweight`, `lbph` and `svi` have nontrivial effect on the response variable `lpsa`, the rest of the variables were judged insignificant at level  $\alpha = 0.1$ . The results in Table 5 also show that indeed non-zero variance estimates can be obtained using the modified bootstrap method, even for predictors with  $\hat{\beta}_{n,j} = 0$ . The choice of the value of  $\lambda_0$  (or equivalently  $\lambda_n$ ) plays an important role, because very large of  $\lambda_0$  will ultimately shrink both  $\tilde{\beta}_{n,j}$  and  $\beta_{n,j}^{**}$  to zero, which will result in zero variance estimates. Similarly important is the choice of the thresholding value  $a_n$ , which will cause similar effects on  $\tilde{\beta}_n$ .

## 6.2 Iowa wheat data

The data gives the pre-season and three growing months precipitation, the mean temperatures for the three growing months and harvest month, the year, and the yield of wheat for the state of Iowa, for the years 1930–1962. The dataset consists of the following components: year of measurement (`year`), pre-season rainfall (`rain0`), mean temp. for the first, second and third growing months: `temp1`, `temp2` and `temp3` respectively; rainfall for the first, second and third growing months: `rain1`, `rain2` and `rain3` respectively; mean temp. for the harvest month (`temp4`). Rainfall is in inches and temperature is in degrees Fahrenheit. The variable of interest is the yield of wheat in Iowa for the given year (`yield`) (in bushels/acre). The data set can also be obtained from the data sets included in the R package `lasso2`.

In our analysis, we will not include the year of measurement (`year`) as a predictor, which implies we are using eight predictors on a dataset of size  $n = 33$ . The predictors were standardized and a linear model was fitted against the `yield` values. The thresholding value was chosen as  $a_n = 10n^{-1/4} \approx 4.172$ . The penalization parameter was selected as  $\lambda_0 = 1$ , which implies  $\lambda_n \approx 5.74$ . The relatively *large* choice of the thresholding parameter  $a_n$  reflects the fact that the magnitude of Lasso coefficients decreased very slowly as a function of the penalization parameter, with smaller values of  $\lambda_n$  yielding estimates comparable to the OLS. We apply the modified bootstrap method at the selected value of  $\lambda_0$  and  $a_n$  and obtain the usual and thresholded Lasso estimates of the true parameter, componentwise variance estimates, 90% confidence intervals and corresponding tests for the hypothesis  $H_0 : \beta_j = 0$ . Based on the bootstrap test as seen in Table 6, only the variables `rain2` and `temp3` have coefficients that are significantly different from zero at level  $\alpha = 0.1$ . The results in Table 6 show that due to the relatively large variances, the confidence intervals also have larger width. It should be noted that although some of the Lasso estimates are zero, the corresponding variance estimates are nonzero, and we are still able to conduct valid inference using the modified bootstrap technique, thereby overcoming earlier drawbacks.

## 7 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 7.1.** *Let  $s_n^2 = n^{-1} \sum_{j=1}^n (\tilde{e}_j - \tilde{e}_n)^2$  and  $m_{3,n} = n^{-1} \sum_{j=1}^n |\tilde{e}_j - \tilde{e}_n|^3$ . Assume that*

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

and condition (C.3) holds. Then

$$|s_n^2 - \sigma^2| + n^{-1/2} m_{3,n} \xrightarrow{a.s.} 0,$$

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

*Proof of Lemma 7.1.* 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 (2009),

$$\|\mathbf{T}_n\| = O(\log n), \quad \text{with probability 1.} \quad (7.1)$$

Hence, using (7.1) and the definition of  $\tilde{\boldsymbol{\beta}}_n$ , we have

$$\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| = O\left(n^{-1/2} \log n\right) \quad \text{with probability 1.} \quad (7.2)$$

By (7.2)

