CONFIDENCE INTERVALS FOR LOW-DIMENSIONAL PARAMETERS IN HIGH-DIMENSIONAL LINEAR MODELS

CUN-HUI ZHANG AND STEPHANIE S. ZHANG

ABSTRACT. The purpose of this paper is to propose methodologies for statistical inference of low-dimensional parameters with high-dimensional data. We focus on constructing confidence intervals for individual coefficients and linear combinations of several of them in a linear regression model, although our ideas are applicable in a much broader context. The theoretical results presented here provide sufficient conditions for the asymptotic normality of the proposed estimators along with a consistent estimator for their finite-dimensional covariance matrices. These sufficient conditions allow the number of variables to far exceed the sample size. The simulation results presented here demonstrate the accuracy of the coverage probability of the proposed confidence intervals, strongly supporting the theoretical results.

Key words: Confidence interval, p-value, statistical inference, linear regression model, high dimension.

1. INTRODUCTION

High-dimensional data is an intense area of research in statistics and machine learning, due to the rapid development of information technologies and their applications in scientific experiments and everyday life. Numerous large, complex datasets have been collected and are waiting to be analyzed; meanwhile, an enormous effort has been mounted in order to meet this challenge by researchers and practitioners in statistics, computer science, and other disciplines. A great number of statistical methods, algorithms, and theories have been developed for the prediction and classification of future outcomes, the estimation of high-dimensional objects, and the selection of important variables or features for further scientific experiments and engineering applications. However, statistical inference with high-dimensional data is still largely untouched, due to the complexity of the sampling distributions of existing estimators. This is particularly the case in the context of the so called large-p-smaller-n problem, where the dimension of the data p is greater than the sample size n.

Regularized linear regression is one of the best understood statistical problems in high-dimensional data. Important work has been done in formulation of problems, development of methodologies and algorithms, and theoretical understanding of their performance under sparsity assumptions on the regression coefficients. This includes ℓ_1 regularized methods [Tib96, CDS01, GR04, Gre06, MB06, Tro06, ZY06, CT07, ZH08, BRT09, Kol09, MY09, vdGB09, Wai09b, Zha09, YZ10, KLT11, SZ11], nonconvex penalized methods [FF93, FL01, FP04, KCO08, Zha10, ZZ11], greedy methods [Zha11a], adaptive methods [Zou06, HMZ08, ZL08, Zha11b, ZZ11], screening methods [FL08], and more. For further discussion, we refer to related sections in [BvdG11] and recent reviews in [FL10, ZZ11].

Among existing results, variable selection consistency is most relevant to statistical inference. An estimator is variable selection consistent if it selects the oracle model composed of exactly the set of variables with nonzero regression coefficients. In the large-p-smaller-n setting, variable selection consistency has been established under incoherence and other ℓ_∞-type conditions on the design matrix for the Lasso [MB06, Tro06, ZY06, Wai09b], and under sparse eigenvalue or ℓ_2-type conditions for nonconvex methods [FP04, Zha10, Zha11a, Zha11b, ZZ11]. Another approach in variable selection with high-dimensional data involves subsampling or randomization, including notably the stability selection method proposed in [MB10]. Since the oracle model is typically assumed to be of smaller order in dimension than the sample size n in selection consistency theory, consistent variable selection allows a great reduction of the complexity of the analysis from a large-p-smaller-n problem to one involving the oracle set of variables only. Consequently, taking the least squares estimator on the selected set of variables if necessary, statistical inference can be justified in the smaller oracle model.

However, statistical inference based on selection consistency theory typically requires a uniform signal strength condition that all nonzero regression coefficients be greater in magnitude than an inflated noise level to take model uncertainty into account. This inflated noise level can be written as $C\sigma\sqrt{(2/n)\log p}$, where σ is the noise level with each response. Based on the sharpest existing results, $C \geq 1/2$ is required for variable selection consistency with a general standardized design matrix [Wai09a, Zha10]. This uniform signal strength condition is, unfortunately, seldom supported by either the data or the underlying science in applications when the presence of weak signals cannot be ruled out. Without this uniform signal strength assumption, consistent estimation of the distribution of the least squares estimator after model selection is impossible [LP06]. Conservative statistical inference after model selection or classification has been considered in [BBZ10, LM11]. However, such conservative methods may not yield sufficiently accurate confidence regions or p-values for common applications with a large number of variables.

We propose a low-dimensional projection (LDP) approach to constructing confidence intervals for regression coefficients without assuming the uniform signal strength condition. We provide theoretical justifications for the use of the proposed confidence interval for a preconceived regression coefficient or a contrast depending on a small number of regression coefficients. We believe that in the presence of potentially many nonzero coefficients of small or moderate magnitude, construction of a confidence interval for such a preconceived parameter is an important problem in and of itself and was open before our paper [LP06], but the proposed method is not limited to this application.

Our theoretical work also justifies the use of LDP confidence intervals simultaneously with multiplicity adjustment. In the absence of a preconceived parameter of interest, the proposed simultaneous confidence intervals provide more information about the unknown regression coefficients than variable selection, but this is not the main point.

