Electronic Journal of Statistics
Vol. 10 (2016) 2124–2156
ISSN: 1935-7524
DOI: 10.1214/16-EJS1160

Thresholding least-squares inference in high-dimensional regression models

Mihai Giurcanu

University of Florida
e-mail: giurcanu@ufl.edu

Abstract: We propose a thresholding least-squares method of inference for high-dimensional regression models when the number of parameters, p, tends to infinity with the sample size, n. Extending the asymptotic behavior of the F-test in high dimensions, we establish the oracle property of the thresholding least-squares estimator when $p = o(n)$. We propose two automatic selection procedures for the thresholding parameter using Scheffé and Bonferroni methods. We show that, under additional regularity conditions, the results continue to hold even if $p = \exp(o(n))$. Lastly, we show that, if properly centered, the residual-bootstrap estimator of the distribution of thresholding least-squares estimator is consistent, while a naive bootstrap estimator is inconsistent. In an intensive simulation study, we assess the finite sample properties of the proposed methods for various sample sizes and model parameters. The analysis of a real world data set illustrates an application of the methods in practice.

MSC 2010 subject classifications: Primary 62J05; secondary 62E20.
Keywords and phrases: Regression models, high-dimensional inference, F-test, thresholding least-squares, residual-bootstrap.

Received January 2016.

Contents

1. Introduction

There is a wide interest in developing statistical and computational methods and theory for high-dimensional regression models when the number of parameters, p, tends to infinity with the sample size, n. In this paper, we propose a thresholding least-squares estimator (TLSE) for high-dimensional linear regression models as a computationally efficient alternative to penalized least-squares. Thresholding inferential methods have been widely used in wavelet nonparametric regression (Donoho and Johnstone, 1994), wavelet nonparametric density estimation (Donoho et al., 1996), and estimation of sparse covariance matrices (Bickel and Levina, 2008; El Karoui, 2008). However, to the best of our knowledge, there is no systematic analysis of thresholding least-squares for high-dimensional linear regression models. The main purpose of this paper is to fill in this gap in the literature.

One of the most popular penalized least-squares estimators for linear regression models is the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) which combines the favorable properties of model selection and ridge regression. Other penalized least-squares estimators are the Bridge estimators (Frank and Friedman, 1993), which include the LASSO as a special case. The asymptotic behavior of Bridge estimators were analyzed by Knight and Fu (2000). The smoothly clipped absolute deviation (SCAD) estimator was proposed by Fan and Li (2001) who also established that the SCAD estimator satisfies the oracle property. An estimator has the oracle property if it is variable selection consistent and the limiting distribution of its subvector corresponding to the non-zero coefficients is the same as if their set were known prior to estimation. Motivated by the fact that the LASSO does not have the oracle property, Zou (2006) proposed the adaptive LASSO (ALASSO) and proved its oracle property. All these methods and theoretical results have been developed under the assumption that p is fixed.

The literature on high-dimensional regression inference dates back to Huber (1973) who showed the asymptotic normality of M-estimators when $p = o(n^{1/2})$, results which were further extended by Portnoy (1984, 1985) for the case when $p\log(n) = o(n^{2/3})$. Asymptotic theory for M-estimators was also developed by Mammen (1989) for the case when $hn^{1/3}(\log(n))^{2/3} \to 0$, where h is the maximum diagonal element of the hat matrix. The consistency of L_2-boosting, which is similar to the forward stagewise least-squares variable selection method, was proven by Bühlmann (2006) when $p = o(\exp(n))$. Asymptotic error rates and power for some multi-stage regression methods were developed by Wasserman and Roeder (2009) for $p = o(\exp(n))$. More recently, van de Geer, Bühlmann and Zhou (2011) compared the ALASSO with the thresholded LASSO in potentially misspecified regression models when $p \geq n$ in terms of prediction error, mean absolute error, mean squared error, and the number of false positive selections. The oracle property of the ALASSO and the Bridge estimators were established by Huang, Ma and Zhang (2008) and Huang, Horowitz and Ma (2008) when $p = o(n)$. Using a (marginal) componentwise estimator as an initial screening of relevant variables, they showed that, under additional regularity conditions, the

results continue to hold even if $p = \exp(o(n))$. The sure-independence screening methodology for ultra-high dimensional feature space was introduced by Fan and Lv (2008). More recently, Wang and Leng (2016) proposed an alternative screening procedure with improved statistical properties and similar computational complexity.

Bootstrap methods (Efron, 1979; Freedman, 1981) are popular computational intensive alternatives to the asymptotic inference which often improve accuracy of inference on small samples (Hall, 1992). The consistency of the residual-bootstrap distribution of the least-squares estimator (LSE) was proved by Bickel and Freedman (1983) when $p = o(n)$. The consistency of residual-bootstrap distributions of M-estimators in general and of the LSE in particular were proved by Mammen (1989, 1993). A parametric bootstrap in conjunction with thresholding inference for a high-dimensional mean with unknown covariance matrix was used by van der Laan and Bryan (2001). More recently, Chatterjee and Lahiri (2011) showed that the residual-bootstrap distribution of the LASSO is inconsistent when the model is sparse, i.e., when some regression coefficients are equal to zero, and that centering the bootstrap distribution at a consistent variable selection estimator provides consistent bootstrap inference. The consistency of the residual-bootstrap distribution of the ALASSO and the oracle property of the residual empirical process for high-dimensional regression models was proved by Chatterjee and Lahiri (2013) and Chatterjee, Gupta and Lahiri (2015).

In this paper, we propose a two-step thresholding least-squares method of inference for high-dimensional regression models. We first show the oracle property of the TLSE when $p = o(n)$ based on an extension of the asymptotic distribution of the F-test for high-dimensional regression models. Similarly to Huang, Ma and Zhang (2008), Huang, Horowitz and Ma (2008), and Fan and Lv (2008), we then show that using a componentwise least-squares estimator as an initial dimension reduction estimator, the resulting TLSE has the oracle property even when $p = \exp(o(n))$. Our theoretical results require that the number of non-zero coefficients, q, be of order $q = o(n)$; this constitutes an advantage of the TLSE compared to multi-stage regression models (Wasserman and Roeder, 2009) which require $q = O(1)$, and the ALASSO and the Bridge estimators (Huang, Ma and Zhang, 2008; Huang, Horowitz and Ma, 2008) which essentially require $q = o(n^{1/2})$. We propose two automatic selection procedures for the thresholding parameter which ensure the oracle property of the TLSE using Scheffé and Bonferroni methods adapted for high-dimensional models. We further show that, when properly centered, the residual-bootstrap distribution of the TLSE is consistent, and when the regression model is sparse, then a naive bootstrap distribution of the TLSE, as a random element in the space of probability distributions on a finite dimensional space, converges *in distribution* to a random normal distribution, and thus, it is inconsistent.

We conclude this section with an outline. In Section 2, we present the large sample properties of the TLSE for both cases: (i) $p < n$ and (ii) $p \geq n$, and present the automatic thresholding parameter selection methods. In Section 3, we study the asymptotic behavior of the bootstrap distribution of TLSE. In

Section 4, we present the results of an empirical study of the finite sample properties of the proposed methods and in Section 5, we analyze a real-world data set to illustrate an application of the methods in practice. The proofs of theoretical results can be found in an Appendix.

2. Thresholding least-squares

Consider the linear regression model:

$$Y_i = X_i^T \beta + \epsilon_i \,, \quad i = 1, \ldots, n \,, \tag{2.1}$$

where $Y_i \in \mathbb{R}$ is the response and $X_i = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$ is the (non-random) explanatory variable corresponding to the ith subject, $\beta \in \mathbb{R}^p$ is the (unknown) regression parameter vector, and $\epsilon_i \in \mathbb{R}$ is the (unobserved) error, with $\epsilon_1, \ldots, \epsilon_n \sim$ i.i.d. P, P is a distribution on $\mathbb{R}$, with $\mathrm{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \sigma^2$, and $\epsilon \sim P$. For identifiability reasons, we assume throughout the paper that q, the number of the non-zero components of β, is smaller than the sample size, i.e., $q < n$. We are interested in statistical inference for β in the case when its dimension, p, increases with the sample size, n, and β is sparse, i.e., when some of its components are zero. For notational convenience, we suppress the dependence on n of p, q, X_i, and Y_i.

Without loss of generality, by centering the response and standardizing the covariates, we assume that the intercept term has been removed from the set of predictors. Thus, $\bar{Y} = 0$, $\bar{X}^{(j)} = 0$, and $\bar{S}^{(j)} = 1$, where

$$\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i \,, \quad \bar{X}^{(j)} = n^{-1} \sum_{i=1}^{n} X_{ij} \,, \quad \bar{S}^{(j)} = n^{-1} \sum_{i=1}^{n} X_{ij}^2 \,, \quad j = 1, \ldots, p \,.$$

Let $I = \{1, \ldots, p\}$. For $b = (b_1, \ldots, b_p)^T \in \mathbb{R}^p$, let $K_b = \{j \in I : b_j = 0\}$, $J_b = \{j \in I : b_j \neq 0\}$, $q_J = \mathrm{card}(J)$ for $J \subset I$ and $\mathrm{card}(J)$ denotes the number of elements of J, $\Theta_1 = \{b \in \mathbb{R}^p : K_b \neq \varnothing\}$, and $\Theta_2 = \mathbb{R}^p \setminus \Theta_1$, where $\varnothing$ is the empty set. The regression model (2.1) is called *sparse* if $\beta \in \Theta_1$, i.e., when some of the regression coefficients are 0.

2.1. Case $p < n$

In this section, we consider the case when $p < n$. Let $\bar{\beta}$ be the least-squares estimator (LSE) of β, i.e.,

$$\bar{\beta} = (X^T X)^{-1} X^T Y = n^{-1} \Omega^{-1} \sum_{i=1}^{n} X_i Y_i = \sum_{i=1}^{n} c_i Y_i \,,$$

where $X = (X_{ij}) \in \mathbb{R}^{n \times p}$ is the design matrix, $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ is the response vector, $c_i = (X^T X)^{-1} X_i \in \mathbb{R}^p$, and

$$\Omega = n^{-1} X^T X = n^{-1} \sum_{i=1}^{n} X_i X_i^T \in \mathbb{R}^{p \times p} \,.$$

Let $\bar{\sigma}^2$ be the (unbiased) LSE of σ^2 given by

$$\bar{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - X_i^T\bar{\beta}\right)^2.$$

Let ρ_1 and ρ_2 denote the minimum and the maximum eigenvalues of Ω, respectively, and let $\|a\|$ denote the Euclidean norm of $a \in \mathbb{R}^p$. We assume the following regularity conditions:

A.1 $\mathrm{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \sigma^2$, and $\mathrm{E}(\epsilon^4) < \infty$, where $\epsilon \sim P$;
A.2 $\rho_1 > 0$, $\max_{1\le i\le n}\|X_i\|^2 = O(p)$, and $p/(n\rho_1) = o(1)$.

Note that since $\mathrm{tr}(\Omega) = p$, where $\mathrm{tr}(\Omega)$ denotes the trace of Ω, then $\rho_1 \le 1$. By condition A.2, it thus follows that $p = o(n)$. Lemma 2.1 shows the consistency of $\bar{\beta}$ and $\bar{\sigma}^2$, that the rate of convergence of $a^T(\bar{\beta} - \beta)$ depends on ρ_1 (which is allowed to tend to 0 but at a slower rate than p/n), and that the rate of convergence of $\bar{\sigma}^2$ to σ^2 is $n^{-1/2}$, exactly the same as in the case of fixed p.

Lemma 2.1. *Suppose that conditions A.1–A.2 hold and let $a \in \mathbb{R}^p$ with $\|a\| = 1$. Then (i) $a^T(\bar{\beta} - \beta) = O_P((\rho_1 n)^{-1/2})$; and (ii) $\bar{\sigma}^2 = \sigma^2 + O_P(n^{-1/2})$.*

Let $\xrightarrow{\mathrm{d}}$ denote convergence in distribution. Lemma 2.2 shows that $a^T\bar{\beta}$ is asymptotically normal for all $a \in \mathbb{R}^p$, with $\|a\| = 1$.

Lemma 2.2. *Suppose that conditions A.1–A.2 hold and let $a \in \mathbb{R}^p$ with $\|a\| = 1$. Then $\bar{s}^{-1/2}a^T(\bar{\beta} - \beta) \xrightarrow{d} N(0,1)$, where $\bar{s} = n^{-1}\bar{\sigma}^2 a^T\Omega^{-1}a$.*

For $J \subseteq I$, let $\beta_J = (\beta_j : j \in J)^T \in \mathbb{R}^{q_J}$, $X_J = (X_{ij} : 1 \le i \le n, j \in J) \in \mathbb{R}^{n\times q_J}$, $\Omega_{JK} = n^{-1}X_J^T X_K \in \mathbb{R}^{q_J\times q_K}$, and let

$$\Sigma_{KK} = \Omega_{KK} - \Omega_{KJ}\Omega_{JJ}^{-1}\Omega_{JK} \in \mathbb{R}^{q_K\times q_K} \tag{2.2}$$

be the Schur complement of the block matrix Ω_{JJ}, where $K = I \setminus J$. An immediate consequence of Lemma 2.2 is that for every fixed $K \subset K_\beta$ (i.e., K does not depend on n), we have

$$n^{1/2}\bar{\sigma}^{-1}\Sigma_{KK}^{1/2}(\bar{\beta}_K - \beta_K) \xrightarrow{\mathrm{d}} N(0, I_{q_K}),$$

where I_{q_K} is the identity matrix in $\mathbb{R}^{q_K\times q_K}$. Note that the F-test statistic for testing the null hypothesis $H_0 : K_\beta = K$ against $H_a : K_\beta \ne K$, where $K \subset I$ is fixed, is given by

$$\hat{F}(K) = n\frac{\bar{\beta}_K^T\Sigma_{KK}\bar{\beta}_K}{q_K\bar{\sigma}^2}.$$

By Lemma 2.2, it follows that under $H_0 : K_\beta = K$, then

$$q_K\hat{F}(K) \xrightarrow{\mathrm{d}} \chi^2_{q_K},$$

where $\chi^2_{q_K}$ is the chi-squared distribution with q_K degrees of freedom. Lemma 2.3 shows the limiting null distribution of the scaled and centered F-test statistic when the cardinality of K increases with n (and thus, q_K increases with n).