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

Since  $\sigma_n^2 \rightarrow \sigma^2$  almost surely, it follows that  $|s_n^2 - \sigma^2| = o(1)$  with probability 1. Next consider  $m_{3,n}$ . Using the condition on the  $\mathbf{x}_i$ 's, we get

$$\max_{1 \leq i \leq n} \|\mathbf{x}_i\| \leq \left( \sum_{i=1}^n \|\mathbf{x}_i\|^2 \right)^{1/2} = O\left(n^{1/2}\right). \quad (7.3)$$

Hence, by the Marcinkiewicz-Zygmund Strong Law of Large Numbers, (7.2), and (7.3), 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 7.2.** *Suppose that conditions (C.1) and (C.3) hold. Then*

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

*Proof of Lemma 7.2.* By Lemma 7.1, 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  $\mathbf{t} = (t_1, \dots, t_p)' \in \mathbb{R}^p$ ,  $\mathbf{t} \neq \mathbf{0}$ . Let  $s_n^2(\mathbf{t}) = \mathbf{Var}_*(n^{-1/2} \sum_{i=1}^n \mathbf{t}' \mathbf{x}_i e_i^{**})$ . Note that by the definition of  $A$ ,  $s_n^2(\mathbf{t}) \rightarrow \mathbf{t}' \mathbf{C} \mathbf{t} \tau^2 \in (0, \infty)$ . By the Berry-Esseen Theorem for independent (but possibly non-identically distributed random variables) (cf. Chapter 12, [Bhattacharya and Ranga Rao \(1986\)](#)),

$$\begin{aligned}
\sup_{x \in \mathbb{R}} \left| \mathbf{P}_* \left( n^{-1/2} \sum_{i=1}^n \mathbf{t}' \mathbf{x}_i e_i^{**} \leq x \right) - \Phi(x/s_n(\mathbf{t})) \right| &\leq (2.75) \frac{\sum_{i=1}^n \mathbf{E}_* \left| n^{-1/2} \mathbf{t}' \mathbf{x}_i e_i^{**} \right|^3}{\left( \sum_{i=1}^n \mathbf{E}_* \left| n^{-1/2} \mathbf{t}' \mathbf{x}_i e_i^{**} \right|^2 \right)^{3/2}} \\
&\leq C \frac{\|\mathbf{t}\|^3 n^{-3/2} \sum_{i=1}^n \|\mathbf{x}_i\|^3 \mathbf{E}_* |e_i^{**}|^3}{s_n^3(\mathbf{t})} \\
&= o(1).
\end{aligned}$$

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

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

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

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

$$\mathbf{W}_n^{**} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i e_i^{**}.$$

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

$$V_n^{**}(\mathbf{u}) = \mathbf{u}'\mathbf{C}_n\mathbf{u} - 2\mathbf{u}'\mathbf{W}_n^{**} + \lambda_n \sum_{j=1}^p \left( \left| \tilde{\beta}_{n,j} + \frac{u_j}{\sqrt{n}} \right| - |\tilde{\beta}_{n,j}| \right). \quad (7.4)$$

Let  $A \in \mathcal{F}$  be such that  $\mathbf{P}(A) = 1$  and for every  $\omega \in A$ , (7.1), (7.2) hold and

$$\mathcal{L}\left(\mathbf{W}_n^{**} \mid \mathcal{E}\right)(\omega) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{C}). \quad (7.5)$$

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  $n \geq N$ ,

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

for all  $\mathbf{u} \in \mathbb{R}^p$  with  $|u_j| \leq \hat{\beta}_{n,j}\sqrt{n}$  for  $j = 1, \dots, p_0$ . Now using (7.5) and the arguments as in the unbootstrapped case (cf. Knight and Fu (2000)), one can establish the weak convergence

$$\mathcal{L}\left(\mathbf{V}_n^{**}(\cdot) \mid \mathcal{E}\right)(\omega) \xrightarrow{d} \mathcal{L}(\mathbf{V}(\cdot))$$