The most important difference between the proposed LDP and existing variable selection approaches concerns the requirement known as the uniform signal strength condition. As we have mentioned earlier, variable selection consistency requires all nonzero regression coefficients be greater than $C\sigma\sqrt{(2/n)\log p}$, with $C \geq 1/2$ at the least. This is a necessity for the simultaneous correct selection of *all zero or nonzero* coefficients. If this criterion is the goal,

we can not do better than technical improvements over existing methods. However, a main complaint about the variable selection approach is the practicality of the uniform signal strength condition, and the crucial difference between the two approaches is precisely in the case where the condition fails to hold. Without the condition, neither large nor zero coefficients are guaranteed to be correctly selected by existing variable selection methods in the presence of potentially many nonzero coefficients below the radar screen, but the proposed method can. The power of the proposed method is small for testing small nonzero coefficients, but this is unavoidable and does not affect the correct selection of other variables. In this sense, the proposed confidence intervals decompose the variable selection problem into multiple marginal testing problems for individual coefficients as Gaussian means.

2. Methodology

We develop methodologies and algorithms for the construction of confidence intervals for the individual regression coefficients and their linear combinations in the linear model

$$(1) \qquad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}),$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is a response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ is a design matrix with columns $\boldsymbol{x}_j$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown regression coefficients. When $\mathrm{rank}(\boldsymbol{X}) < p$, $\boldsymbol{\beta}$ is unique under proper conditions on the sparsity of $\boldsymbol{\beta}$ and regularity of $\boldsymbol{X}$, but not in general. To simplify the discussion, we standardize the design to $\|\boldsymbol{x}_j\|_2^2 = n$. The design matrix $\boldsymbol{X}$ is assumed to be deterministic throughout the paper, except in Subsection 3.4.

The following notation will be used. For real numbers x and y, $x \wedge y = \min(x, y)$, $x \vee y = \max(x, y)$, $x_+ = x \vee 0$, and $x_- = (-x)_+$. For vectors $\boldsymbol{v} = (v_1, \ldots, v_m)$ of any dimension, $\mathrm{supp}(\boldsymbol{v}) = \{j : v_j \neq 0\}$, $\|\boldsymbol{v}\|_0 = |\mathrm{supp}(\boldsymbol{v})| = \#\{j : v_j \neq 0\}$, and $\|\boldsymbol{v}\|_q = \{\sum_j |v_j|^q\}^{1/q}$, with the usual extension to $q = \infty$. For $A \subset \{1, \ldots, p\}$, $\boldsymbol{v}_A = (v_j, j \in A)^T$ and $\boldsymbol{X}_A = (\boldsymbol{x}_k, k \in A)$, including $A = -j = \{1, \ldots, p\} \setminus \{j\}$.

2.1. **Bias corrected linear estimators.** In the classical theory of linear models, the least squares estimator of an estimable regression coefficient β_j can be written as

$$(2) \qquad \widehat{\beta}_j^{(lse)} := (\boldsymbol{x}_j^{\perp})^T \boldsymbol{y} / (\boldsymbol{x}_j^{\perp})^T \boldsymbol{x}_j,$$

where $\boldsymbol{x}_j^{\perp}$ is the projection of $\boldsymbol{x}_j$ to the orthogonal complement of the column space of $\boldsymbol{X}_{-j} = (\boldsymbol{x}_k, k \neq j)$. Since this is equivalent to solving the equations $(\boldsymbol{x}_j^{\perp})^T(\boldsymbol{y} - \beta_j \boldsymbol{x}_j) = (\boldsymbol{x}_j^{\perp})^T \boldsymbol{x}_k = 0 \ \forall \ k \neq j$ in the score system $\boldsymbol{v} \to (\boldsymbol{x}_j^{\perp})^T \boldsymbol{v}$, $\boldsymbol{x}_j^{\perp}$ can be viewed as the score vector for the least squares estimation of β_j. For estimable β_j and β_k,

$$(3) \qquad \mathrm{Cov}(\widehat{\beta}_j^{(lse)}, \widehat{\beta}_k^{(lse)}) = \sigma^2 (\boldsymbol{x}_j^{\perp})^T \boldsymbol{x}_k^{\perp} / (\|\boldsymbol{x}_j^{\perp}\|_2^2 \|\boldsymbol{x}_k^{\perp}\|_2^2).$$

In the high-dimensional case $p > n$, $\mathrm{rank}(\boldsymbol{X}_{-j}) = n$ for all j when $\boldsymbol{X}$ is in general position. Consequently, $\boldsymbol{x}_j^{\perp} = 0$ and (2) is undefined. However, it may still be interesting to preserve certain properties of the least squares estimator. This can be done by retaining the main equation $\boldsymbol{z}_j^T(\boldsymbol{y} - \beta_j \boldsymbol{x}_j) = 0$ in a score system $\boldsymbol{z}_j : \boldsymbol{v} \to \boldsymbol{z}_j^T \boldsymbol{v}$ and relaxing the constraint $\boldsymbol{z}_j^T \boldsymbol{x}_k = 0$ for $k \neq j$, resulting in a linear estimator. One advantage of (2) is the explicit formula (3) for the covariance structure. This feature holds for all linear estimators of $\boldsymbol{\beta}$.