Lemma 2.3. *Suppose conditions A.1–A.2 hold and* $n\rho_1^2 \to \infty$. *Then, under* $H_0 : K_\beta = K$,

$$\frac{q_K \hat{F}(K) - q_K}{(2q_K)^{1/2}} \xrightarrow{d} N(0,1)\,. \tag{2.3}$$

Let $\hat{K}$ be a thresholding estimator of the index set of the zero components of β, K_β, given by:

$$\hat{K} = \left\{ j \in I : |\bar{\beta}_j| \le \gamma \bar{\sigma}_{jj} \right\},$$

where $\bar{\beta} = (\bar{\beta}_j : j \in I)^T \in \mathbb{R}^p$, γ is the *thresholding parameter*, $\bar{\sigma}_{jj} = n^{-1/2}\bar{\sigma}\omega_{jj}^{1/2}$, and $\Omega^{-1} = (\omega_{ij}) \in \mathbb{R}^{p\times p}$. We assume that γ satisfies the following conditions:

$$\frac{\rho_1 \rho_2^{-1}\gamma^2 - q_{K_\beta}}{q_{K_\beta}^{1/2}} \to \infty \tag{2.4a}$$

and

$$\frac{\rho_1 \rho_2^{-1}\left(n^{1/2}\sigma^{-1}\rho_1^{1/2}\min_{j\in J_\beta}|\beta_j| - \gamma\right)^2 - q}{q^{1/2}} \to \infty\,. \tag{2.4b}$$

To get an intuition about conditions (2.4a) and (2.4b), suppose for the moment that $\liminf_n \rho_1 > 0$, $\liminf_n \rho_2^{-1} > 0$, $\liminf_n \min_{j\in J_\beta}|\beta_j| > 0$, and $q_{K_\beta} \sim n^\tau$, where $0 < \tau < 1$; here and elsewhere, we use the standard notation that $a_n \sim b_n$ if and only if $a_n = O(b_n)$ and $b_n = O(a_n)$. In this case, we can choose $\gamma \sim n^{\tau_0/2}$, where $\tau < \tau_0 < 1$. Our default choice for γ is $\gamma = (p\log(p))^{1/2}$, and as long as $p = o(n/\log(n))$, then conditions (2.4a) and (2.4b) hold. In Section 2.2, we will take up this problem in more detail and present two automatic thresholding parameter selection procedures.

The TLSE of β is defined as follows:

$$\hat{\beta} = \operatorname*{argmin}_{b\in\mathbb{R}^p} \left\{ (Y - Xb)^T (Y - Xb) : b_{\hat{K}} = 0 \right\}.$$

Note that $\hat{\beta}_{\hat{K}} = 0$ and $\hat{\beta}_{\hat{J}} = (X_{\hat{J}}^T X_{\hat{J}})^{-1} X_{\hat{J}}^T Y$, where $\hat{J} = I \setminus \hat{K}$. Let further

$$\hat{\sigma}^2 = \frac{1}{n - q_{\hat{J}}} \sum_{i=1}^n \left(Y_i - X_i^T \hat{\beta}\right)^2$$

be the TLSE of σ^2. Similarly to Lemma 2.1, it readily follows that $\hat{\sigma}^2 = \sigma^2 + O_P(n^{-1/2})$. Theorem 2.1 shows that $\hat{\beta}$ has the oracle property, and thus, the TLSE has similar asymptotic properties as the LASSO-type estimators that have the oracle property.

Theorem 2.1. *Suppose that conditions A.1–A.2,* (2.4a) *and* (2.4b) *hold, and that* $n\rho_1^2 \to \infty$. *Then* $\hat{\beta}$ *has the oracle property, i.e., (i)* $\Pr(\hat{K} = K_\beta) \to 1$; *and (ii)* $\hat{s}^{-1/2} a^T(\hat{\beta} - \beta) \xrightarrow{d} N(0,1)$, *where* $a \in \mathbb{R}^p$, $\|a\| = 1$, *and* $\hat{s} = n^{-1}\hat{\sigma}^2 a_{\hat{J}}^T \Omega_{\hat{J}\hat{J}}^{-1} a_{\hat{J}}$.

Note that the regularity conditions of Theorem 2.1 are less restrictive than those imposed for Bridge estimators by Huang, Horowitz and Ma (2008). To this end, assume that $\liminf_n \rho_1 > 0$, $\liminf_n \rho_2^{-1} > 0$, and $\liminf_n \min_{j\in J_\beta}|\beta_j| > 0$. Then, by condition (A3)(b) of Huang, Horowitz and Ma (2008), we obtain $\lambda(p/n)^{\gamma/2}p^{-1} \to \infty$, where λ is the regularization parameter and $\gamma \in (0,1)$ is the power of the penalty component of the Bridge estimator. Hence $\lambda/p \to \infty$. By their condition (A2)(b), then $\lambda q/n \to 0$. However, this is a restrictive condition, since, for example, if $p \sim n/\log(n)$, then $q = o(\log(n))$; however, in this case, our regularity conditions require only that $p = o(n/\log(n))$. See the paragraph above for more details. Moreover, if the covariates are uniformly bounded, then condition (A5)(b) holds only if $q = o(n^{1/2})$ while our corresponding regularity condition $\max_{1\le i\le p}\|X_i\|^2 = O(p)$ holds without additional restrictions.

By Lemma 2.2 and Theorem 2.1, when $\beta \in \Theta_2$, the asymptotic distributions of $a^T(\bar\beta-\beta)$ and $a^T(\hat\beta-\beta)$ are the same for all $a \in \mathbb{R}^p$, with $\|a\| = 1$. Corollary 2.1 is a direct consequence of Theorem 2.1 and shows that $\hat\beta$ is more efficient than $\bar\beta$ in the sense that the asymptotic variance of $a^T(\hat\beta-\beta)$ is smaller than or equal to the asymptotic variance of $a^T(\bar\beta-\beta)$.

Corollary 2.1. *Suppose conditions A.1–A.2, (2.4a) and (2.4b) hold, and that* $n\rho_1^2 \to \infty$. *Then* $\hat\beta$ *is asymptotically at least as efficient as* $\bar\beta$.

2.2. Thresholding parameter selection

Theoretically, any sequence γ satisfying (2.4a) and (2.4b) will ensure the oracle property of $\hat\beta$. In this section, we describe two automatic selection procedures for γ with good numerical and statistical properties. These procedures are based on extensions of Scheffé and Bonferroni methods adapted for high dimensional regression models.

Method I The first method is based on Scheffé procedure (see, e.g., Khuri, 2010, Section 6.6.1.1). We assume that conditions of Lemma 2.3 hold. Hence,

$$\frac{n(\bar\beta-\beta)^T\Omega(\bar\beta-\beta)/\bar\sigma^2 - p}{(2p)^{1/2}} \xrightarrow{\mathrm{d}} N(0,1)\,.$$

Let $\alpha > 0$ be a sequence of nominal levels such that $\alpha = o(1)$ and $\log(\alpha) = o(n^{1/2}/p^{1/4})$. Using an approximation of the upper tail probabilities of $N(0,1)$ (see, e.g., Zelen and Severo, 1972, Example 26.2.12, p. 932), then $\xi_\alpha \to \infty$ and $\xi_\alpha = o(n/p^{1/2})$, where ξ_α is the upper α-quantile of $N(0,1)$. Since $\xi_\alpha \to \infty$, then

$$\Pr\Big(n(\bar\beta-\beta)^T\Omega(\bar\beta-\beta) \le p\bar\sigma^2 + (2p)^{1/2}\bar\sigma^2\xi_\alpha\Big) \to 1\,.$$

Let $\zeta = \Omega^{1/2}(\bar\beta-\beta)$ and $c^2 = p\bar\sigma^2/n + (2p)^{1/2}\bar\sigma^2\xi_\alpha/n$. It is known that (see, e.g., Khuri, 2010, Lemma 6.1),

$$\zeta^T\zeta \le c^2 \iff \big|b^T\zeta\big| \le c(b^Tb)^{1/2} \text{ for all } b \in \mathbb{R}^p\,.$$

Hence,

$$\Pr\Big(\big|b^T\Omega^{1/2}(\bar\beta-\beta)\big|\le (b^Tb)^{1/2}\big(p\bar\sigma^2/n+(2p)^{1/2}\bar\sigma^2\xi_\alpha/n\big)^{1/2}\text{ for all } b\in\mathbb{R}^p\Big)\to 1\,.$$

Letting $l=\Omega^{1/2}b$ and substituting $b=\Omega^{-1/2}l$ for b in the expression above, we obtain

$$\Pr\Big(\big|l^T(\bar\beta-\beta)\big|\le (l^T\Omega^{-1}l)^{1/2}\big(p\bar\sigma^2/n+(2p)^{1/2}\bar\sigma^2\xi_\alpha/n\big)^{1/2}\text{ for all } l\in\mathbb{R}^p\Big)\to 1\,.$$

Let

$$\hat K_{\mathrm{S}}=\big\{j\in I:|\bar\beta_j|\le\omega_{jj}^{1/2}\big(p\bar\sigma^2/n+(2p)^{1/2}\bar\sigma^2\xi_\alpha/n\big)^{1/2}\big\}\,.$$

Let $\hat\beta(\hat K_{\mathrm{S}})$ be the TLSE of β corresponding to the Scheffé dimension reduction set $\hat K_{\mathrm{S}}$. Since $\xi_\alpha\to\infty$, then $\Pr(K_\beta\subseteq\hat K_{\mathrm{S}})\to 1$ and since $\xi_\alpha=o(n/p^{1/2})$, then $\Pr(J_\beta\cap\hat K_{\mathrm{S}})\to 0$. Hence, $\Pr(\hat K_{\mathrm{S}}=K_\beta)\to 1$. Similarly to the proof of Theorem 2.1, then $\hat\beta(\hat K_{\mathrm{S}})$ has the oracle property.

Method II The second method is based on Bonferroni procedure (see, e.g., Khuri, 2010, Section 7.5.3) for normal models, i.e., under the additional assumption that $\epsilon\sim N(0,\sigma^2)$. Let

$$\hat K_{\mathrm{B}}=\big\{j\in I:|\bar\beta_j|\le\bar\sigma_{jj}t_{n-p;\alpha/p}\big\}\,,$$

where $t_{p;\alpha}$ is the upper α-quantile of the t-distribution with p degrees of freedom, $\alpha=o(1)$, and $\log(\alpha)=o\big((n^{1/2}\rho_1^{1/2}\min_{j\in J_\beta}|\beta_j|)^{1/2}\big)$. Using the same normal tail approximation as above, $t_{n-p;\alpha/p}\to\infty$ and $t_{n-p;\alpha/p}=o\big(n^{1/2}\rho_1^{1/2}\min_{j\in J_\beta}|\beta_j|\big)$. Thus

$$\begin{aligned}\liminf_{n\to\infty}\Pr\big(K_\beta\subset\hat K_{\mathrm{B}}\big)&=\liminf_{n\to\infty}\Pr\big(\cap_{j\in K_\beta}\{|\bar\beta_j|\le\bar\sigma_{jj}t_{n-p;\alpha/p}\}\big)\\&\ge 1-\limsup_n\alpha=1\,.\end{aligned}$$

Similarly to (A.17), by Lemma 2.3, we obtain

$$\begin{aligned}&\limsup_n\Pr\Big(\min_{j\in J_\beta}\frac{|\bar\beta_j|}{\bar\sigma_{jj}}\le t_{n-p;\alpha/p}\Big)\\&\quad\le\limsup_n\Pr\Big(\max_{j\in J_\beta}\frac{|\bar\beta_j-\beta_j|}{\bar\sigma_{jj}}\ge n^{1/2}\bar\sigma^{-1}\rho_1^{1/2}\min_{j\in J_\beta}|\beta_j|-t_{n-p;\alpha/p}\Big)=0\end{aligned}$$

Therefore, $\Pr\big(\hat K_{\mathrm{B}}=K_\beta\big)\to 1$. Let $\hat\beta(\hat K_{\mathrm{B}})$ be the TLSE corresponding to the Bonferroni dimension reduction set $\hat K_{\mathrm{B}}$. Similarly to the proof of Theorem 2.1, then $\hat\beta(\hat K_{\mathrm{B}})$ has the oracle property. In our simulation study and data analysis, we have taken $\alpha=1/(2\log(n))$, and thus, for sample sizes $100\le n\le 1000$, we have $0.07<\alpha<0.10$.

2.3. Case $p \geq n$

In this section, we consider the case when $p \geq n$ and $q < n$, where recall that $q = q_{J_\beta}$ is the number of non-zero components of $\beta \in \mathbb{R}^p$. We will show that in this case, we can make oracle inference about β in two steps using a similar initial variable screening method as Huang, Horowitz and Ma (2008), Huang, Ma and Zhang (2008), and Fan and Lv (2008). First, we select an index set $\hat{J}_0 \subseteq I$, with $\text{card}(\hat{J}_0) < n$, with the property that it contains the indices $i \in I$ for which the absolute values of the t-statistics of the (marginal) componentwise least-squares estimator (CLSE) of β are larger than an appropriately chosen threshold value. In the second step, we perform the thresholding least-squares inference presented in Sections 2.1–2.2 using only the covariates from the index set $\hat{J}_0$. We will show that, under additional regularity conditions, the resulting TLSE has the oracle property even when p increases almost exponentially with respect to n.

Since the columns of the design matrix $X \in \mathbb{R}^{n\times p}$ are standardized and the response vector $Y \in \mathbb{R}^n$ is centered at 0, then the CLSE of β is given by $\tilde{\beta} = (\tilde{\beta}_j : j \in I)^T \in \mathbb{R}^p$, where

$$\tilde{\beta}_j = \frac{\sum_{i=1}^n X_{ij}Y_i}{\sum_{i=1}^n X_{ij}^2} = n^{-1}\sum_{i=1}^n X_{ij}Y_i\,, \quad j \in I\,.$$

Let $\tilde{\Gamma} = \{\tilde{\gamma}_j : j \in I\}$ be the set of the absolute values of the t-statistics corresponding to $\tilde{\beta}$, i.e.,

$$\tilde{\gamma}_j = \frac{|\tilde{\beta}_j|}{\tilde{\sigma}_j}\,, \quad \text{where} \quad \tilde{\sigma}_j^2 = \frac{1}{n(n-1)}\sum_{i=1}^n \big(Y_i - X_{ij}\tilde{\beta}_j\big)^2\,, \quad j \in I\,.$$

Let $\tilde{\gamma}_{(j)}$ denote the jth order statistic of $\tilde{\Gamma}$, $j \in I$. Then $\hat{J}_0$ is defined as the index set corresponding to the largest m absolute values of the t-statistics corresponding to the CLSE of β, where $q \leq m$ and $m = o(n)$ is a pre-specified value corresponding to the hypothesized maximum number of non-zero regression coefficients. Thus,

$$\hat{J}_0 = \big\{j \in I : \tilde{\gamma}_j > \tilde{\gamma}_{(p-m)}\big\}\,.$$

The TLSE is defined in the same way as in Section 2.1 for the response vector Y and design matrix $X_{\hat{J}_0}$. With a slight abuse of notation, let ρ_1 and ρ_2 denote the minimum and the maximum eigenvalues of $\Omega_{\hat{J}_0\hat{J}_0} \in \mathbb{R}^{m\times m}$.