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  $\mathcal{L}(\mathbf{T}_n^{**} \mid \mathcal{E})(\omega) \xrightarrow{d} \mathcal{L}(\mathbf{T}_\infty)$  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  $F_n \xrightarrow{d} F$  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  $\left\{ \|\mathbf{T}_n^{**}\|^2 \right\}_{n \geq 1}$  is uniformly integrable almost surely, i.e.,

$$\lim_{\alpha \rightarrow \infty} \sup_{n \geq 1} \mathbf{E}_* \|\mathbf{T}_n^{**}\|^2 \mathbf{1}(\|\mathbf{T}_n^{**}\| \geq \alpha) = 0, \quad a.s. (\mathbf{P}).$$

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

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

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

$$\mathbf{T}_n^{**} \in \left\{ \mathbf{u} \mid \|\mathbf{u}\| \leq \frac{2}{\eta_{1,n}} \left( 2w_0 + \frac{\lambda_n}{\sqrt{n}}\sqrt{p} \right) \right\}.$$

This leads to the observation that

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

For any fixed  $t_0 = 2\eta_{1,n}^{-1} \left( 2w_0 + \frac{\lambda_n}{\sqrt{n}}\sqrt{p} \right)$ , we obtain  $w_0 = \frac{1}{2} \left( \frac{t_0\eta_{1,n}}{2} - \frac{\lambda_n}{\sqrt{n}}\sqrt{p} \right)$ . And using the above relation we can write

$$\left[ \|\mathbf{T}_n^{**}\| > t_0 \right] \Rightarrow \left[ \|\mathbf{W}_n^{**}\| \geq \frac{1}{2} \left( \frac{t_0\eta_{1,n}}{2} - \frac{\lambda_n}{\sqrt{n}}\sqrt{p} \right) \right]. \quad (7.6)$$

Using the conditions (C.1) and (C.2) we can say,  $\eta_{1,n} \rightarrow \eta_1 (> 0)$  and  $\frac{\lambda_n}{\sqrt{n}}\sqrt{p} \rightarrow \lambda_0\sqrt{p}$ , which is finite. Using (7.6), we can write

$$\begin{aligned}
\mathbf{E}_* \|\mathbf{T}_n^{**}\|^2 \mathbf{1}(\|\mathbf{T}_n^{**}\| \geq \alpha) &\leq \left[ \alpha^2 \mathbf{P}_* (\|\mathbf{T}_n^{**}\| > \alpha) + \int_{\alpha}^{\infty} t \mathbf{P}_* (\|\mathbf{T}_n^{**}\| > t) dt \right] \\
&\leq 4 \int_{\alpha/2}^{\infty} t \mathbf{P}_* (\|\mathbf{T}_n^{**}\| > t) dt \\
&\leq 4 \int_{\alpha/2}^{\infty} t \mathbf{P}_* \left( \|\mathbf{W}_n^{**}\| \geq \frac{1}{2} \left( \frac{t\eta_{1,n}}{2} - \frac{\lambda_n}{\sqrt{n}}\sqrt{p} \right) \right) dt \\
&\leq 4 \int_{\alpha/2}^{\infty} t \mathbf{P}_* \left( \|\mathbf{W}_n^{**}\| \geq \frac{t\eta_1}{4} \right) dt, \quad (7.7)
\end{aligned}$$

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 (7.5) and the continuous mapping theorem, we can say that for any fixed  $c \in (0, \infty) \setminus D$ , where  $D$  is a countable set,

$$\|\mathbf{W}_n^{**}\|^2 \mathbf{1}(\|\mathbf{W}_n^{**}\| \geq c) \xrightarrow{d} \|\mathbf{W}_{\infty}\|^2 \mathbf{1}(\|\mathbf{W}_{\infty}\| \geq c), \quad a.s. (\mathbf{P}),$$

and using Lemma 7.1 we have

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

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

$$\mathbf{E}_* \|\mathbf{W}_n^{**}\|^2 \mathbf{1}(\|\mathbf{W}_n^{**}\| \geq c) \xrightarrow{a.s.} \mathbf{E} \|\mathbf{W}_\infty\|^2 \mathbf{1}(\|\mathbf{W}_\infty\| \geq c) \quad a.s. (\mathbf{P}), \quad (7.8)$$

for every  $c \in (0, \infty) \setminus D$ . Further, the right side of (7.8) 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 (7.7) 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  $\{\|\mathbf{T}_n^{**}\|^2\}_{n \geq 1}$ , we have

$$\mathbf{E}_* \|\mathbf{T}_n^{**}\|^2 \rightarrow \mathbf{E} \|\mathbf{T}_\infty\|^2, \quad a.s. (\mathbf{P}).$$