For any score vector $\boldsymbol{z}_j$ not orthogonal to $\boldsymbol{x}_j$, the corresponding univariate linear regression estimator satisfies

$$\widehat{\beta}_j^{(lin)} = \frac{\boldsymbol{z}_j^T \boldsymbol{y}}{\boldsymbol{z}_j^T \boldsymbol{x}_j} = \beta_j + \frac{\boldsymbol{z}_j^T \boldsymbol{\varepsilon}}{\boldsymbol{z}_j^T \boldsymbol{x}_j} + \sum_{k \neq j} \frac{\boldsymbol{z}_j^T \boldsymbol{x}_k \beta_k}{\boldsymbol{z}_j^T \boldsymbol{x}_j}$$

with a similar covariance structure to (3). A problem with this linear estimator is its bias. For every $k \neq j$ with $\boldsymbol{z}_j^T \boldsymbol{x}_k \neq 0$, the contribution of β_k to the bias is linear in β_k. Thus, under the assumption of $\|\boldsymbol{\beta}\|_0 \leq 2$, which is very strong, the bias of $\widehat{\beta}_j^{(lin)}$ is still unbounded when $\boldsymbol{z}_j^T \boldsymbol{x}_k \neq 0$ for at least one $k \neq j$. We note that for $\text{rank}(\boldsymbol{X}_{-j}) = n$, it is impossible to have $\boldsymbol{z}_j \neq 0$ and $\boldsymbol{z}_j^T \boldsymbol{x}_k = 0$ for all $k \neq j$, so that bias is unavoidable. Still, this analysis of the linear estimator suggests a bias correction with a nonlinear initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$:

(4) $$\widehat{\beta}_j = \widehat{\beta}_j^{(lin)} - \sum_{k \neq j} \frac{\boldsymbol{z}_j^T \boldsymbol{x}_k \widehat{\beta}_k^{(init)}}{\boldsymbol{z}_j^T \boldsymbol{x}_j} = \frac{\boldsymbol{z}_j^T \boldsymbol{y}}{\boldsymbol{z}_j^T \boldsymbol{x}_j} - \sum_{k \neq j} \frac{\boldsymbol{z}_j^T \boldsymbol{x}_k \widehat{\beta}_k^{(init)}}{\boldsymbol{z}_j^T \boldsymbol{x}_j}.$$

One may also interpret (4) as a one-step self bias correction from the initial estimator and write

$$\widehat{\beta}_j := \widehat{\beta}_j^{(init)} + \frac{\boldsymbol{z}_j^T \{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)}\}}{\boldsymbol{z}_j^T \boldsymbol{x}_j}.$$

The estimation error of (4) can be decomposed as a sum of the noise and the approximation errors:

(5) $$\widehat{\beta}_j - \beta_j = \frac{\boldsymbol{z}_j^T \boldsymbol{\varepsilon}}{\boldsymbol{z}_j^T \boldsymbol{x}_j} + \frac{1}{\boldsymbol{z}_j^T \boldsymbol{x}_j} \sum_{k \neq j} \boldsymbol{z}_j^T \boldsymbol{x}_k (\beta_k - \widehat{\beta}_k^{(init)}).$$

We require that $\boldsymbol{z}_j$ be a vector depending on $\boldsymbol{X}$ only, so that $\boldsymbol{z}_j^T \boldsymbol{\varepsilon}/\|\boldsymbol{z}_j\|_2 \sim N(0, \sigma^2)$. A full description of (4) still requires the specification of the score vector $\boldsymbol{z}_j$ and the initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$. These choices will be discussed in the following two subsections.

2.2. **Low-dimensional projections.** We propose to use as $\boldsymbol{z}_j$ a relaxed orthogonalization of $\boldsymbol{x}_j$ against other design vectors. Recall that $\boldsymbol{z}_j$ aims to play the role of $\boldsymbol{x}_j^\perp$, the projection of $\boldsymbol{x}_j$ to the orthogonal complement of the column space of $\boldsymbol{X}_{-j} = (\boldsymbol{x}_k, k \neq j)$. In the trivial case where $\|\boldsymbol{x}_j^\perp\|_2$ is not too small, we may simply take $\boldsymbol{z}_j = \boldsymbol{x}_j^\perp$. In addition to the case of $\text{rank}(\boldsymbol{X}_{-j}) = n$, where $\boldsymbol{x}_j^\perp = 0$, a relaxed projection could be useful when $\|\boldsymbol{x}_j^\perp\|_2$ is positive but small. Since a relaxed projection $\boldsymbol{z}_j$ is used and the estimator (4) is a bias-corrected projection of $\boldsymbol{y}$ to the direction of $\boldsymbol{z}_j$, hereafter we call (4) the low-dimensional projection estimator (LDPE) for easy reference.