We assume the following regularity conditions.

B.1 $\text{E}(\epsilon) = 0$, $\text{E}(\epsilon^2) = \sigma^2$, where $\epsilon \sim P$, and P has sub-Gaussian tails, that is, there exists constants $c_0, C_0 > 0$ such that

$$\Pr\big(|\epsilon| \geq x\big) \leq C_0\exp\big(-c_0x^2\big) \quad \text{for all } x \geq 0\,;$$

B.2 $p = q^{-1}\exp(o(n))$, $q \leq m$, $m = o(n)$, $\rho_1 > 0$, $m/(n\rho_1) = o(1)$, and $\max_{1\leq i\leq n}\|X_{i,\hat{J}_0}\|^2 = o(m)$, where $X_{i,\hat{J}_0} = (X_{i,j} : j \in \hat{J}_0)^T \in \mathbb{R}^m$;

B.3 (i) The maximum correlation between the covariates in J_β and K_β is of order $o(q^{-1})$, i.e.,

$$\max_{j\in J_\beta, k\in K_\beta} n^{-1}\Big|\sum_{i=1}^{n} X_{ij}X_{ik}\Big| = o(q^{-1})\,;$$

(ii) There exists a constant $c_2 > 0$ such that

$$\liminf_n \min_{j\in J_\beta} |\zeta_j| \geq c_2\,,$$

where $\zeta_j = \mathrm{E}(\tilde{\beta}_j) = n^{-1}\sum_{i=1}^{n} X_i^T \beta X_{ij}$;
(iii) There exists a constant $C_3 > 0$ such that

$$\limsup_n \max_{j\in J_\beta} |\beta_j| \leq C_3\,.$$

Theorem 2.2 shows that, under additional regularity conditions, the TLSE has the oracle property even if p grows almost exponentially with n.

Theorem 2.2. *Assume that conditions B.1–B.3, (2.4a) and (2.4b) hold, and that* $n\rho_1^2 \to \infty$. *Then (i)* $\Pr(J_\beta \subset \hat{J}_0) \to 1$; *and (ii)* $\hat{s}^{-1/2}a^T(\hat{\beta} - \beta) \xrightarrow{d} N(0,1)$, *where* $a \in \mathbb{R}^p$, $\|a\| = 1$, *and* $\hat{s} = n^{-1}\sigma^2 a_{\hat{J}}^T \Omega_{\hat{J}\hat{J}}^{-1} a_{\hat{J}}$.

The regularity conditions of Theorem 2.2 are similar to those imposed for the Bridge estimators by Huang, Horowitz and Ma (2008). However, we can highlight an important difference. Specifically, the *partial orthogonality* condition (B2)(a) of Huang, Horowitz and Ma (2008) requires the correlation coefficients between the covariates corresponding to the zero and the non-zero coefficients be of order $O(n^{-1/2})$, while our corresponding condition B.3(i) requires to be only of order $o(q^{-1})$. Under their condition (B3)(a), $q = o(n^{1/2})$, and thus, our condition is less restrictive. Note that van de Geer, Bühlmann and Zhou (2011) developed regularity conditions for an analytical comparison between ALASSO and LASSO with thresholding in terms of prediction error, mean absolute error, mean squared error, and the number of false positive selections. Since van de Geer, Bühlmann and Zhou (2011) did not prove the oracle property of the ALASSO and the LASSO with thresholding, we cannot compare our regularity conditions with theirs.

We can relax the condition of sub-Gaussian tails of the errors in condition B.1 on the expense of a slower growth of p. Specifically, assuming only finite fourth order moments of ϵ given by condition A.1, by Markov's inequality, we obtain

$$\Pr\Big(n^{-1}\Big|\sum_{i=1}^{n} X_{ij}\epsilon_i\Big| \geq c_2\Big) = O(n^{-2})\,.$$

Analysis of the proof of Theorem 3.1 shows that part (i) and (ii) of the theorem hold provided that $pq = o(n^2)$. Note further that we can also provide more primitive conditions for B.3(ii). Specifically, we could request instead that

$$\liminf_n \min_{j\in J_\beta} |\beta_j| \geq c_2 \quad \text{and} \quad \max_{j\neq j'\in J_\beta} n^{-1}\Big|\sum_{i=1}^{n} X_{ij}X_{ij'}\Big| = o(q^{-1})\,.$$

To this end, note that for $j \in J_\beta$, we have

$$|\zeta_j| = \Big|n^{-1}\sum_{i=1}^n X_{ij}X_i^T\beta\Big| = \Big|n^{-1}\sum_{i=1}^n \sum_{j'\in J_\beta} X_{ij}X_{ij'}\beta_j\Big|$$
$$\geq c_2 - q\Big\{\max_{j\in J_\beta}|\beta_j|\Big\} \max_{j\neq j'\in J_\beta} n^{-1}\Big|\sum_{i=1}^n X_{ij}X_{ij'}\Big|,$$

and thus, B.3(ii) holds.

In practice, we can set $m = \lfloor n/\log(n)\rfloor$, where $\lfloor n/\log(n)\rfloor$ is the integer part of $n/\log(n)$. However, we can also select m via a k-fold cross-validation procedure. Specifically, for $i = 1,\ldots,k$, let T_i and V_i denote the index sets corresponding to the ith training and validation data sets, respectively, where $T_i, V_i \subset I$, with $\text{card}(V_i) = n_k$ and $T_i = I \setminus V_i$. Let $L \subset \{1,\ldots,n\}$ be a search grid for m. For $l \in L$, let $\hat{v}_i(l)$ denote the cross-validation estimate of the mean squared prediction error of the submodel corresponding to $m = l$ using the ith validation set, i.e.,

$$\hat{v}_i(l) = n_k^{-1}\sum_{j\in V_i}\big(Y_j - \hat{Y}_{j,l}^{T_i}\big)^2,$$

where $\hat{Y}_{j,l}^{T_i}$ is the predicted value of Y_j using the training data set T_i and the design matrix $X_{T_i,J_l} = (X_{t,j} : t \in T_i, j \in J_l)$ and $J_l = \{j \in I : \tilde{\gamma}_j > \tilde{\gamma}_{(p-l)}\}$. Let $\hat{v}(l) = k^{-1}\sum_{i=1}^k \hat{v}_i(l)$, and set $m = \hat{m}$, where $\hat{m} = \text{argmin}_{l\in L}\,\hat{v}(l)$. In the simulation study and data analysis, we search l on a grid L of 20 equally spaced values on the log-scale (so that the grid is finer for smaller values of l), and we use $k = 20$.

3. Bootstrap inference

Let $\hat{\beta}$ be the bootstrap version of β in the context of thresholding least-squares inference. Let $\hat{\epsilon}_i = Y_i - X_i^T\hat{\beta}$ be the ith (raw) residual; since $\bar{X} = 0$ and $\bar{Y} = 0$, then $\sum_{j=1}^n \hat{\epsilon}_j = 0$, and thus, the residuals are inherently "centered" at zero. Let $\hat{E}_{1:n} = \{\hat{\epsilon}_1,\ldots,\hat{\epsilon}_n\}$ be the sample of residuals and let $\mathbb{P} = n^{-1}\sum_{i=1}^n \delta_{\hat{\epsilon}_i}$ denote its empirical distribution, where δ_x denotes the unit mass at $x \in \mathbb{R}$. The residual-bootstrap method (Freedman, 1981) is performed by first sampling, with replacement, the residuals which are then added to the fitted values to obtain a bootstrap sample. Specifically, given $\hat{E}_{1:n}$, let $\hat{E}_{1:n}^* = \{\hat{\epsilon}_1^*,\ldots,\hat{\epsilon}_n^*\}$ be a conditionally i.i.d. sample from $\mathbb{P}$, i.e., $\hat{E}_{1:n}^*$ is a with replacement random sample of size n from $\hat{E}_{1:n}$. Then, for each $i = 1,\ldots,n$, let $Y_i^* = X_i^T\hat{\beta} + \hat{\epsilon}_i^*$, and let $Y^* = (Y_1^*,\ldots,Y_n^*)^T \in \mathbb{R}^n$. The bootstrap version of $\hat{\beta}$ is

$$\hat{\beta}^* = \underset{b\in\mathbb{R}^p}{\text{argmin}}\big\{(Y^* - Xb)^T(Y^* - Xb) : b_{\hat{K}} = 0\big\}.$$

Similarly to its sample version, the bootstrap version of $\hat{\beta}$ has a closed form solution given by $\hat{\beta}^*_{\hat{K}} = 0$ and $\hat{\beta}^*_{\hat{J}} = (X^T_{\hat{J}} X_{\hat{J}})^{-1} X_{\hat{J}} Y^*$. Let

$$\hat{\sigma}^{*2} = \frac{1}{n - q_{\hat{J}}} \sum_{i=1}^{n} (Y_i^* - X_i^T \hat{\beta})^2 \quad \text{and} \quad \hat{s}^* = n^{-1} \hat{\sigma}^{*2} a^T_{\hat{J}} \Omega^{-1}_{\hat{J}\hat{J}} a_{\hat{J}}$$

be the bootstrap versions of $\hat{\sigma}^2$ and $\hat{s}$, respectively.

The bootstrap estimator of $\mathscr{L}(\hat{s}^{-1/2} a^T (\hat{\beta} - \beta))$ is the conditional distribution given Y of $\hat{s}^{*-1/2} a^T (\hat{\beta}^* - \hat{\beta})$, which we denote as $\mathscr{L}(\hat{s}^{*-1/2} a^T (\hat{\beta}^* - \hat{\beta}) | Y)$. Another option is to use $\bar{\beta}$ as a centering value, and in this case, $\mathscr{L}(\hat{s}^{*-1/2} a^T (\hat{\beta}^* - \bar{\beta}) | Y)$ is the "naive" bootstrap distribution estimator of $\mathscr{L}(\hat{s}^{-1/2} a^T (\hat{\beta} - \beta))$. Since it is determined by Y, the bootstrap distribution $\mathscr{L}(\hat{s}^{*-1/2} a^T (\hat{\beta}^* - \hat{\beta}) | Y)$ is a random element in $\mathscr{P}$, the space of distributions on $\mathbb{R}$. We equip $\mathscr{P}$ with the Prohorov metric, which metrizes the weak convergence, and with the corresponding Borel sigma-field generated by the topology of weak convergence (see, e.g., Dudley, 2002, p. 393–399). Theorem 3.1 shows that, under the regularity conditions of Theorem 2.1, the bootstrap distribution $\mathscr{L}(\hat{s}^{*-1/2} a^T (\hat{\beta}^* - \hat{\beta}) | Y)$ is consistent, and if $\beta \in \Theta_1$, the "naive" version $\mathscr{L}(\hat{s}^{*-1/2} a^T (\hat{\beta}^* - \bar{\beta}) | Y)$ converges *in distribution* to a random distribution, and thus is inconsistent. This result is obtained for the case when $p < n$. Careful analysis of the proof of Theorem 3.1 shows that a similar result continue to hold also in the case when $p \geq n$ under the regularity conditions of Theorem 2.2.

Theorem 3.1. *Suppose that conditions A.1–A.2, (2.4a) and (2.4b) hold, and that* $n\rho_1^2 \to \infty$. *Then (i)* $\mathscr{L}(\hat{s}^{*-1/2} a^T (\hat{\beta}^* - \hat{\beta}) | Y) \xrightarrow{\Pr} N(0,1)$ *for* $\beta \in \Theta$. *Suppose further that* $\limsup_n \rho_1^{-1} \rho_2 < \infty$; *then (ii)* $\mathscr{L}(\hat{s}^{*-1/2} a^T (\hat{\beta}^* - \bar{\beta}) | Y) \xrightarrow{d} N(Z,1)$ *for* $\beta \in \Theta_1$, *where* $Z \sim N(0, \sigma_\eta^2)$ *and* σ_η^2 *is defined in the proof.*

Another residual-bootstrap method is based on resampling the scaled residuals $\hat{r}_i$, where $\hat{r}_i = r_i - \bar{r}$, $r_i = (1 - h_{ii})^{-1/2} \hat{\epsilon}_i$, $\bar{r} = n^{-1} \sum r_i$, and $h_{ii} = X_i^T (X^T X)^{-1} X_i$ is the ith element of the "hat matrix" (Davison and Hinkley, 1997, p. 259). By the regularity conditions of Theorem 3.1, $\max_{1 \leq i \leq n} h_{ii} = O(p/(n\rho_1)) = o(1)$, and thus, $n^{-1} \sum_{i=1}^n r_i^2 \xrightarrow{\Pr} \sigma^2$. Careful analysis of the proof of Theorem 3.1 shows that this version of residual-bootstrap is also consistent.

4. Empirical results

4.1. Simulation models

In this section, we present the results of a simulation study of the finite sample behavior of the thresholding least-squares inference in sparse low and high-dimensional regression models. We assess our results in terms of computational and numerical efficiency of the ordinary least-squares, LASSO, and thresholding least-squares inferential methods. The computations are done in the R language and we use the glmnet package (Friedman, Hastie and Tibshirani,

2010; El Ghaoui, Viallon and Rabbani, 2012) to compute the LASSO estimator (with the regularization parameter selected via the 10-fold cross-validation procedure implemented in the package). We have implemented the thresholding least-squares methods in an R package, called TLSE, which is available from the author upon request. Simulations were performed on the HiPerGator cluster hosted by the High Performance Computing Center at University of Florida.