Now note that for any fixed  $1 \leq j \leq p$  and any  $\alpha > 0$ , we have

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

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

$$\mathbf{E}_* |T_{n,j}^{**}|^2 \rightarrow \mathbf{E} |T_{\infty,j}|^2, \quad a.s. (\mathbf{P}). \quad (7.9)$$

Since  $L_2$  convergence implies  $L_1$  convergence,  $\mathbf{E}_* T_{n,j}^{**} \rightarrow \mathbf{E} T_{\infty,j}$ , almost surely ( $\mathbf{P}$ ). Hence, for all  $1 \leq j \leq p$ ,

$$\mathbf{Var}_* (T_{n,j}^{**}) \rightarrow \mathbf{Var} (T_{\infty,j}) \quad a.s. (\mathbf{P}). \quad (7.10)$$

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

$$T_{n,j}^{**} T_{n,k}^{**} \xrightarrow{d} T_{\infty,j} T_{\infty,k}, \quad \text{almost surely } (\mathbf{P}),$$

for any  $j \neq k$ . Also note that

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

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

$$\mathbf{E}_* (T_{n,j}^{**} T_{n,k}^{**}) \rightarrow \mathbf{E} (T_{\infty,j} T_{\infty,k}), \quad a.s. (\mathbf{P}). \quad (7.11)$$

Combining (7.10) and (7.11), 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 (7.9).  $\square$

## References

- Bach, F. (2008). Bolasso: model consistent lasso estimation through the bootstrap. *CoRR*, abs/0804.1302.
- Bhattacharya, R. N. and Ranga Rao, R. (1986). *Normal Approximation and Asymptotic Expansions*. Robert E. Krieger Publishing Co. Inc., Melbourne, FL. Reprint of the 1976 original.
- Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2008). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*
- Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *Ann. Statist.*, 35(6):2313–2351.
- Chatterjee, A. and Lahiri, S. N. (2009). Asymptotic properties of the residual bootstrap for lasso estimators. (*submitted*), preprint available at <http://www.stat.tamu.edu/cha/lasso-residual-bootstrap-pams-09.pdf>.
- Draper, N. R. and Smith, H. (1998). *Applied regression analysis*. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons Inc., New York, third edition.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):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. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.
- Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.*, 32(3):928–961.
- Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist.*, 9(6):1218–1228.
- Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.*, 7(3):397–416.
- Huang, J., Horowitz, J. L., and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.*, 36(2):587–613.
- Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378.
- Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

- Meinshausen, N. and Yu, B. (2008). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*
- Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2):319–337.
- Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. Radical prostatectomy treated patients. *J. Urol.*, 141(5):1076–1083.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.
- Wainwright, M. J. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. Technical report, Dept. of Statistics, UC Berkeley.
- Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.
- Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36:1567.
- Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.

Table 1: Table comparing the bias and MSE of the bootstrap based optimal value  $\widehat{\lambda}_0^{opt}$  at different sample sizes  $n$  and thresholding values  $a_n = cn^{-1/4}$  with  $c = \{1/5, 1, 10\}$  (cf. (2.4)). Here,  $\epsilon \sim N(0, 1)$ ,  $\beta = (5, 0, -12, 0, 4, 0)'$ .