A proper relaxed projection $\boldsymbol{z}_j$ should control both the noise and approximation error terms in (5), given suitable conditions on $\{\boldsymbol{X}, \boldsymbol{\beta}\}$ and an initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$. By (5), the approximation error of (4) can be bounded by

(6) $$\left| \sum_{k \neq j} \boldsymbol{z}_j^T \boldsymbol{x}_k (\beta_k - \widehat{\beta}_k^{(init)}) \right| \leq \left(\max_{k \neq j} \left| \boldsymbol{z}_j^T \boldsymbol{x}_k \right| \right) \|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1.$$

This conservative bound is conveniently expressed as the product of a known function of z_j and the initial estimation error independent of j. For score vectors z_j, define

$$(7) \qquad \eta_j = \max_{k \neq j} \left| z_j^T x_k \right| / \|z_j\|_2, \quad \tau_j = \|z_j\|_2 / |z_j^T x_j|.$$

We refer to η_j as the bias factor since $\eta_j \|\widehat{\beta}^{(init)} - \beta\|_1$ controls the approximation error in (6) relative to the length of the score vector. We refer to τ_j as the noise factor, since $\tau_j \sigma$ is the standard deviation of the noise component in (5). Since $z_j^T \varepsilon \sim N(0, \sigma^2 \|z_j\|_2^2)$, (5) yields

$$(8) \qquad \eta_j \|\widehat{\beta}^{(init)} - \beta\|_1 / \sigma = o(1) \;\Rightarrow\; \tau_j^{-1} (\widehat{\beta}_j - \beta_j) \approx N(0, \sigma^2).$$

Thus, we would like to pick a z_j with a small η_j for the asymptotic normality and a small τ_j for estimation efficiency. Confidence intervals for β_j and linear functionals of them can be constructed provided the condition in (8) and a consistent estimator of σ.

We still need a suitable z_j, a relaxed orthogonalization of x_j against other design vectors. When the unrelaxed $x_j^{\perp}$ is nonzero, it can be viewed as the residual of the least squares fit of x_j on X_{-j}. A familiar relaxation of the least squares method is to add an ℓ_1 penalty. This leads to the choice of z_j as the residual of the Lasso. Let $\widehat{\gamma}_j$ be the vector of coefficients from the Lasso regression of x_j on X_{-j}. The Lasso-generated score is

$$(9) \qquad z_j = x_j - X_{-j}\widehat{\gamma}_j, \; \widehat{\gamma}_j = \arg\min_b \left\{ \frac{\|x_j - X_{-j}b\|_2^2}{2n} + \lambda_j \|b\|_1 \right\}.$$

It follows from the Karush-Kuhn-Tucker conditions for (9) that $|x_k^T z_j / n| \leq \lambda_j$ for all $k \neq j$, so that (7) holds with $\eta_j \leq n\lambda_j / \|z_j\|_2$. This gives many choices of z_j with different $\{\eta_j, \tau_j\}$. Explicit choices of such a z_j, or equivalently a λ_j, are described in the next subsection. A rationale for the use of a common penalty level λ_j for all components of b in (9) is the standardization of all design vectors. In an alternative in Subsection 2.3 called the restricted LDPE (R-LDPE), the penalty is set to zero for certain components of b in (9).

2.3. **Specific implementations.** We have to pick $\widehat{\beta}^{(init)}$, $\widehat{\sigma}$, and the λ_j in (9). Since consistent estimation of σ and fully automatic choices of λ_j are needed, we use methods based on the scaled Lasso and the least squares estimator in the model selected by the scaled Lasso (scaled Lasso-LSE).

The scaled Lasso [Ant10, SZ10, SZ11] is a joint convex minimization method given by

$$(10) \qquad \{\widehat{\beta}^{(init)}, \widehat{\sigma}\} = \arg\min_{b,\sigma} \left\{ \frac{\|y - Xb\|_2^2}{2\sigma n} + \frac{\sigma}{2} + \lambda_0 \|b\|_1 \right\},$$

with a preassigned penalty level λ_0. This automatically provides an estimate of the noise level in addition to the initial estimator of β. We use $\lambda_0 = \lambda_{univ} = \sqrt{(2/n)\log p}$ in our simulation study. Existing error bounds for the estimation of both β and σ require $\lambda_0 = A\sqrt{(2/n)\log(p/\epsilon)}$ with certain $A > 1$ and $0 < \epsilon \leq 1$ [SZ11].

The estimator (10) has appeared in the literature in different forms. The joint minimization formulation was given in [Ant10], and an equivalent algorithm in [SZ10]. If the minimum over b is taken first in (10), the resulting $\widehat{\sigma}$ appeared earlier in [Zha10]. The square root Lasso [BCW11] gives the same $\beta^{(init)}$ with a different formulation, but not joint

estimation. The formulations in [Zha10] and [SZ10] allow concave penalties and a degrees of freedom adjustment.