Our simulation models are similar to those considered by Huang, Horowitz and Ma (2008) for the Bridge estimators and we assess the finite sample performance of the methods in terms of (i) variable selection, (ii) prediction accuracy, and (iii) estimation efficiency. The variable selection performance is measured by the relative frequency of correct identification of the set of zero and non-zero regression parameters. Specifically, for $j \in K_\beta$, the relative frequency of correct identification is

$$\hat{p}_j = S^{-1} \sum_{s=1}^{S} I(\hat{\beta}_j^s = 0)\,,$$

where $\hat{\beta}^s = (\hat{\beta}_1^s, \ldots, \hat{\beta}_p^s)^T \in \mathbb{R}^p$ is the TLSE of β on the sth simulated sample, $s = 1, \ldots, S$, and S is the total number of simulated samples. For $j \in J_\beta$, the relative frequency of correct identification is

$$\hat{p}_j = S^{-1} \sum_{s=1}^{S} I(\hat{\beta}_j^s \neq 0)\,.$$

The prediction performance is measured by the (empirical) root mean squared prediction error, which is calculated as:

$$\text{RMSPE} = \Big\{ S^{-1} \sum_{s=1}^{S} n^{-1} \sum_{i=1}^{n} \big(Y_i^s - \hat{Y}_i^s\big)^2 \Big\}^{1/2},$$

where Y_i^s is the ith response of the sth simulated sample, and $\hat{Y}_i^s$ is the predicted response for the ith observation in the sth simulated sample using the parameter estimates obtained on the sth independent simulated sample of size n from the regression model. The estimation efficiency is measured by the empirical root mean squared error (RMSE) of the parameter estimates, which is calculated as:

$$\text{RMSE}(\hat{\beta}_j) = \Big\{ S^{-1} \sum_{s=1}^{S} \big(\hat{\beta}_j^s - \beta_j\big)^2 \Big\}^{1/2}, \quad j = 1, \ldots, p\,,$$

and the root average mean squared error (RAMSE), which is calculated as:

$$\text{RAMSE}(\hat{\beta}) = \Big\{ p^{-1} \sum_{j=1}^{p} S^{-1} \sum_{s=1}^{S} \big(\hat{\beta}_j^s - \beta_j\big)^2 \Big\}^{1/2}.$$

The samples are generated from the regression model (2.1), the design matrix $X \in \mathbb{R}^{n \times p}$ is generated once and then kept fixed across simulations, and the

errors are generated from the normal distribution $N(0, \sigma^2)$. We have compared the runtime of TLSE and LASSO on a data set generated from the model 6, with $n = 20000$, on a Intel i5 ultrabook, 3.6 GHz, 64 bit, 8GB RAM, running under Ubuntu 14.04. The runtime for the LASSO (we used the cv.glmnet function implemented in the glmnet package) was 12.7 sec, for the LSE (we used the lm function) was 6.2 sec, and for the TLSE was 6.6 sec. We anticipate a significant runtime improvement of the TLSE compared to the LASSO for ultra-high dimensional data sets.

We generated the samples from the following six regression models.

Model 1. We set $p = 30$, $\sigma = 0.5$, $X_i \sim$ i.i.d. $N(0, \Sigma)$, $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{30\times 30}$ with $\sigma_{ij} = r^{|i-j|}$ and $r = 0.5$, $\beta_j = 0$ for $1 \le j \le 15$, $\beta_j = 0.5$ for $16 \le j \le 20$, $\beta_j = 1.5$ for $21 \le j \le 25$, $\beta_j = 2.5$ for $26 \le j \le 30$.

Model 2. The same as Model 1 with $r = 0.9$.

Model 3. We set $p = 30$, $\sigma = 0.5$, $X_{ij} \sim$ i.i.d. $N(0,1)$, $1 \le j \le 15$, $X_{ij} = Z_{i1} + \eta_{ij}$, $Z_{i1} \sim$ i.i.d. $N(0,1)$, $\eta_{ij} \sim$ i.i.d. $N(0, 0.25)$, $16 \le j \le 20$, $X_{ij} = Z_{i2} + \eta_{ij}$, $Z_{i2} \sim$ i.i.d. $N(0,1)$, $\eta_{ij} \sim$ i.i.d. $N(0, 0.25)$, $21 \le j \le 25$, $X_{ij} = Z_{i3} + \eta_{ij}$, $Z_{i3} \sim$ i.i.d. $N(0,1)$, $\eta_{ij} \sim$ i.i.d. $N(0, 0.25)$, $26 \le j \le 30$, $i = 1, \ldots, n$, $\beta_j = 0$ for $1 \le j \le 15$ and $\beta_j = 1.5$ for $16 \le j \le 30$.

Model 4. We set $p = 200$, $\sigma = 0.5$, $X_{i,1:185} \sim$ i.i.d. $N(0, \Sigma_1)$, with $\Sigma_1 = (\sigma_{1,ij}) \in \mathbb{R}^{185\times 185}$ and $\sigma_{1,ij} = r^{|i-j|}$, $r = 0.5$, $X_{i,186:200} \sim$ i.i.d. $N(0, \Sigma_2)$, $\Sigma_2 = (\sigma_{2,ij}) \in \mathbb{R}^{15\times 15}$, $\sigma_{2,ij} = r^{|i-j|}$, $r = 0.5$, $X_{i,1:185}$ and $X_{i,186:200}$ are independent, $i = 1, \ldots, n$, $\beta_j = 0$ for $1 \le j \le 185$, $\beta_j = 0.5$ for $186 \le j \le 190$, $\beta_j = 1.5$ for $191 \le j \le 195$, and $\beta_j = 2.5$ for $196 \le j \le 200$.

Model 5. The same as Model 4 with $r = 0.9$.

Model 6. We set $p = 500$, $\sigma = 0.5$, $X_{i,16:500} \sim$ i.i.d. $N(0, I_{485})$, $X_{i,486:500}$ are generated in the same way as in Model 5, $X_{i,1:485}$ and $X_{i,486:500}$ are independent, $i = 1, \ldots, n$, $\beta_j = 0$ for $1 \le j \le 485$, and $\beta_j = 1.5$ for $486 \le j \le 500$.

4.2. Simulation results

In this section, we present the results of the simulation study. The number of simulations is $S = 5000$ and the sample sizes are $n = 100, 200, 400, 800$. Figures 1–2 show the empirical frequency of correct identification of the zero and non-zero regression parameters for the LASSO and the TLSE, respectively. Note that the empirical frequency of correct identification for models 3 and 6 is about 1 for all regression parameters and sample sizes for both the LASSO and the TLSE. The empirical frequency of correct identification for model 1 is about 1 for the TLSE for all parameters and sample sizes and smaller than 1 for the LASSO in the case of the zero regression parameters (with values of about 0.8 for $n = 400$). For model 5, the empirical frequency of correct identification is about 1 for the zero regression parameters and all sample sizes for both the LASSO and the TLSE; however, even though the empirical frequency is below 1 for smaller coefficients and smaller sample sizes, the TLSE

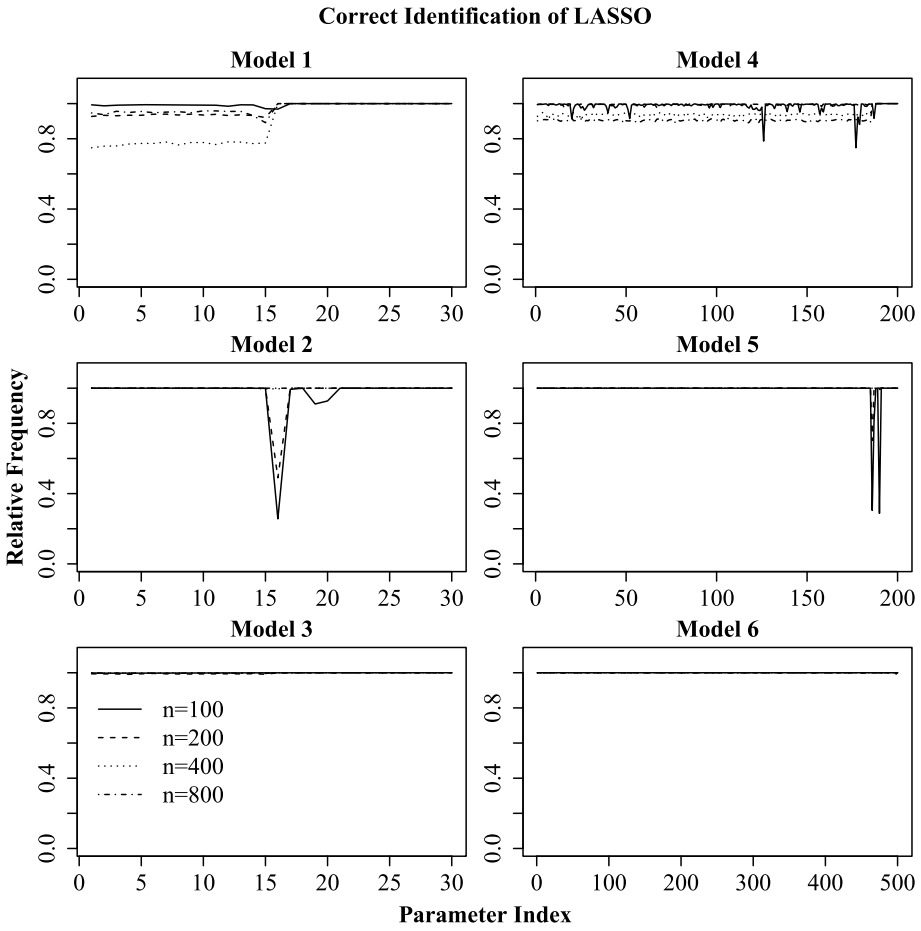

FIG 1. *Empirical frequency of correct identification of the zero and the non-zero regression parameters for models 1–6 using the LASSO estimator. The number of simulations is* $S = 5000$ *and the sample sizes are* $n = 100, 200, 400, 800$.

outperforms the LASSO for all cases. For models 2 and 4, we have mixed results. Specifically, for model 2, while the LASSO outperforms the TLSE for smaller non-zero regression coefficients for $n = 100$, the TLSE outperforms the LASSO for all other sample sizes for both the zero and non-zero parameters. For model 4, while the LASSO outperforms the TLSE for small non-zero parameters for $n = 100$, the TLSE has higher frequency of correct identification for both the zero and non-zero regression parameters for all other sample sizes.

Figures 3–5 show the root mean squared errors (RMSE) of the LSE, LASSO, and TLSE, respectively. For models 4–6, the LSE is calculated using the first $m = \lfloor n/5 \rfloor$ variables in the model. Note that for models 1 and 2, the RMSE

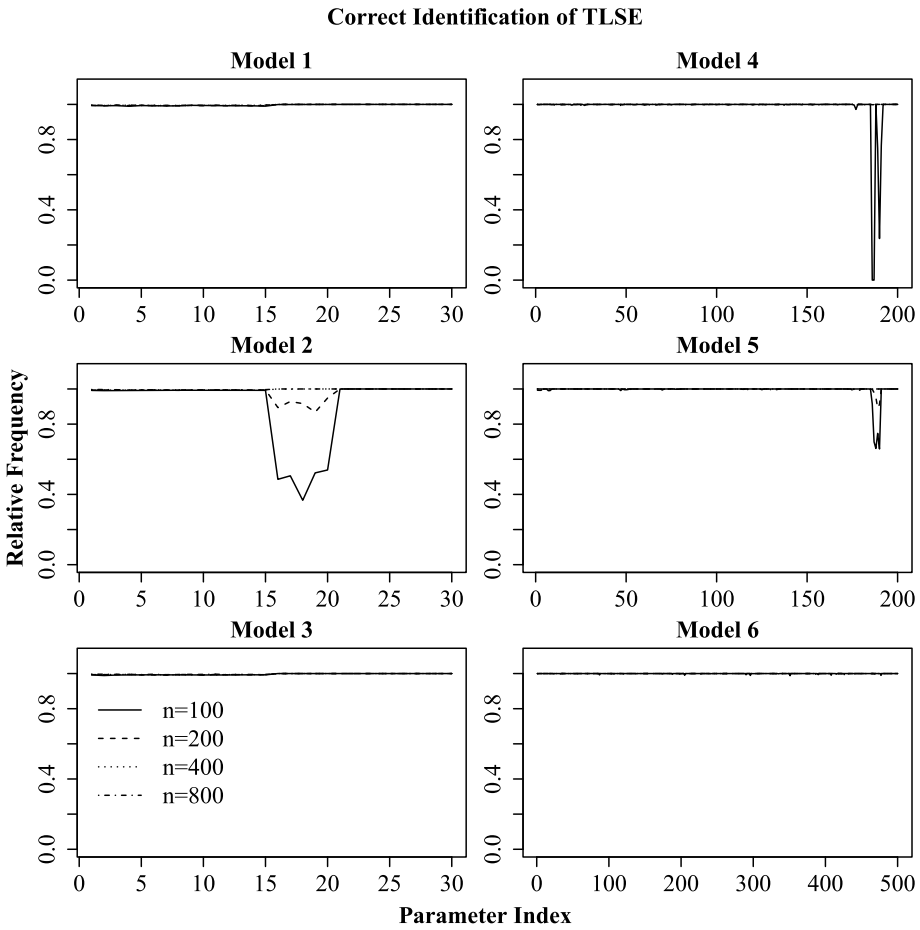

FIG 2. *Empirical frequency of correct identification of the zero and the non-zero regression parameters for models 1–6 using the TLSE. The number of simulations is* $S = 5000$ *and the sample sizes are* $n = 100, 200, 400, 800$.

of LSE are larger for the zero regression parameters and smaller for the non-zero regression parameters than of the LASSO and the TLSE; the RMSE of the TLSE are generally smaller than of the LASSO for all parameters and sample sizes, with one exception, for $n = 100$, where the RMSE of TLSE are larger than of the LASSO for smaller regression parameters. For model 3, the RMSE of all estimators are similar. The RMSE of the LASSO and the TLSE are significantly smaller than of the LSE for all parameters and sample sizes for the models 4–6. In these cases, the RMSE of LSE for some regression parameters are as high as 12 for model 5 due to the high dimension of the model and high correlation among the variables. Note that, for these models the TLSE generally outperforms the LASSO for all samples and regression parameters.