Comparison of bias and MSE of $\widehat{\lambda}_0^{opt}$ at different $n$ and $c$							
$n$	$\lambda_0^{opt*}$	$c = 1/5$		$c = 1$		$c = 10$	
		bias	MSE	bias	MSE	bias	MSE
100	0.4888	-0.0832	0.0184	-0.0535	0.0091	-0.0520	0.0083
300	0.4157	0.0060	0.0122	0.0175	0.0068	0.0238	0.0061
600	0.4960	-0.0226	0.0150	-0.0143	0.0060	-0.0162	0.0065
2000	0.4599	-0.0102	0.0100	-0.0120	0.0066	-0.0125	0.0062

\* true optimal  $\lambda_0$  (which depends on the choice of  $\mathbf{X}_n$ )

Table 2: Empirical coverage probabilities based on the modified bootstrap confidence region  $I_{n,\alpha}$ , at the *triplet* of  $\lambda_0$  values:  $\{\lambda_0^{opt}, \lambda_0^{opt} \pm 0.2\}$  and at different sample sizes  $n$ . Here  $\epsilon \sim N(0, 1)$ ,  $\beta = (5, 0, -12, 0, 4, 0)'$  and a fixed thresholding value  $a_n = cn^{-1/4}$  with  $c = 1$ .

Coverage probabilities using $I_{n,\alpha}$ at $\alpha = 0.9$					
$n = 100$		$n = 600$		$n = 2000$	
$\lambda_0$	Coverage	$\lambda_0$	Coverage	$\lambda_0$	Coverage
0.259	0.952	0.258	0.966	0.206	0.982
0.459*	0.938	0.458*	0.952	0.406*	0.956
0.659	0.924	0.658	0.938	0.606	0.936

\* true optimal  $\lambda_0$

n = 600 : boxplot showing deviation of modified bootstrap based optimal  $\lambda_0$  values from the true optimal value at various thresholding constants c