The Lasso is biased, as is the scaled Lasso. Let $\widehat{S}^{(init)}$ be the set of nonzero estimated coefficients by the scaled Lasso. When $\widehat{S}^{(init)}$ catches most large $|\beta_j|$, the bias of (10) can be reduced by the least squares estimator in the selected model $\widehat{S}^{(init)}$:

$$(11) \qquad \{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\} = \arg\min_{\boldsymbol{b}, \sigma} \left\{ \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2}{2\sigma(n - |\widehat{S}^{(init)}|)} + \frac{\sigma}{2} : b_j = 0 \ \forall \ j \notin \widehat{S}^{(init)} \right\}.$$

This defines the scaled Lasso-LSE. We use the same notation in (10) and (11) since they both give initial estimates for the LDPE (4) and a noise level estimator for statistical inference based on the LDPE. The specific estimators will henceforth be referred to by their names or as (10) and (11). The scaled Lasso-LSE enjoys similar analytical error bounds as the scaled Lasso and outperformed scaled Lasso in a simulation study [SZ11].

The scaled Lasso can be also used to determine λ_j for the $\boldsymbol{z}_j$ in (9). However, the penalty level for the scaled Lasso, set to guarantee performance bounds for the estimation of regression coefficients and noise level, may not be the best for controlling the bias and the standard error of the LDPE. By (7) and (8), it suffices to find a $\boldsymbol{z}_j$ with small bias factor η_j and small noise factor τ_j. These quantities are always available. This is quite different from the estimation of $\{\boldsymbol{\beta}, \sigma\}$ in (10) where the effect of over-fitting is unobservable.

We choose λ_j by tracking η_j and τ_j in the Lasso path. One of our ideas is to reduce η_j by allowing some over fitting of $\boldsymbol{x}_j$ as long as τ_j is reasonably small. Ideally, this slightly more conservative approach will lead to confidence intervals with more accurate coverage probability. Along the Lasso path for regressing $\boldsymbol{x}_j$ against $\boldsymbol{X}_{-j}$, let

$$(12) \qquad \widehat{\boldsymbol{\gamma}}_j(\lambda) = \arg\min_{\boldsymbol{b}} \left\{ \|\boldsymbol{x}_j - \boldsymbol{X}_{-j}\boldsymbol{b}\|_2^2/(2n) + \lambda\|\boldsymbol{b}\|_1 \right\},$$

$$\boldsymbol{z}_j(\lambda) = \boldsymbol{x}_j - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}_j(\lambda),$$

$$\eta_j(\lambda) = \max_{k \neq j} |\boldsymbol{x}_k^T \boldsymbol{z}_j(\lambda)|/\|\boldsymbol{z}_j(\lambda)\|_2,$$

$$\tau_j(\lambda) = \|\boldsymbol{z}_j(\lambda)\|_2/|\boldsymbol{x}_j^T \boldsymbol{z}_j(\lambda)|,$$

be the coefficient estimator $\widehat{\boldsymbol{\gamma}}_j$, residual $\boldsymbol{z}_j$, the bias factor η_j, and the noise factor τ_j, as functions of λ. We compute $\boldsymbol{z}_j$ according to the algorithm in Table 1.

In Table 1, Step 1 finds a feasible upper bound η_j^* for the bias factor and the corresponding noise factor τ_j^*. Step 2 seeks $\boldsymbol{z}_j = \boldsymbol{z}_j(\lambda_j)$ in (12) at a certain level $\lambda = \lambda_j$ with a smaller $\eta_j = \eta_j(\lambda_j)$, subject to the constraint $\tau(\lambda_j) \leq (1 + \kappa_0)\tau_j^*$ on the noise factor. It follows from Proposition 1 (i) below that $\eta_j(\lambda)$ is non-decreasing in λ, so that searching for the smallest $\eta_j(\lambda)$ is equivalent to searching for the smallest λ in Step 2, subject to the constraint.

In the search for $\boldsymbol{z}_j$ with smaller η_j in Step 2, the relative increment in the noise factor τ_j is no greater than κ_0. This corresponds to a loss of relative efficiency no greater than $1 - 1/(1 + \kappa_0)^2$ for the estimation of β_j. In our simulation experiments, $\kappa_0 = 1/4$ provides a suitable choice, compared with $\kappa_0 = 0$ and $\kappa_0 = 1/2$. We would like to emphasize here that the score vectors $\boldsymbol{z}_j$ computed by the algorithm in Table 1 are completely determined by the design $\boldsymbol{X}$.

TABLE 1. Computation of $\boldsymbol{z}_j$ from the Lasso (12)

Input:	an upper bound η_j^* for the bias factor, with default value $\eta_j^* = \sqrt{2\log p}$, tuning parameters $\kappa_0 \in [0,1]$ and $\kappa_1 \in (0,1]$;
Step 1:	(verify/adjust η_j^* and compute the corresponding noise factor τ_j^*) If $\eta_j(\lambda) > \eta_j^*$ for all $\lambda > 0$, $\eta_j^* \leftarrow (1+\kappa_1)\inf_{\lambda>0}\eta_j(\lambda)$; $\lambda \leftarrow \max\{\lambda : \eta_j(\lambda) \leq \eta_j^*\}$, $\eta_j^* \leftarrow \eta_j(\lambda)$, $\tau_j^* \leftarrow \tau_j(\lambda)$;
Step 2:	(further reduction of the bias factor) $\lambda_j \leftarrow \min\{\lambda : \tau_j(\lambda) \leq (1+\kappa_0)\tau_j^*\}$;
Output:	λ_j, $\boldsymbol{z}_j \leftarrow \boldsymbol{z}_j(\lambda_j)$, $\tau_j \leftarrow \tau_j(\lambda_j)$, $\eta_j \leftarrow \eta_j(\lambda_j)$