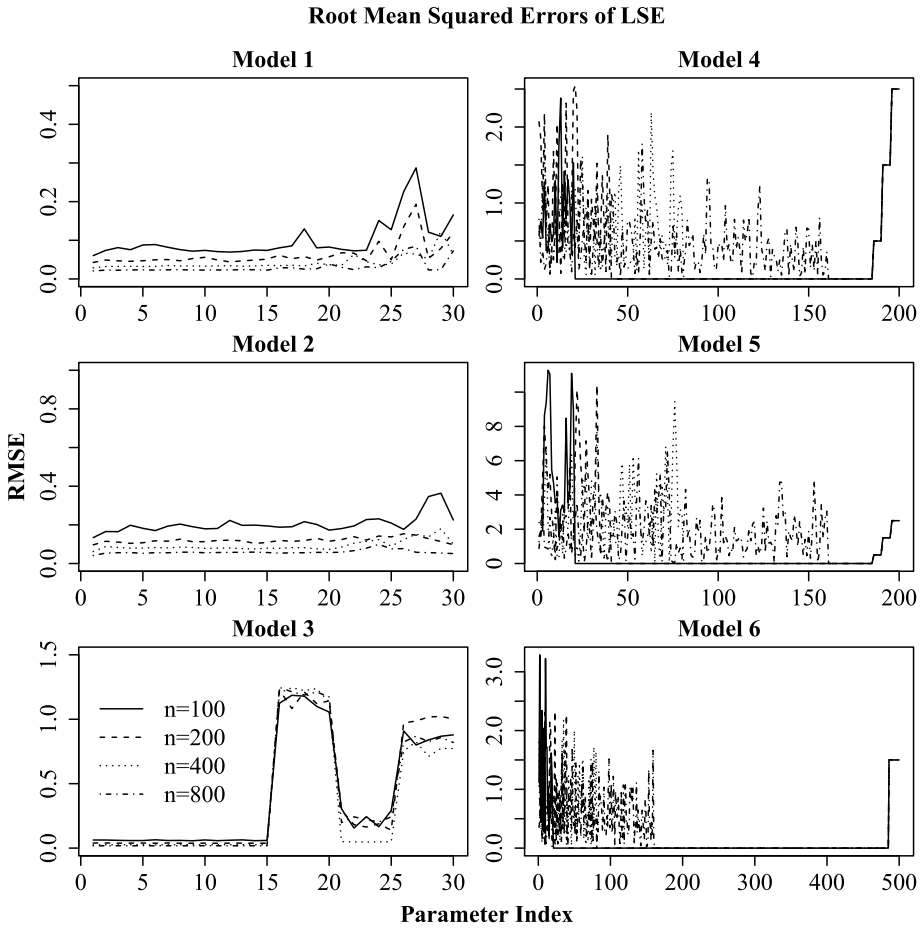

FIG 3. *Empirical root mean squared errors of the LSE for models 1–6. The number of simulations is* $S = 5000$ *and the sample sizes are* $n = 100, 200, 400, 800$. *For models 4–6, the LSE is calculated using the first* $m = \lfloor n/5 \rfloor$ *variables in the models.*

Tables 1–2 show the empirical root mean squared prediction errors (RMSPE) and the empirical root average mean squared errors (RAMSE) of the LSE, the LASSO, and the TLSE, respectively. Note that for models 1–2, the RMSPE of the LSE and the TLSE are similar and significantly smaller than of the LASSO for all sample sizes. For model 3, the RMSPE of the LASSO is smaller than of the LSE and the TLSE. Generally, the RAMSE of all estimators are similar, with smaller values for the LSE and the TLSE. The situation changes dramatically for the high-dimensional models 4–6. Specifically, the RMSPE and RAMSE of the LASSO and the TLSE are significantly smaller than of the LSE, and that generally, the TLSE outperforms the LASSO (with one exception, for model 5 with $n = 200$). The results of the simulation study shows a slightly better per-

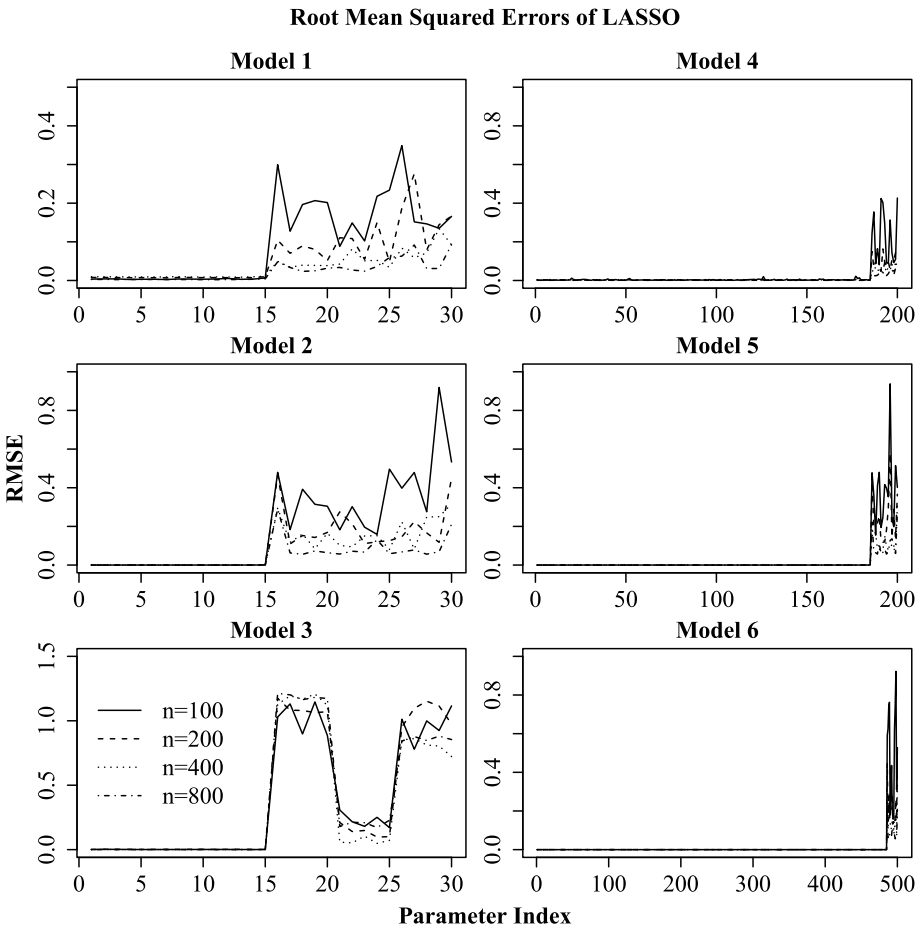

FIG 4. *Empirical root mean squared errors of LASSO for models 1–6. The number of simulations is* $S = 5000$ *and the sample sizes are* $n = 100, 200, 400, 800$.

formance of the LSE and the TLSE over the LASSO in sparse low-dimensional regression models, and a significant better performance of the LASSO and the TLSE for high-dimensional regression models, with slightly better results for the TLSE on smaller and moderate samples.

5. Data analysis

In this section, we use ordinary least-squares and thresholding least-squares methods of inference to analyze a high-dimensional data set. The data set consists of the data collected on intervention chemicals (chemicals given by a ketogenic diet) and seizure load response (measured as the relative percent change

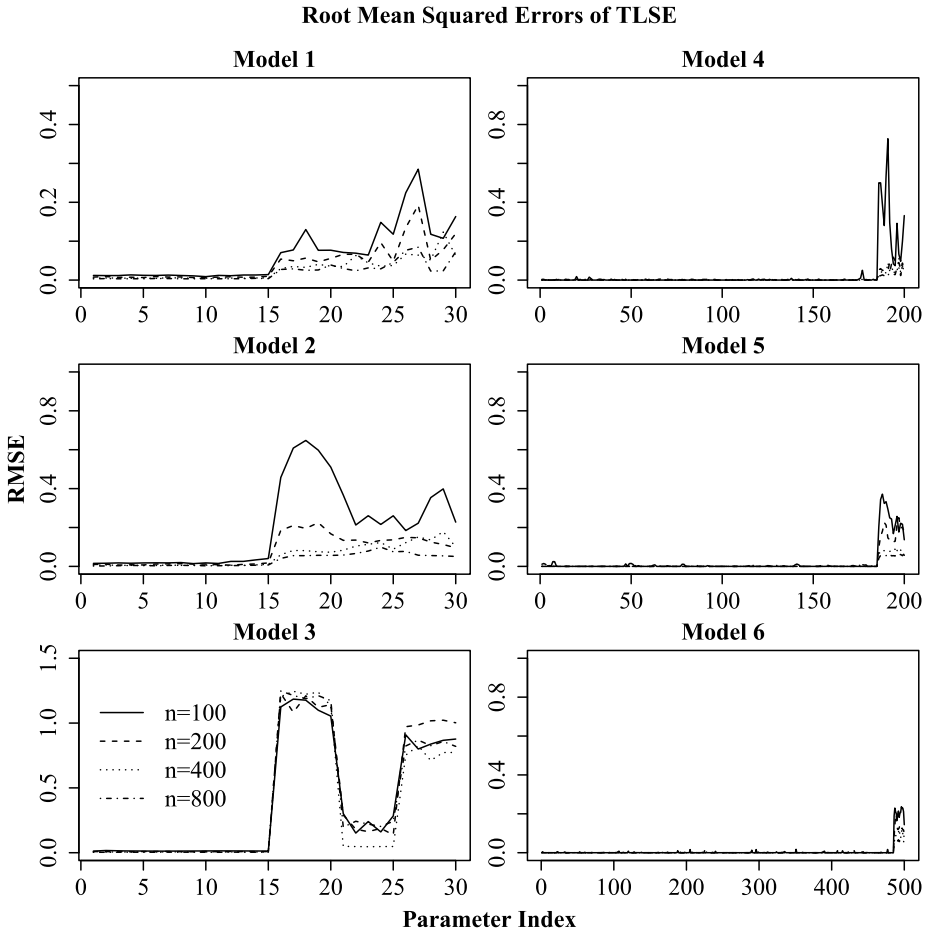

FIG 5. *Empirical root mean squared errors of the TLSE for models 1–6. The number of simulations is* $S = 5000$ *and the sample sizes are* $n = 100, 200, 400, 800$.

in seizure day from the baseline seizure day over a two-week period) for a group of 55 children suffering from epilepsy. After removing incomplete cases as well as explanatory variables with no variation, the pre-processed data set has 6830 observations and 116 explanatory variables. Since the correlation matrix of explanatory variables is nearly singular, with more than 25 eigenvalues smaller than 0.001, we fist perform a hierarchical cluster algorithm to identify groups of highly correlated variables. The (distance) dissimilarity between variables is calculated as 1 minus the absolute value of the correlation coefficient of the variables and we use a group average agglomerative clustering algorithm (see, e.g., Hastie, Tibshirani and Friedman, 2008, Section 14.3.12). The dendrogram is cut at the height of $h = 0.30$; this choice implies that the average dissimi-

TABLE 1

Empirical root mean squared prediction errors of the LSE, the LASSO, and the TLSE for regression models 1–6. The number of simulations is $S = 5000$ and the samples sizes are $n = 100, 200, 400, 800$. For models 4–6, the LSE is calculated using the first $m = \lfloor n/5 \rfloor$ covariates in the models.

Estimate	n	Mod 1	Mod 2	Mod 3	Mod 4	Mod 5	Mod 6
LSE	100	0.73	1.45	1.54	10.72	19.99	19.75
LASSO	100	1.05	2.53	0.92	1.25	2.05	1.8
TLSE	100	0.69	1.5	1.52	1.23	0.9	0.86
LSE	200	0.68	0.62	1.4	12.99	22.27	19.86
LASSO	200	0.88	0.95	1	0.71	0.76	1.26
TLSE	200	0.66	0.62	1.39	0.57	1.18	0.64
LSE	400	0.56	0.94	1.62	12.43	21.23	19.82
LASSO	400	0.59	1.45	1.35	0.57	0.89	0.99
TLSE	400	0.55	0.93	1.62	0.53	0.52	0.57
LSE	800	0.52	0.58	1.75	11.78	20.75	18.95
LASSO	800	0.53	0.62	1.58	0.54	0.75	0.59
TLSE	800	0.52	0.57	1.75	0.54	0.51	0.55

TABLE 2

Empirical root average mean squared errors of the LSE, the LASSO, and the TLSE for regression models 1–6. The number of simulations is $S = 5000$ and the sample sizes are $n = 100, 200, 400, 800$. For models 4–6, the LSE is calculated using the first $m = \lfloor n/5 \rfloor$ covariates in the models.

Estimate	n	Mod 1	Mod 2	Mod 3	Mod 4	Mod 5	Mod 6
LSE	100	0.112	0.209	0.59	0.577	2.102	0.4
LASSO	100	0.139	0.296	0.584	0.071	0.117	0.094
TLSE	100	0.096	0.283	0.587	0.098	0.071	0.036
LSE	200	0.071	0.121	0.631	0.749	2.108	0.412
LASSO	200	0.092	0.161	0.625	0.029	0.075	0.046
TLSE	200	0.062	0.11	0.63	0.021	0.048	0.023
LSE	400	0.046	0.097	0.585	0.714	2.288	0.457
LASSO	400	0.047	0.131	0.58	0.017	0.037	0.028
TLSE	400	0.039	0.078	0.585	0.015	0.022	0.016
LSE	800	0.035	0.061	0.607	0.66	2.038	0.483
LASSO	800	0.035	0.081	0.604	0.013	0.033	0.018
TLSE	800	0.031	0.045	0.607	0.013	0.016	0.01

larity (distance) between the clusters is larger than 0.30, and thus, the average absolute correlation between the clusters is smaller than 0.70. For each group of variables determined by the hierarchical clustering algorithm, we perform a principal component analysis and use the scores of the first principal components as covariates in the regression analysis. Our final design matrix has $p = 74$ columns and $n = 6830$ rows, and its minimum eigenvalue is $\rho_1 = 0.004$ (which is greater than $\log(n)/n = 0.001$).