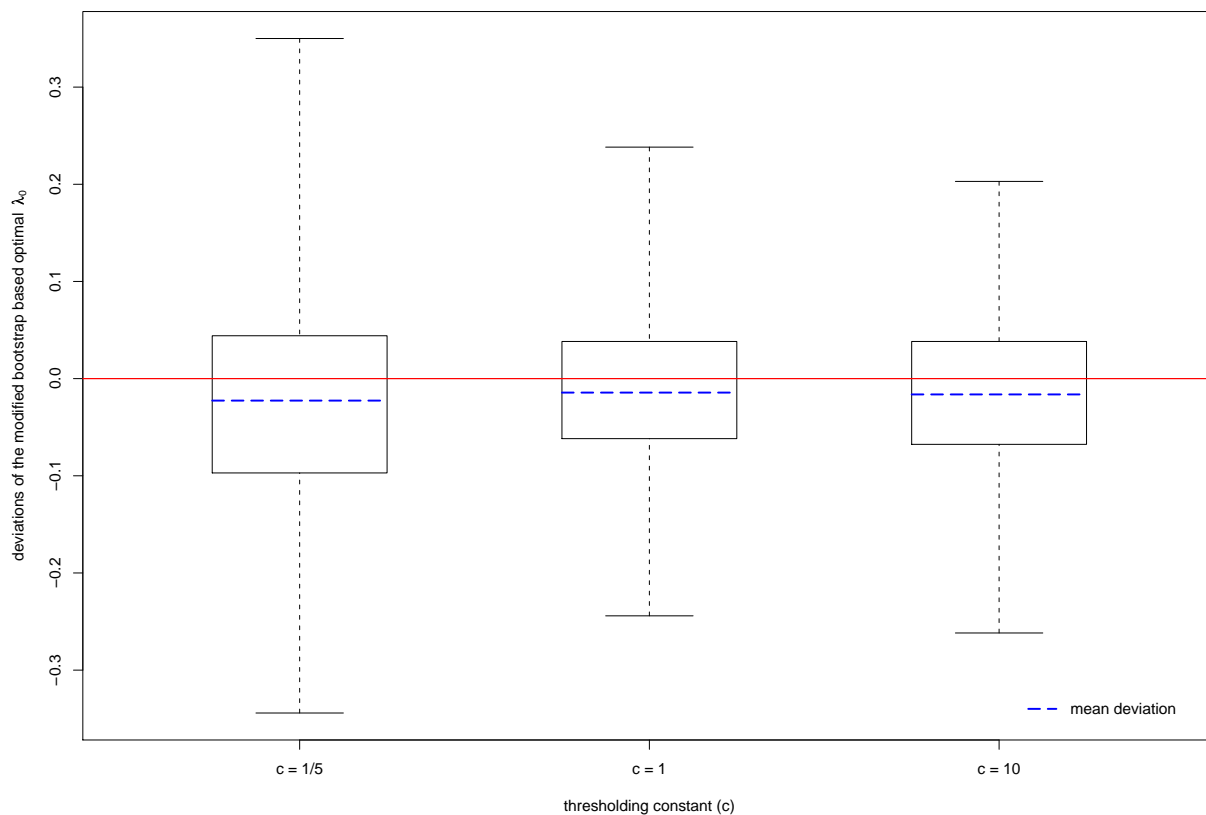


Figure 1: Figure comparing the **deviations** of the modified bootstrap based optimal  $\lambda_0$  ( $\widehat{\lambda}_0^{opt}$ ) from the true optimal value ( $\lambda_0^{opt}$ ) at thresholding values  $a_n = cn^{-1/4}$  with  $c = \{1/5, 1, 10\}$ . Here  $n = 600$ ,  $\epsilon \sim N(0, 1)$ ,  $\beta = (5, 0, -12, 0, 4, 0)'$ .

Table 3: Table comparing the overall MSE criterion  $\tau_M(., \lambda_0)$  (cf. (5.3)) for the usual and modified bootstrap methods, at the *triplet* of  $\lambda_0$  values:  $\{\lambda_0^{opt}, \lambda_0^{opt} \pm 0.2\}$  and at different sample sizes  $n$ . Here  $\epsilon \sim N(0, 1)$ ,  $\beta = (5, 0, -12, 0, 4, 0)'$  and a fixed thresholding value  $a_n = cn^{-1/4}$  with  $c = 1$ .

Overall accuracy of bootstrap moment estimates: comparison using $\tau_M(a, \lambda_0)$								
$n = 100$			$n = 600$			$n = 2000$		
$\lambda_0$	Usual	Modified	$\lambda_0$	Usual	Modified	$\lambda_0$	Usual	Modified
0.259	0.358	0.277	0.258	0.182	0.105	0.206	0.121	0.0734
0.459*	0.491	0.250	0.458*	0.284	0.0989	0.406*	0.209	0.0651
0.659	0.687	0.249	0.658	0.434	0.106	0.606	0.348	0.0679

\* true optimal  $\lambda_0$

Table 4: Table comparing the MSE's of usual and modified (below in **bold**) bootstrap based estimates of  $\sigma_n(i, j : \lambda_0)$  (cf. (5.2)) for five specific pairs of covariates  $(i, j)$ , at the *triplet* of  $\lambda_0$  values for different sample sizes  $n$ . Here  $\epsilon \sim N(0, 1)$ ,  $\beta = (5, 0, -12, 0, 4, 0)'$  and fixed thresholding value  $a_n = cn^{-1/4}$  with  $c = 1$ .