A main objective of the algorithm in Table 1 is to find a $\boldsymbol{z}_j$ with a bias factor $\eta_j \leq C\sqrt{\log p}$ to allow a uniform bias bound via (6), (7), and (8). It is ideal if $C = \sqrt{2}$ is attainable, but a reasonably small C also works with the argument. When $\eta_j^* = \sqrt{2\log p}$ is not feasible, Step 1 finds a larger upper bound η_j^* for the bias factor. When $\sup_\lambda \eta_j(\lambda) < \sqrt{2\log p}$, $\eta_j^* < \sqrt{2\log p}$ after the adjustment in Step 1, resulting in an even smaller η_j in Step 2. This does happen in our simulation experiments. The choice of the target upper bound $\sqrt{2\log p}$ for η_j is based on its feasibility as well as the sufficiency of $\eta_j \leq \sqrt{2\log p}$ for the verification of the condition in (8) based on the existing ℓ_1 error bounds for the estimation of $\boldsymbol{\beta}$. Proposition 1 below asserts that $\max_{j\leq p}\eta_j^* \leq C\sqrt{\log p}$ is feasible when $\boldsymbol{X}$ allows an optimal rate of sparse recovery. In our simulation experiments, we are able to use $\eta_j^* \leq \sqrt{2\log p}$ in all replications and settings for all variables, a total of more than 1 million instances. Moreover, the theoretical results in Subsection 3.4 prove that for the η_j^* in Table 1, $\max_{j\leq p}\eta_j^* \leq 3\sqrt{\log p}$ with high probability under proper conditions on random $\boldsymbol{X}$. It is worthwhile to note that both η_j and τ_j are computed, and control of $\max_k \eta_k$ is not required for the LDPE to apply to variables with small η_j.

We have also experimented with an LDPE using a restricted Lasso relaxation for $\boldsymbol{z}_j$. This R-LDPE (restricted LDPE) can be viewed as a special case of a more general weighted low dimensional projection with different levels of relaxation for different variables $\boldsymbol{x}_k$ according to their correlation to $\boldsymbol{x}_j$. Although we have used (6) to bound the bias, the summands with larger absolute correlation $|\boldsymbol{x}_j^T\boldsymbol{x}_k/n|$ are likely to have a greater contribution to the bias due to the initial estimation error $|\widehat{\beta}_k^{(init)} - \beta_k|$. A remedy for this phenomenon is to force smaller $|\boldsymbol{z}_j^T\boldsymbol{x}_k/n|$ for large $|\boldsymbol{x}_j^T\boldsymbol{x}_k/n|$ with a weighted relaxation. For the Lasso (9), this weighted relaxation can be written as

$$\boldsymbol{z}_j = \boldsymbol{x}_j - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}_j, \ \ \widehat{\boldsymbol{\gamma}}_j = \arg\min_{\boldsymbol{b}}\left\{\frac{\|\boldsymbol{x}_j - \boldsymbol{X}_{-j}\boldsymbol{b}\|_2^2}{2n} + \lambda_j\sum_{k\neq j}w_k|b_k|\right\},$$

with w_k being a decreasing function of the absolute correlation $|\boldsymbol{x}_j^T\boldsymbol{x}_k/n|$. For the R-LDPE, we simply set $w_k = 0$ for large $|\boldsymbol{x}_j^T\boldsymbol{x}_k/n|$ and $w_k = 1$ for other k.

Here is an implementation of the R-LDPE. Let $K_{j,m}$ be the index set of the m largest $|\boldsymbol{x}_j^T\boldsymbol{x}_k|$ with $k \neq j$ and $\boldsymbol{P}_{j,m}$ be the orthogonal projection to the linear span of $\{\boldsymbol{x}_k, k \in$

$K_{j,m}\}$. Let $\boldsymbol{z}_j = f(\boldsymbol{x}_j, \boldsymbol{X}_{-j})$ denotes the algorithm in Table 1 as a mapping $(\boldsymbol{x}_j, \boldsymbol{X}_{-j}) \to \boldsymbol{z}_j$. We compute the R-LDPE by taking the projection of all design vectors to the orthogonal complement of $\{\boldsymbol{x}_k, k \in K_{j,m}\}$ before the application of the procedure in (12) and Table 1. The resulting score vector can be written as

$$(13) \qquad \boldsymbol{z}_j = f(\boldsymbol{P}_{j,m}^{\perp}\boldsymbol{x}_j, \boldsymbol{P}_{j,m}^{\perp}\boldsymbol{X}_{-j}).$$