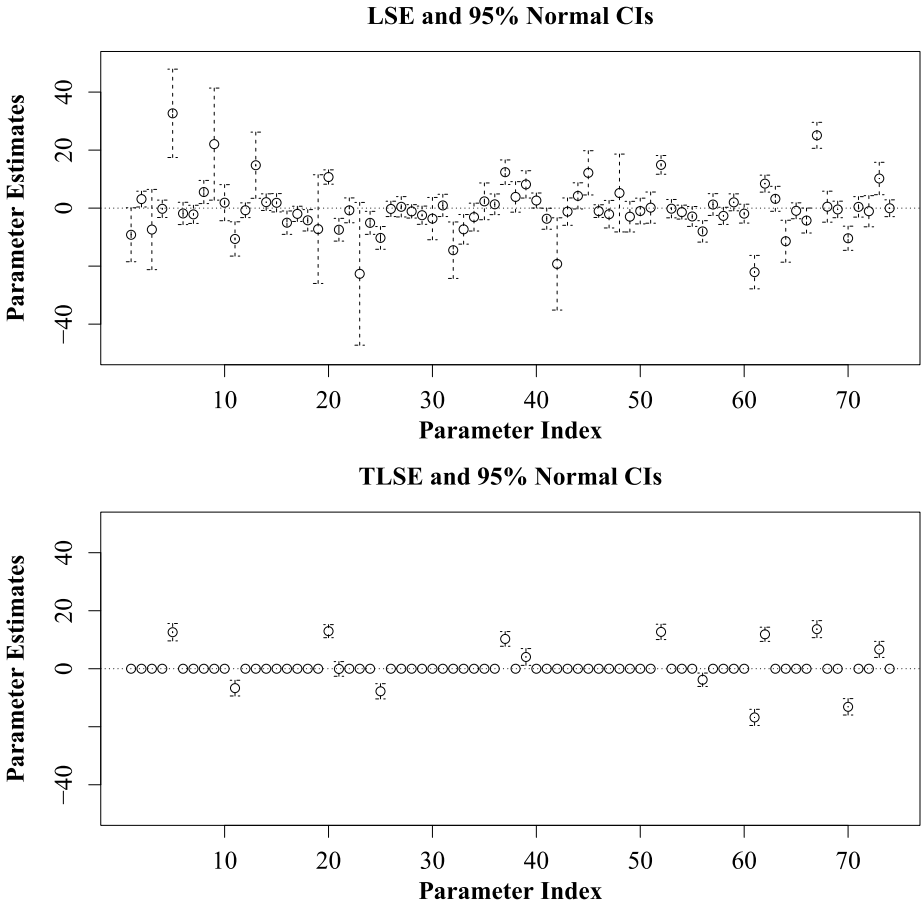

FIG 6. *Least-squares estimates (LSE) and thresholding least-squares estimates (TLSE) of regression parameters with the corresponding 95% normal confidence intervals (CIs).*

Figure 6 shows the LSE and TLSE of the regression parameters with the corresponding 95% individual normal confidence intervals. The cardinality of the non-zero regression parameter estimates of the TLSE is 14 (and thus, $74 - 14 = 60$ components of the TLSE are set equal to zero). Note that, the zero components of the TLSE correspond to non-significant components of the LSE, and the widths of the corresponding confidence intervals are significantly smaller. This is in agreement with the theoretical and simulation results which shows that the TLSE is more efficient than the LSE for sparse regression models.

Appendix

Proof of Lemma 2.1. Note first that

$$\operatorname{var}\big(a^T(\bar\beta-\beta)\big)=\sigma^2a^T(X^TX)^{-1}a\le n^{-1}\sigma^2\rho_1^{-1}=O\big((n\rho_1)^{-1}\big)\,.$$

Hence $a^T(\bar\beta-\beta)=O_P((\rho_1 n)^{-1/2})$, as stated. To prove (ii), note that

$$\bar\sigma^2=(n-p)^{-1}Y^T(I-H)Y=(n-p)^{-1}\epsilon_{1:n}^T(I-H)\epsilon_{1:n}\,,$$

where $H=X\big(X^TX\big)^{-1}X^T$ is the "hat matrix", $H=(h_{ij})\in\mathbb{R}^{n\times n}$, and $\epsilon_{1:n}=(\epsilon_1,\dots,\epsilon_n)^T\in\mathbb{R}^n$. Since H is idempotent of rank p, then $\sum_{j=1}^n h_{jj}=p$. Thus,

$$\begin{aligned}\sum_{j=1}^n h_{jj}^2&\le p\max_{1\le j\le n}h_{jj}=p\max_{1\le j\le n}X_j^T(X^TX)^{-1}X_j\\&\le\big(p/(\rho_1 n)\big)\max_{1\le j\le n}\|X_j\|^2=O\big(p^2/(n\rho_1)\big)=o(p)\,.\end{aligned}\tag{A.1}$$

Hence,

$$\begin{aligned}\operatorname{var}(\bar\sigma^2)&=(n-p)^{-2}\operatorname{var}\big(\epsilon_{1:n}^T(I-H)\epsilon_{1:n}\big)\\&=(n-p)^{-2}\Big(2\sigma^4\operatorname{tr}\big((I-H)^2\big)+(\mu_4-3\sigma^4)\sum_{j=1}^n(1-h_{jj})^2\Big)\\&=(n-p)^{-2}\Big[2\sigma^4(n-p)+(\mu_4-3\sigma^4)\Big(n-2p+\sum_{j=1}^n h_{jj}^2\Big)\Big]=O\big(n^{-1}\big)\,,\end{aligned}$$

where $\mu_4=\mathrm{E}(\epsilon^4)$. Thus, $\bar\sigma^2=\sigma^2+O_P(n^{-1/2})$, as stated. □

Proof of Lemma 2.2. Since $\bar\sigma^2=\sigma^2+o_P(1)$, by Slutsky's theorem, it is enough to show that $s^{-1/2}a^T(\bar\beta-\beta)\xrightarrow{\mathrm{d}}N(0,1)$, where $s=n^{-1}\sigma^2a^T\Omega^{-1}a$. To this end, note that

$$s^{-1/2}a^T(\bar\beta-\beta)=s^{-1/2}a^T(X^TX)^{-1}X^T\epsilon_{1:n}=\sum_{i=1}^n\alpha_i\epsilon_i\,,\tag{A.2}$$

where $\alpha_i=s^{-1/2}a^T(X^TX)^{-1}X_i\in\mathbb{R}$. Thus, we have to show that $\sum_{i=1}^n\alpha_i\epsilon_i\xrightarrow{\mathrm{d}}N(0,1)$. To this end, note first that $\mathrm{E}(\alpha_i\epsilon_i)=0$. Let $\sigma_i^2=\operatorname{var}(\alpha_i\epsilon_i)$. Since

$$\sigma_i^2=s^{-1}\sigma^2a^T(X^TX)^{-1}X_iX_i^T(X^TX)^{-1}a\,,$$

then $\sum_{i=1}^n\sigma_i^2=1$. Thus, it is enough to show that the Lindeberg condition holds (see, e.g., van der Vaart, 1998, p. 20):

$$\sum_{i=1}^n\mathrm{E}\big(|\alpha_i\epsilon_i|^2I(|\alpha_i\epsilon_i|\ge\delta)\big)\to0\quad\text{for all}\quad\delta>0\,.\tag{A.3}$$

Since $\alpha_i^2 = s^{-1}a^T(X^TX)^{-1}X_iX_i^T(X^TX)^{-1}a$, then $\sum_{i=1}^n \alpha_i^2 = 1/\sigma^2$. Thus, to prove (A.3), it is enough to show that

$$\max_{1\le i\le n} \mathrm{E}\big(\epsilon^2 I(|\alpha_i\epsilon| \ge \delta)\big) \to 0\,. \tag{A.4}$$

Since

$$\max_{1\le i\le n} \mathrm{E}\big(\epsilon^2 I(|\alpha_i\epsilon| \ge \delta)\big) \le \mathrm{E}\Big(\epsilon^2 I(|\epsilon| \max_{1\le i\le n}|\alpha_i| \ge \delta)\Big)\,,$$

then (A.4) holds provided that $\max_{1\le i\le n}|\alpha_i| = o(1)$. To this end, note that

$$\begin{aligned}\max_{1\le i\le n} \alpha_i^2 &= \max_{1\le i\le n}\left\{\frac{a^T(X^TX)^{-1}X_iX_i^T(X^TX)^{-1}a}{\sigma^2 a^T(X^TX)^{-1}a}\right\}\\ &\le \frac{\max_{1\le i\le n}\|X_i\|^2}{n\sigma^2\rho_1} = O\big(p/(n\rho_1)\big) = o(1)\,.\end{aligned}$$

This concludes the proof of the lemma. □

Proof of Lemma 2.3. Without loss of generality, we state (and prove) this result for the case when $K = I$. By rescaling, we further assume without loss of generality that $\sigma^2 = 1$. Since $H_0 : K_\beta = I$ holds, we have

$$\begin{aligned}n\bar\beta^T\Omega\bar\beta &= n\epsilon_{1:n}^T X(X^TX)^{-1}\Omega(X^TX)^{-1}X^T\epsilon_{1:n}\\ &= \left(\sum_{i=1}^n \epsilon_i X_i\right)^T (X^TX)^{-1}\left(\sum_{i=1}^n \epsilon_i X_i\right).\end{aligned}$$

Let

$$Y_{j,n} = \left(\sum_{i=1}^j \epsilon_i X_i\right)^T (X^TX)^{-1}\left(\sum_{i=1}^j \epsilon_i X_i\right) - \sum_{i=1}^j h_{ii}\,.$$

Since $\sum_{i=1}^n h_{ii} = p$, then (2.3) holds provided that

$$\frac{Y_{n,n}}{(2p)^{1/2}} \xrightarrow{\mathrm{d}} N(0,1)\,. \tag{A.5}$$

Note that $\{(Y_{j,n}, \mathscr{F}_{j,n}) : 1 \le j \le n\}$ is a martingale array, where $\mathscr{F}_{j,n} = \sigma_a\{\epsilon_i : 1 \le i \le j\}$ is the *natural filtration* and $\sigma_a\{\epsilon_i : 1 \le i \le j\}$ denotes the σ-field generated by $\{\epsilon_1, \ldots, \epsilon_j\}$. To this end, since $Y_{j,n}$ is $\mathscr{F}_{j,n}$-measurable by definition, we have to show that $\mathrm{E}(Y_{j,n}|\mathscr{F}_{j-1,n}) = Y_{j-1,n}$ almost surely (a. s.), $j = 1, \ldots, n$, where $Y_{0,n} = 0$ a. s., and $\mathscr{F}_{0,n}$ is the trivial σ-field. Let $T_j = \sum_{i=1}^j \epsilon_i X_i$. Then, we write

$$\begin{aligned}Y_{j,n} &= (T_{j-1} + \epsilon_j X_j)^T (X^TX)^{-1}(T_{j-1} + \epsilon_j X_j) - \sum_{i=1}^{j-1} h_{ii} - h_{jj}\\ &= Y_{j-1,n} + 2\epsilon_j X_j^T (X^TX)^{-1}T_{j-1} + \epsilon_j^2 h_{jj} - h_{jj}\\ &= Y_{j-1,n} + 2\epsilon_j \sum_{i=1}^{j-1} h_{ji}\epsilon_i + (\epsilon_j^2 - 1)h_{jj}\,.\end{aligned} \tag{A.6}$$

Since $\mathrm{E}(\epsilon_j|\mathscr{F}_{j-1,n}) = 0$ a. s. and $\mathrm{E}(\epsilon_j^2|\mathscr{F}_{j-1,n}) = 1$ a. s., then $\mathrm{E}(Y_{j,n}|\mathscr{F}_{j-1,n}) = Y_{j-1,n}$ a. s.. Hence $\{(Y_{j,n}, \mathscr{F}_{j,n}) : 1 \le j \le n\}$ is a martingale array. Consider the martingale difference array $\{(Z_{j,n}, \mathscr{F}_{j,n}) : 1 \le j \le n\}$, where

$$\begin{aligned} Z_{j,n} &= (2p)^{-1/2}\big(Y_{j,n} - Y_{j-1,n}\big) \\ &= (2p)^{-1/2}\left(2\epsilon_j \sum_{i=1}^{j-1} h_{ji}\epsilon_i + (\epsilon_j^2 - 1)h_{jj}\right). \end{aligned} \tag{A.7}$$

Let $\nu_{j,n}^2 = \mathrm{E}(Z_{j,n}^2)$ and $\nu_n^2 = \sum_{j=1}^n \nu_{j,n}^2$. Note that

$$Z_{j,n}^2 = \frac{1}{2p}\left(4\epsilon_j^2\left(\sum_{i=1}^{j-1} h_{ji}\epsilon_i\right)^2 + (\epsilon_j^2 - 1)^2 h_{jj}^2 + 4\epsilon_j(\epsilon_j^2 - 1)h_{jj}\sum_{i=1}^{j-1} h_{ji}\epsilon_i\right), \tag{A.8}$$

and thus,

$$\nu_{j,n}^2 = \frac{1}{2p}\left(4\sum_{i=1}^{j-1} h_{ji}^2 + (\mu_4 - 1)h_{jj}^2\right). \tag{A.9}$$

By (A.1), we further obtain

$$\begin{aligned} \nu_n^2 &= \frac{1}{2p}\sum_{j=1}^n\left(4\sum_{i=1}^{j-1} h_{ji}^2 + (\mu_4 - 1)h_{jj}^2\right) \\ &= \frac{1}{2p}\sum_{j=1}^n\left(2\sum_{i=1}^{n} h_{ji}^2 + (\mu_4 - 3)h_{jj}^2\right) \\ &= p^{-1}\operatorname{tr}(H^2) + (2p)^{-1}(\mu_4 - 3)\sum_{j=1}^n h_{jj}^2 = 1 + o(1)\,. \end{aligned} \tag{A.10}$$

By the central limit theorem for martingale difference arrays (see, e.g., Chow and Teicher, 1997) and (Athreya and Lahiri, 2006, p. 510), then (A.5) holds provided that the following conditions hold:

$$\sum_{j=1}^n \mathrm{E}\left|Z_{j,n}^3\right| \to 0\,, \tag{A.11a}$$

and

$$\sum_{j=1}^n \mathrm{E}\big(Z_{j,n}^2\big|\mathscr{F}_{j-1,n}\big) \xrightarrow{\Pr} 1\,, \tag{A.11b}$$

where $\xrightarrow{\Pr}$ denotes convergence in probability. To prove (A.11a), we have to show that

$$\frac{1}{(2p)^{3/2}}\sum_{j=1}^n \mathrm{E}\left(\left|2\epsilon_j\sum_{i=1}^{j-1} h_{ji}\epsilon_i + (\epsilon_j^2 - 1)h_{jj}\right|^3\right) \to 0\,. \tag{A.12}$$

Using Jensen's inequality $|x+y|^3 \le 4(|x|^3+|y|^3)$ for $x, y \in \mathbb{R}$, we obtain

$$\frac{1}{p^{3/2}}\sum_{j=1}^{n}\mathrm{E}\Big(\Big|2\epsilon_j\sum_{i=1}^{j-1}h_{ji}\epsilon_i+(\epsilon_j^2-1)h_{jj}\Big|^3\Big) \le \frac{4}{p^{3/2}}\sum_{j=1}^{n}\mathrm{E}\Big(\Big|2\epsilon_j\sum_{i=1}^{j-1}h_{ji}\epsilon_i\Big|^3\Big)+\frac{4}{p^{3/2}}\sum_{j=1}^{n}\mathrm{E}\Big(\big|(\epsilon_j^2-1)^3h_{jj}^3\big|\Big). \quad (A.13)$$