Component wise MSE's of bootstrap moment estimates						
$n$	$\lambda_0$	Components $(i, j)$				
		(1, 1)	(4, 4)	(1, 3)	(2, 4)	(3, 6)
100	0.259	0.0217	0.034	0.0063	0.00172	0.0045
		<b>0.0205</b>	<b>0.0239</b>	<b>0.0064</b>	<b>0.001</b>	<b>0.00274</b>
	0.459*	0.0217	0.0537	0.00547	0.00272	0.0149
		<b>0.0213</b>	<b>0.0162</b>	<b>0.00458</b>	<b>0.000495</b>	<b>0.00249</b>
	0.659	0.0234	0.069	0.00607	0.00371	0.0327
		<b>0.0255</b>	<b>0.0078</b>	<b>0.00457</b>	<b>0.000331</b>	<b>0.00265</b>
300	0.22	0.0145	0.0183	0.00604	0.00127	0.00419
		<b>0.0151</b>	<b>0.00873</b>	<b>0.00529</b>	<b>0.00101</b>	<b>0.00355</b>
	0.42*	0.0179	0.0385	0.00541	0.00184	0.00681
		<b>0.0178</b>	<b>0.00616</b>	<b>0.00611</b>	<b>0.000537</b>	<b>0.00233</b>
	0.62	0.0205	0.0538	0.00616	0.0028	0.0161
		<b>0.0197</b>	<b>0.00345</b>	<b>0.00591</b>	<b>0.000232</b>	<b>0.00178</b>
600	0.258	0.0146	0.0289	0.00176	0.00156	0.00249
		<b>0.0134</b>	<b>0.00829</b>	<b>0.00202</b>	<b>0.000893</b>	<b>0.00134</b>
	0.458*	0.0208	0.0412	0.00254	0.00191	0.0078
		<b>0.0189</b>	<b>0.00568</b>	<b>0.00258</b>	<b>0.000441</b>	<b>0.00113</b>
	0.658	0.0277	0.0487	0.00366	0.00242	0.0182
		<b>0.0285</b>	<b>0.00381</b>	<b>0.00401</b>	<b>0.000242</b>	<b>0.000981</b>
2000	0.206	0.00592	0.013	0.00376	0.0028	0.00212
		<b>0.00592</b>	<b>0.00489</b>	<b>0.00377</b>	<b>0.0025</b>	<b>0.00167</b>
	0.406*	0.00768	0.034	0.00393	0.00174	0.00532
		<b>0.00653</b>	<b>0.00274</b>	<b>0.00407</b>	<b>0.000916</b>	<b>0.00115</b>
	0.606	0.0103	0.0493	0.00428	0.00204	0.0154
		<b>0.00953</b>	<b>0.00194</b>	<b>0.00451</b>	<b>0.000356</b>	<b>0.000865</b>

\* true optimal  $\lambda_0$

Table 5: 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 thresholding value  $a_n = 0.4$  and  $\lambda_0 = 0.07$ .

Modified bootstrap based variance estimates, confidence intervals and hypothesis tests for different predictors					
predictor	$\hat{\beta}_{n,j}$	Variance estimate	90% Confidence Interval		$H_0 : \beta_j = 0$
			lower	upper	
lcavol	5.9223	0.8263	4.1820	7.6625	<b>Reject</b>
lweight	1.6269	0.5162	0.1266	3.1272	<b>Reject</b>
age	0	0.1198	-0.6044	0.6044	Accept
lbph	0.6380	0.2061	0.0001	1.2760	<b>Reject</b>
svi	2.2349	0.6818	0.4894	3.9803	<b>Reject</b>
lcp	0	0.2806	-1.0708	1.0708	Accept
gleason	0	0.0717	-0.4494	0.4494	Accept
pgg45	0.3631	0.0940	-0.1222	0.8486	Accept

Table 6: Analysis of Iowa wheat data from [Draper and Smith \(1998\)](#). Table showing componentwise Lasso estimates, modified bootstrap based variance estimates, 90% confidence intervals and tests for  $H_0 : \beta_j = 0$ .  $n = 33$  observations were used with thresholding value  $a_n = 4.17$  and  $\lambda_0 = 1$ .

Modified bootstrap based variance estimates, confidence intervals and hypothesis tests for different predictors					
predictor	$\hat{\beta}_{n,j}$	Variance estimate	90% Confidence Interval		$H_0 : \beta_j = 0$
			lower	upper	
rain0	1.2717	34.4477	-9.3390	-11.8825	Accept
temp1	-1.2703	29.8067	-10.6683	8.1276	Accept
rain1	-3.0075	32.7032	-12.8731	6.8580	Accept
temp2	0	26.7683	-9.3775	9.3775	Accept
rain2	24.2459	118.1731	4.4609	44.0310	<b>Reject</b>
temp3	-24.4779	120.5945	-44.0817	-4.8781	<b>Reject</b>
rain3	6.6335	39.1086	-2.7599	16.0271	Accept
temp4	0	34.2267	-11.5824	11.5824	Accept