We use the rest of this subsection to present some useful properties of the Lasso path (12) for the implementation of the algorithm in Table 1 and some sufficient conditions for the uniform bound $\max_j \eta_j^* \le C\sqrt{\log p}$ for the bias factors in the output. Let

$$(14) \qquad \widehat{\sigma}_j(\lambda) = \arg\min_{\sigma} \min_{\boldsymbol{b}} \left\{ \frac{\|\boldsymbol{x}_j - \boldsymbol{X}_{-j}\boldsymbol{b}\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda\|\boldsymbol{b}\|_1 \right\}$$

be the solution of $\widehat{\sigma}$ in (10) with $\{\boldsymbol{X}, \boldsymbol{y}, \lambda_0\}$ replaced by $\{\boldsymbol{X}_{-j}, \boldsymbol{x}_j, \lambda\}$.

Proposition 1. *(i) In the Lasso path (12), $\|\boldsymbol{z}_j(\lambda)\|_2$, $\eta_j(\lambda)$, and $\widehat{\sigma}_j(\lambda)$ are nondecreasing functions of λ, and $\tau_j(\lambda) \le 1/\|\boldsymbol{z}_j(\lambda)\|_2$. Moreover, $\widehat{\boldsymbol{\gamma}}_j(\lambda) \ne 0$ implies $\eta_j(\lambda) = \lambda n/\|\boldsymbol{z}_j(\lambda)\|_2$. (ii) Let $\lambda_{univ} = \sqrt{(2/n)\log p}$. Then,*

$$(15) \qquad \widehat{\sigma}_j(C\lambda_{univ}) > 0 \ \textit{iff} \ \{\lambda > 0 : \eta_j(\lambda) \le C\sqrt{2\log p}\} \ne \emptyset,$$

and in this case, the algorithm in Table 1 provides

$$(16) \qquad \eta_j \le \eta_j^* \le (1 + \kappa_1 I_{\{C>1\}})(1 \vee C)\sqrt{2\log p}, \quad \tau_j \le n^{-1/2}(1 + \kappa_0)/\widehat{\sigma}_j(C\lambda_{univ}).$$

Moreover, when $\boldsymbol{z}_j(0) = \boldsymbol{x}_j^{\perp} = 0$, $\eta_j(0+)\inf\{\|\boldsymbol{\gamma}_j\|_1 : \boldsymbol{X}_{-j}\boldsymbol{\gamma}_j = \boldsymbol{x}_j\} = \sqrt{n}$.
(iii) Let $0 < a_0 < 1 \le C_0 < \infty$. Suppose that for $s = a_0 n/\log p$

$$\inf_{\boldsymbol{\delta}} \sup_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{\delta}(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{\beta}\|_2^2 : \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}, \sum_{j=1}^p \min(|\beta_j|/\lambda_{univ}, 1) \le s + 1 \right\} \le 2C_0 s(\log p)/n.$$

Then, $\max_{j \le p} \eta_j^* \le (1 + \kappa_1)\sqrt{(4C_0/a_0)\log p}$ for the algorithm in Table 1.

The monotonicity of $\|\boldsymbol{z}_j(\lambda)\|_2$ and $\eta_j(\lambda)$ in Proposition 1 (i) provides directions of search in both steps of the algorithm in Table 1.

Proposition 1 (ii) provides mild conditions for controlling the bias factor at $\eta_j \le \eta_j^* \le C\sqrt{2\log p}$ and the standard error to the order $\tau_j = O(n^{-1/2})$. It asserts that $\eta_j^* \le \sqrt{2\log p}$ when the scaled Lasso (14) with $\lambda = \lambda_{univ}$ yields a positive $\widehat{\sigma}_j$. In the completely collinear case where $\boldsymbol{x}_k = \boldsymbol{x}_j$ for some $k \ne j$, $\inf\{\|\boldsymbol{\gamma}_j\|_1 : \boldsymbol{x}_j = \boldsymbol{X}_{-j}\boldsymbol{\gamma}_j\} = 1$ gives the largest $\eta_j = \sqrt{n}$. This suggests a connection between the minimum feasible η_j and certain "near estimability" of β_j, with small η_j for nearly estimable β_j. It also provides a connection between the smallest $\eta_j(\lambda)$ and an ℓ_1 recovery problem, leading to Proposition 1 (iii).

Proposition 1 (iii) asserts that the validity of the upper bound $\max_j \eta_j^* \le C\sqrt{\log p}$ for the bias factor is a consequence of the existence of an estimator $\boldsymbol{\delta}$ with the ℓ_2 recovery bound in the noiseless case of $\boldsymbol{\varepsilon} = 0$. In the more difficult case of $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I})$, ℓ_2 error bounds of the same type have been proven under sparse eigenvalue conditions on $\boldsymbol{X}$, and by Proposition 1 (iii), $\max_j \eta_j^* \le C\sqrt{\log p}$ is also a consequence of such conditions.

2.4. **Confidence intervals.** In Section 3, we will provide sufficient conditions on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ under which the approximation error in (5) is of smaller order than the standard deviation of the noise component. We construct approximate confidence intervals for such configurations of $\{\boldsymbol{X}, \boldsymbol{\beta}\}$ as follows.