We first prove that the first term on the right side of (A.13) tends to 0. To this end, using the identity $\sum_{j_1=1}^{n}h_{jj_1}^2=h_{jj}$ (since H is an idempotent matrix) and Holder's inequality $\mathrm{E}(|Y|^3)\le\big(\mathrm{E}(|Y|^4)\big)^{3/4}$, we have

$$\mathrm{E}\Big(\Big|\epsilon_j\sum_{i=1}^{j-1}h_{ji}\epsilon_i\Big|^3\Big) \le \Big\{\mathrm{E}\Big(\Big|\epsilon_j\sum_{i=1}^{j-1}h_{ji}\epsilon_i\Big|^4\Big)\Big\}^{3/4} = \Big\{\mu_4\,\mathrm{E}\Big(\sum_{j_1=1}^{j-1}\sum_{j_2=1}^{j-1}\sum_{j_3=1}^{j-1}\sum_{j_4=1}^{j-1}h_{jj_1}h_{jj_2}h_{jj_3}h_{jj_4}\epsilon_{j_1}\epsilon_{j_2}\epsilon_{j_3}\epsilon_{j_4}\Big)\Big\}^{3/4} \le \Big\{\mu_4\sigma^4\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}h_{jj_1}^2h_{jj_2}^2\Big\}^{3/4}=\mu_4^{3/4}\sigma^3h_{jj}^{3/2}.$$

By Cauchy-Schwarz inequality, we thus obtain:

$$\Big(\frac{1}{p^{3/2}}\sum_{j=1}^{n}\mathrm{E}\Big|\epsilon_j\sum_{i=1}^{j-1}h_{ji}\epsilon_i\Big|^3\Big)^2 \le \frac{n}{p^3}\sum_{j=1}^{n}\Big\{\mathrm{E}\Big(\Big|\epsilon_j\sum_{i=1}^{j-1}h_{ji}\epsilon_i\Big|^3\Big)\Big\}^2 \le \mu_4^{3/2}\sigma^6\frac{n}{p^3}\sum_{j=1}^{n}h_{jj}^3 \le \mu_4^{3/2}\sigma^6\frac{n}{p^3}\sum_{j=1}^{n}h_{jj}\Big\{\max_{1\le j\le n}h_{jj}^2\Big\} = \frac{n}{p^3}pO\big(p^2/(n\rho_1)^2\big)=O\big(1/(n\rho_1^2)\big)=o(1).$$

We now show that the second term on the right side of (A.13) tends to 0. To this end, note that

$$\frac{1}{p^{3/2}}\sum_{j=1}^{n}|h_{jj}^3| \le \frac{1}{p^{3/2}}p\max_{1\le j\le n}h_{jj}^2 = p^{-1/2}O\big(p^2/(n\rho_1)^2\big)=o(1),$$

and thus, (A.11a) holds, as stated. Lastly, to prove (A.11b), note first that

$$\mathrm{E}\big(Z_{j,n}^2|\mathscr{F}_{j-1,n}\big)=\frac{1}{2p}\Big(4\Big(\sum_{i=1}^{j-1}h_{ji}\epsilon_i\Big)^2+(\mu_4-1)h_{jj}^2+4\mu_3\sum_{i=1}^{j-1}h_{jj}h_{ji}\epsilon_i\Big), \quad (A.14)$$

where $\mu_3=\mathrm{E}(\epsilon^3)$. By Cauchy-Schwarz inequality and (A.1), we have

$$p^{-2}\,\mathrm{var}\Big(\sum_{j=1}^{n}\sum_{i=1}^{j-1}h_{jj}h_{ji}\epsilon_i\Big)=p^{-2}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\sum_{i=1}^{\min(j_1,j_2)}h_{j_1j_1}h_{j_1i}h_{j_2j_2}h_{j_2i}$$

$$\leq p^{-2} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} h_{j_1 j_1} h_{j_2 j_2} \sum_{i=1}^{n} h_{j_1 i}^2 \sum_{i=1}^{n} h_{j_2 i}^2$$
$$= p^{-2} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} h_{j_1 j_1}^2 h_{j_2 j_2}^2 = p^{-2} o(p^2) = o(1)\,.$$

Hence (A.11b) holds provided that

$$\operatorname{var}\Bigg\{ p^{-1} \sum_{j=1}^{n} \Bigg(\sum_{i=1}^{j-1} h_{ji} \epsilon_i \Bigg)^2 \Bigg\} = o(1)\,. \tag{A.15}$$

To this end, note that

$$\operatorname{var}\Bigg\{ p^{-1} \sum_{j=1}^{n} \Bigg(\sum_{i=1}^{j-1} h_{ji} \epsilon_i \Bigg)^2 \Bigg\}$$
$$= p^{-2} \sum_{j=1}^{n} \sum_{k=1}^{n} \operatorname{cov}\Bigg(\sum_{j_1, j_2=1}^{j-1} h_{jj_1} h_{jj_2} \epsilon_{j_1} \epsilon_{j_2}, \sum_{k_1, k_2=1}^{k-1} h_{kk_1} h_{kk_2} \epsilon_{k_1} \epsilon_{k_2} \Bigg)\,.$$

Note that in the sum above, only the terms for which pairs of indices are equal are non-zero. Consider the terms for which $j_1 = j_2$ and $k_1 = k_2$ (all other terms are treated similarly). Hence, we have to show that

$$p^{-2} \sum_{j=1}^{n} \sum_{k=1}^{n} \operatorname{cov}\Bigg(\sum_{j_1=1}^{j-1} h_{jj_1}^2 \epsilon_{j_1}^2, \sum_{k_1=1}^{k-1} h_{kk_1}^2 \epsilon_{k_1}^2 \Bigg) = o(1)\,.$$

Note further that in the sum above, only the terms for which $j_1 = k_1$ are non-zero, and thus, by (A.1), we have

$$p^{-2} \sum_{j=1}^{n} \sum_{k=1}^{n} \operatorname{cov}\Bigg(\sum_{j_1=1}^{j-1} h_{jj_1}^2 \epsilon_{j_1}^2, \sum_{k_1=1}^{k-1} h_{kk_1}^2 \epsilon_{k_1}^2 \Bigg)$$
$$= p^{-2} (\mu_4 - 1) \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{j_1=1}^{\min(j,k)-1} h_{jj_1}^2 h_{kj_1}^2$$
$$\leq p^{-2} (\mu_4 - 1) \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{j_1=1}^{n} h_{jj_1}^2 h_{kj_1}^2$$
$$\leq p^{-2} (\mu_4 - 1) \sum_{j_1=1}^{n} h_{j_1 j_1}^2 = p^{-2} O\big(p^2/(n\rho_1)\big) = o(1)\,.$$

This completes the proof. □

Proof of Theorem 2.1. Let $\Xi_{K_\beta K_\beta} = \operatorname{diag}\big(\Sigma_{K_\beta K_\beta}^{-1}\big) \in \mathbb{R}^{q_{K_\beta} \times q_{K_\beta}}$ be the diagonal matrix of $\Sigma_{K_\beta K_\beta}^{-1} = (\Omega^{-1})_{K_\beta K_\beta} \in \mathbb{R}^{q_{K_\beta} \times q_{K_\beta}}$, where recall that $q_{K_\beta} = \operatorname{card}(K_\beta)$.

Note first that since $\Sigma^{-1}_{K_\beta K_\beta}$ is a diagonal submatrix of Ω^{-1}, then $\|\Sigma^{-1}_{K_\beta K_\beta}\| \leq \|\Omega^{-1}\| = \rho_1^{-1}$, where $\|A\|$ is the spectral norm of a matrix A (see, e.g., Bhatia, 1997). Since $\Xi_{K_\beta K_\beta}$ is the diagonal matrix of $\Sigma^{-1}_{K_\beta K_\beta}$ and $\Sigma^{-1}_{K_\beta K_\beta}$ is a diagonal submatrix of Ω^{-1}, then

$$\operatorname{mineig}(\Xi_{K_\beta K_\beta}) \geq \operatorname{mineig}\big(\Sigma^{-1}_{K_\beta K_\beta}\big) \geq \operatorname{mineig}(\Omega^{-1}) = \rho_2^{-1}\,,$$

where $\operatorname{mineig}(A)$ is the minimum eigenvalue of a symmetric matrix A. Hence, $\|\Xi^{-1}_{K_\beta K_\beta}\| \leq \rho_2$. Since $\|AB\| \leq \|A\|\|B\|$, we further obtain:

$$\begin{aligned}
\Pr\Big(\max_{j\in K_\beta}\frac{|\bar\beta_j|}{\bar\sigma_{jj}} \geq \gamma\Big) &\leq \Pr\Big(\sum_{j\in K_\beta}\frac{\bar\beta_j^2}{\bar\sigma_{jj}^2} \geq \gamma^2\Big)\\
&= \Pr\Big(n\bar\sigma^{-2}\bar\beta^T_{K_\beta}\Xi^{-1}_{K_\beta}\bar\beta_{K_\beta} \geq \gamma^2\Big)\\
&\leq \Pr\Big(q_{K_\beta}\hat F(K_\beta)\|\Xi^{-1}_{K_\beta}\|\|\Sigma^{-1}_{K_\beta}\| \geq \gamma^2\Big)\\
&\leq \Pr\Big(q_{K_\beta}\rho_2\rho_1^{-1}\hat F(K_\beta) \geq \gamma^2\Big)\\
&= \Pr\Big(q_{K_\beta}\hat F(K_\beta) \geq \rho_1\rho_2^{-1}\gamma^2\Big)\,.
\end{aligned} \tag{A.16}$$

By (2.4a) and Lemma 2.3, then $\Pr\big(q_{K_\beta}\hat F(K_\beta) \geq \rho_1\rho_2^{-1}\gamma^2\big) \to 0$. Hence, by (A.16),

$$\Pr\Big(\frac{|\bar\beta_j|}{\bar\sigma_{jj}} \leq \gamma \text{ for all } j \in K_\beta\Big) \to 1\,,$$

and thus, $\Pr\big(K_\beta \subset \hat K\big) \to 1$.

Since $\Sigma^{-1}_{J_\beta J_\beta}$ is a diagonal submatrix of Ω^{-1}, then $\max_{j\in J_\beta}\omega_{jj} \leq \|\Sigma^{-1}_{J_\beta J_\beta}\| \leq \|\Omega^{-1}\| = \rho_1^{-1}$. Hence

$$\begin{aligned}
\Pr\Big(\min_{j\in J_\beta}\frac{|\bar\beta_j|}{\bar\sigma_{jj}} \leq \gamma\Big) &\leq \Pr\Big(\min_{j\in J_\beta}\frac{|\beta_j| - |\bar\beta_j - \beta_j|}{\bar\sigma_{jj}} \leq \gamma\Big)\\
&\leq \Pr\Big(\min_{j\in J_\beta}\frac{|\beta_j|}{\bar\sigma_{jj}} - \max_{j\in J_\beta}\frac{|\bar\beta_j - \beta_j|}{\bar\sigma_{jj}} \leq \gamma\Big)\\
&= \Pr\Big(\max_{j\in J_\beta}\frac{|\bar\beta_j - \beta_j|}{\bar\sigma_{jj}} \geq \min_{j\in J_\beta}\frac{|\beta_j|}{\bar\sigma_{jj}} - \gamma\Big)\\
&= \Pr\Big(\max_{j\in J_\beta}\frac{|\bar\beta_j - \beta_j|}{\bar\sigma_{jj}} \geq n^{1/2}\bar\sigma^{-1}\min_{j\in J_\beta}\frac{|\beta_j|}{\omega_{jj}^{1/2}} - \gamma\Big)\\
&\leq \Pr\Big(\max_{j\in J_\beta}\frac{|\bar\beta_j - \beta_j|}{\bar\sigma_{jj}} \geq n^{1/2}\bar\sigma^{-1}\rho_1^{1/2}\min_{j\in J_\beta}|\beta_j| - \gamma\Big)\,.
\end{aligned} \tag{A.17}$$

By (2.4b) and Lemma 2.3, similarly to (A.16), we obtain

$$\Pr\Big(\min_{j\in J_\beta}\frac{|\bar\beta_j|}{\bar\sigma_{jj}} \leq \gamma\Big) \to 0\,.$$

Hence, $\Pr\left(J_\beta \cap \hat{K} \neq \varnothing\right) \to 0$, and thus part (i) holds, as stated. Part (ii) follows similarly to the proof of Lemma 2.2. □

Proof of Corollary 2.1. Since $\Sigma_{\hat{J}\hat{J}} \preceq \Omega_{\hat{J}\hat{J}}$, then $\Omega^{-1}_{\hat{J}\hat{J}} \preceq \Sigma^{-1}_{\hat{J}\hat{J}}$, where $\preceq$ denotes the Loewner partial ordering of symmetric matrices, where recall that $\Omega^{-1}_{\hat{J}\hat{J}}$ is the inverse of $\Omega_{\hat{J}\hat{J}}$. This further implies that $\hat{s} \leq \bar{s}$ a. s.. This completes the proof. □

Proof of Theorem 2.2. Since $m = o(n)$ and $m \geq q$, it readily follows that (i) holds provided that $\Pr(\min_{j\in J_\beta} \tilde{\gamma}_j \geq \max_{k\in K_\beta} \tilde{\gamma}_k) \to 1$. To this end, note that

$$\begin{aligned}
\Pr\Big(\min_{j\in J_\beta} \tilde{\gamma}_j \geq \max_{k\in K_\beta} \tilde{\gamma}_k\Big) &\geq 1 - \sum_{j\in J_\beta}\sum_{k\in K_\beta} \Pr\big(\tilde{\gamma}_j \leq \tilde{\gamma}_k\big) \\
&= 1 - \sum_{j\in J_\beta}\sum_{k\in K_\beta} \Pr\left(\left|n^{-1}\sum_{i=1}^n X_{ij}Y_i\right| \leq \left|n^{-1}\sum_{i=1}^n X_{ik}Y_i\right|\right) \\
= 1 - \sum_{j\in J_\beta}\sum_{k\in K_\beta} \Pr\left(\left|n^{-1}\sum_{i=1}^n X_{ij}\big(X_i^T\beta + \epsilon_i\big)\right| \leq \left|n^{-1}\sum_{i=1}^n X_{ik}\big(X_i^T\beta + \epsilon_i\big)\right|\right) \\
&= 1 - \sum_{j\in J_\beta}\sum_{k\in K_\beta} \Pr\left(\left|\zeta_j + n^{-1}\sum_{i=1}^n X_{ij}\epsilon_i\right| \leq \left|\zeta_k + n^{-1}\sum_{i=1}^n X_{ik}\epsilon_i\right|\right).
\end{aligned}$$

Note further that for $k \in K_\beta$, we have

$$\begin{aligned}
|\zeta_k| = \left|n^{-1}\sum_{i=1}^n X_{ik}X_i^T\beta\right| &= \left|n^{-1}\sum_{i=1}^n \sum_{j\in J_\beta} X_{ik}X_{ij}\beta_j\right| \\
&\leq q\Big\{\max_{j\in J_\beta}|\beta_j|\Big\} \max_{j\in J_\beta, k\in K_\beta} n^{-1}\left|\sum_{i=1}^n X_{ik}X_{ij}\right| = o(1)\,.
\end{aligned}$$

Hence, for n sufficiently large,

$$\begin{aligned}
&\Pr\Big(\min_{j\in J_\beta} \tilde{\gamma}_j \geq \max_{k\in K_\beta} \tilde{\gamma}_k\Big) \\
&\quad\geq 1 - \sum_{j\in J_\beta}\sum_{k\in K_\beta} \Pr\left(|\zeta_j| - \left|n^{-1}\sum_{i=1}^n X_{ij}\epsilon_i\right| \leq |\zeta_k| + \left|n^{-1}\sum_{i=1}^n X_{ik}\epsilon_i\right|\right) \\
&\quad\geq 1 - \sum_{j\in J_\beta}\sum_{k\in K_\beta} \Pr\left(\left|n^{-1}\sum_{i=1}^n X_{ij}\epsilon_i\right| + \left|n^{-1}\sum_{i=1}^n X_{ik}\epsilon_i\right| \geq c_2 - o(1)\right) \\
&\qquad\geq 1 - 2q(p-q)\max_{j\in I}\left\{\Pr\left(\left|n^{-1}\sum_{i=1}^n X_{ij}\epsilon_i\right| \geq c_2/4\right)\right\} \\
&\qquad\quad= 1 - 2q(p-q)\max_{j\in I}\left\{\Pr\Big(n^{-1/2}\big|\epsilon_{1:n}^T X^{(j)}\big| \geq n^{1/2}c_2/4\Big)\right\},
\end{aligned}$$

where $X^{(j)} = (X_{1j}, \ldots, X_{nj})^T \in \mathbb{R}^n$. By Lemma 4 (i) of Huang, Horowitz and Ma (2008), $n^{-1/2}\epsilon_{1:n}^T X^{(j)}$ is sub-Gaussian, and thus, there exists K_1 and K_2 such that

$$\Pr\big(n^{-1/2}\big|\epsilon_{1:n}^T X^{(j)}\big| \geq x\big) \leq K_1 \exp(-K_2 x^2)\,,$$

where K_1 and K_2 do not depend on j. Hence,

$$\Pr\Big(\min_{j\in J_\beta} \tilde{\gamma}_j \geq \max_{k\in K_\beta} \tilde{\gamma}_k\Big) \geq 1 - 2q(p-q)K_1 \exp\Big(-K_2 n (c_2/4)^2\Big) \to 1\,.$$

Thus, (i) follows, as stated. Part (ii) follows as in the proof of Theorem 2.1. □

Proof of Theorem 3.1. Since $\hat{\sigma}^{*2} = \sigma^2 + o_P(1)$, by the (conditional) Slutsky's theorem, it is enough to show the consistency result for $\mathscr{L}(\hat{s}^{-1/2} a^T(\hat{\beta}^* - \hat{\beta})|Y)$. To this end, note first that

$$\hat{s}^{-1/2} a^T(\hat{\beta}^* - \hat{\beta}) = \hat{s}^{-1/2} a_{\hat{J}}^T (X_{\hat{J}}^T X_{\hat{J}})^{-1} X_{\hat{J}}^T \hat{\epsilon}_{1:n}^* = \sum_{i=1}^n \hat{\alpha}_i \hat{\epsilon}_i^*\,, \tag{A.18}$$

where $\hat{\alpha}_i = \hat{s}^{-1/2} a_{\hat{J}}^T (X_{\hat{J}}^T X_{\hat{J}})^{-1} X_{i,\hat{J}} \in \mathbb{R}$, $X_{i,\hat{J}} = (X_{ij} : j \in \hat{J})^T \in \mathbb{R}^{q_{\hat{J}}}$, and $\hat{\epsilon}_{1:n}^* = (\hat{\epsilon}_1^*, \ldots, \hat{\epsilon}_n^*)^T \in \mathbb{R}^n$. Thus, we have to show that

$$\mathscr{L}\Big(\sum_{i=1}^n \hat{\alpha}_i \hat{\epsilon}_i^* \big| Y\Big) \xrightarrow{\Pr} N(0,1)\,.$$

To this end, note first that $\mathrm{E}(\hat{\alpha}_i \hat{\epsilon}_i^*|Y) = 0$ a. s.. Let $\hat{\sigma}_i^2 = \mathrm{var}(\hat{\alpha}_i \hat{\epsilon}_i^*|Y)$. Since

$$\hat{\sigma}_i^2 = \hat{s}^{-1} \hat{\sigma}^2 a_{\hat{J}}^T \big(X_{\hat{J}}^T X_{\hat{J}}\big)^{-1} X_{i,\hat{J}} X_{i,\hat{J}}^T \big(X_{\hat{J}}^T X_{\hat{J}}\big)^{-1} a_{\hat{J}}\,,$$

then $\sum_{i=1}^n \hat{\sigma}_i^2 \xrightarrow{\Pr} 1$. Thus, part (i) follows provided that the (conditional) Lindeberg condition holds:

$$\sum_{i=1}^n \mathrm{E}\big(|\hat{\alpha}_i \hat{\epsilon}_i^*|^2 I(|\hat{\alpha}_i \hat{\epsilon}_i^*| \geq \delta)|Y\big) \xrightarrow{\Pr} 0 \quad \text{for all} \quad \delta > 0\,. \tag{A.19}$$

Since

$$\hat{\alpha}_i^2 = \hat{s}^{-1} a_{\hat{J}}^T \big(X_{\hat{J}}^T X_{\hat{J}}\big)^{-1} X_{i,\hat{J}} X_{i,\hat{J}}^T \big(X_{\hat{J}}^T X_{\hat{J}}\big)^{-1} a_{\hat{J}}\,,$$

then $\sum_{i=1}^n \hat{\alpha}_i^2 \xrightarrow{\Pr} 1/\sigma^2$. Thus, by (A.19), it is enough to show that

$$\max_{1\leq i\leq n} \mathrm{E}\big(|\hat{\epsilon}^*|^2 I(|\hat{\alpha}_i \hat{\epsilon}^*| \geq \delta)|Y\big) \xrightarrow{\Pr} 0\,, \tag{A.20}$$

where, conditionally on Y, $\hat{\epsilon}^* \sim \mathbb{P}$. Since

$$\max_{1\leq i\leq n} \mathrm{E}\big(|\hat{\epsilon}^*|^2 I(|\hat{\alpha}_i \hat{\epsilon}^*| \geq \delta)|Y\big) \leq \mathrm{E}\big(\hat{\epsilon}^{*2} I(|\hat{\epsilon}^*| \max_{1\leq i\leq n} |\hat{\alpha}_i| \geq \delta)|Y\big) \quad \text{a. s.}\,,$$

then (A.20) holds provided that $\max_{1\le i\le n}|\hat{\alpha}_i| = o_P(1)$. To this end, note that

$$\max_{1\le i\le n}\hat{\alpha}_i^2 = \max_{1\le i\le n}\left\{\frac{a_{\hat{J}}^T(X_{\hat{J}}^TX_{\hat{J}})^{-1}X_{i,\hat{J}}X_{i,\hat{J}}^T(X_{\hat{J}}^TX_{\hat{J}})^{-1}a_{\hat{J}}}{\hat{\sigma}^2a_{\hat{J}}^T(X_{\hat{J}}^TX_{\hat{J}})^{-1}a_{\hat{J}}}\right\}$$
$$\le \frac{\max_{1\le i\le n}\|X_{i,\hat{J}}\|^2}{n\hat{\sigma}^2\rho_1} = O_P\big(p/(n\rho_1)\big) = o_P(1)\,.$$

Lastly, to prove (ii), note first that

$$\hat{s}^{-1/2}a^T(\hat{\beta}^* - \bar{\beta}) = \hat{s}^{-1/2}a^T(\hat{\beta}^* - \hat{\beta}) + \hat{s}^{-1/2}a^T(\hat{\beta} - \bar{\beta})\,.$$

Note further that

$$\hat{s}^{-1/2}a^T(\hat{\beta} - \bar{\beta}) = \sum_{i=1}^n \eta_i\epsilon_i + o_P(1)\,,$$

where

$$\eta_i = s_{J_\beta}^{-1/2}\big(a_{J_\beta}^T(X_{J_\beta}^TX_{J_\beta})^{-1}X_{i,J_\beta} - a^T(X^TX)^{-1}X_i\big)$$

and $s_{J_\beta} = n^{-1}\sigma^2a_{J_\beta}^T\Omega_{J_\beta J_\beta}^{-1}a_{J_\beta}$. Since

$$s = n^{-1}\sigma^2a^T\Omega^{-1}a \le n^{-1}\sigma^2a^Ta\rho_1^{-1} \le n^{-1}\sigma^2a^T\Omega_{J_\beta J_\beta}^{-1}a\rho_1^{-1}\rho_2 = \rho_2\rho_1^{-1}s_{J_\beta}$$

and $\limsup_n \rho_1^{-1}\rho_2 < \infty$, it follows that $\sigma_\eta^2 < \infty$, where

$$\sigma_\eta^2 = \lim_{n\to\infty}\sum_{i=1}^n \eta_i^2\,.$$

Using the same approach as in the proof of Lemma 2.2 and Theorem 2.1, it follows that

$$\hat{s}^{-1/2}a^T(\hat{\beta} - \bar{\beta}) \xrightarrow{\mathrm{d}} N(0, \sigma_\eta^2)\,.$$

By Slutsky's theorem and the continuous mapping theorem on $\mathscr{P}$, the space of distributions on $\mathbb{R}$, it readily follows that (ii) holds, as stated. This completes the proof. □

Acknowledgments

The author wishes to thank the Editor, the Associate Editor, and the referees for helpful comments. He also wishes to thank to Dr Peggy Borum, University of Florida, for sharing the data set.

This research was supported in part by the seed grant "Integrated Interdisciplinary Omics System for Collaborative Biomedical Research" awarded by the Informatics Institute, University of Florida.

References

Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. Springer, New York. MR2247694

Bhatia, R. (1997). *Matrix Analysis*. Springer, New York. MR1477662

Bickel, P. J. and Freedman, D. A. (1983). Bootstrapping regression models with many parameters. In *A Festschrift for Erich L. Lehmann in Honour of his Sixty-fifth Birtday* (P. J. Bickel, K. A. Doksum and J. L. Hodges, eds.) 28–48. Wadsworth, Belmont, CA. MR0689736

Bickel, P. J. and Levina, E. (2008). Covariance Regularization by thresholding. *The Annals of Statistics* **36** 2577–2604. MR2485008

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics* **34** 559–583. MR2281878

Chatterjee, A., Gupta, S. and Lahiri, S. N. (2015). On the residual empirical process based on the ALASSO in high dimensions and its functional oracle property. *Journal of Econometrics* **186** 317–324. MR3343789

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping Lasso estimators. *Journal of the American Statistical Association* **106** 608-625. MR2847974

Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the adaptive Lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics* **41** 1232-1259. MR3113809

Chow, Y. S. and Teicher, H. (1997). *Probability Theory: independence, interchangeability, martingales*. Springer, New York. MR1476912

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge. MR1478673

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425-455. MR1311089

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *The Annals of Statistics* **24** 508-539. MR1394974

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, Cambridge. MR1932358

Efron, B. (1979). Bootstrap methods: another look at Jackknife. *The Annals of Statistics* **7** 1–26. MR0515681

El Ghaoui, L., Viallon, V. and Rabbani, T. (2012). Safe feature elimination for the lasso and sparse supervised learning problems. *Pacific Journal of Optimization* **8(4)** 667–698. MR3026449

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* **36** 2717–2756. MR2485011

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. MR1946581

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B* **96** 1348–1360. MR2530322

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35** 109–148.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics* **9** 1218–1228. MR0630104

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1-22.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer, New York. MR2722294

Huang, J., Horowitz, J. L. and Ma, S. (2008). Asymptotic properties of Bridge estimators in sparse high-dimensional regression models. *Annals of Statistics* **36** 587–613. MR2396808

Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive LASSO for sparse high-dimensional regression models. *Statistica Sinica* **18** 1603–1618. MR2469326

Huber, P. (1973). Robust regression: asymptotics, conjectures, and Monte Carlo. *The Annals of Statistics* **1** 799–821. MR0356373

Khuri, A. I. (2010). *Linear Model Methodology*. Chapman & Hall/CRC, Boca Raton, Florida. MR3052842

Knight, K. and Fu, W. (2000). Asymptotics for LASSO-type estimators. *The Annals of Statistics* **5** 1356–1378. MR1805787

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics* **17** 382–400. MR0981457

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics* **21** 255–285. MR1212176

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; I. Consistency. *The Annals of Statistics* **12** 1298-1309. MR0760690

Portnoy, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; II. Normal approximation. *The Annals of Statistics* **13** 1403-1417. MR0811499

Tibshirani, R. (1996). Regression shrinkage and selection estimator via the Lasso. *Journal of the Royal Statistical Society: Series B* **58** 267–288. MR1379242

van de Geer, S., Bühlmann, P. and Zhou, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics* **5** 688–749. MR2820636

van der Laan, M. and Bryan, J. (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics* **2** 445–461.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Springer-Verlag, New York. MR1652247

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares pro-

jection for screening variables. *Journal of the Royal Statistical Society. Series B.* **78** 589–611.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics* **37** 2178–2201. MR2543689

Zelen, M. and Severo, N. (1972). Probability Functions. In *Handbook of Mathematical Functions with Formula, Graphs, and Mathematical Tables*, (M. Abramowitz and I. Stegun, eds.). *Applied Mathematics Series* **55** 925–997. National Bureau of Standards, Washington. MR0208798

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469