The covariance of the noise component in (5) is proportional to

$$(17) \qquad \boldsymbol{V} = (V_{jk})_{p \times p}, \quad \text{where} \quad V_{jk} = \frac{\boldsymbol{z}_j^T \boldsymbol{z}_k}{|\boldsymbol{z}_j^T \boldsymbol{x}_j||\boldsymbol{z}_k^T \boldsymbol{x}_k|} = \sigma^{-2} \text{Cov}\Big(\frac{\boldsymbol{z}_j^T \boldsymbol{\varepsilon}}{\boldsymbol{z}_j^T \boldsymbol{x}_j}, \frac{\boldsymbol{z}_k^T \boldsymbol{\varepsilon}}{\boldsymbol{z}_k^T \boldsymbol{x}_k} \Big).$$

Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \dots, \widehat{\beta}_p)^T$ be the vector of LDPEs $\widehat{\beta}_j$ in (4). For sparse vectors $\boldsymbol{a}$ with bounded $\|\boldsymbol{a}\|_0$, e.g. $\|\boldsymbol{a}\|_0 = 2$ for a contrast between two regression coefficients, an approximate $(1 - \alpha)100\%$ confidence interval is

$$(18) \qquad\qquad\qquad \left| \boldsymbol{a}^T \widehat{\boldsymbol{\beta}} - \boldsymbol{a}^T \boldsymbol{\beta} \right| \leq \widehat{\sigma} \Phi^{-1}(1 - \alpha/2)(\boldsymbol{a}^T \boldsymbol{V} \boldsymbol{a})^{1/2},$$

where Φ is the standard normal distribution function. We may choose $\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\}$ in (10) or (11) and $\boldsymbol{z}_j$ in Table 1 or (13) in the construction of $\widehat{\boldsymbol{\beta}}$ and the confidence intervals. An alternative, larger estimate of σ, producing more conservative approximate confidence intervals, is the penalized maximum likelihood estimator of [SBvdG10].

3. Theoretical Results

In this section, we prove that when the ℓ_1 loss of the initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$ is of an expected magnitude and the noise level estimator $\widehat{\sigma}$ is consistent, the LDPE based confidence interval has approximately the preassigned coverage probability for statistical inference of linear combinations of β_j with sufficiently small η_j. Under proper conditions on X such as those given in Proposition 1, the width of such confidence intervals is of the order $\tau_j \asymp n^{-1/2}$. The accuracy of the approximation for the coverage probability is sufficiently sharp to allow simultaneous interval estimation of all β_j and sharp error bounds for the estimation and selection errors of thresholded LDPE. We use existing error bounds to verify the conditions on $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$ under a capped-ℓ_1 relaxation of the sparsity condition $\|\boldsymbol{\beta}\|_0 \leq s$, provided that $s \log p \ll n^{1/2}$. Random matrix theory is used in Subsection 3.4 to check regularity conditions.

3.1. **Confidence intervals for preconceived parameters, deterministic design.** Here we establish the asymptotic normality of the LDPE (4) and the validity of the resulting confidence interval (18) for a preconceived parameter. This result is new and useful in and of itself since high-dimensional data often present a few effects known to be of high interest in advance. Examples include treatment effects in clinical trials, or the effect of education on income in social-economical studies. Simultaneous confidence intervals for all individual β_j and thresholded LDPE for the entire vector $\boldsymbol{\beta}$ will be considered in the next subsection as consequences of this result.

Let $\lambda_{univ} = \sqrt{(2/n) \log p}$. Suppose (1) holds with a vector $\boldsymbol{\beta}$ satisfying the following capped-ℓ_1 sparsity condition:

$$(19) \qquad\qquad\qquad \sum_{j=1}^p \min \{|\beta_j|/(\sigma \lambda_{univ}), 1\} \leq s.$$

This condition holds if $\boldsymbol{\beta}$ is ℓ_0 sparse with $\|\boldsymbol{\beta}\|_0 \le s$ or ℓ_q sparse with $\|\boldsymbol{\beta}\|_q^q/(\sigma\lambda_{univ})^q \le s$, $0 < q \le 1$. Let $\sigma^* = \|\boldsymbol{\varepsilon}\|_2/\sqrt{n}$. A generic condition we impose on the initial estimator is

$$(20) \qquad P\Big\{\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 \ge C_1 s\sigma^* \sqrt{(2/n)\log(p/\epsilon)}\Big\} \le \epsilon$$

for a certain fixed constant C_1 and all $\alpha_0/p^2 \le \epsilon \le 1$, where $\alpha_0 \in (0,1)$ is a preassigned constant. We also impose a similar generic condition on an estimator $\widehat{\sigma}$ for the noise level:

$$(21) \qquad P\Big\{|\widehat{\sigma}/\sigma^* - 1| \ge C_2 s(2/n)\log(p/\epsilon)\Big\} \le \epsilon, \ \forall \alpha_0/p^2 \le \epsilon \le 1,$$

with a fixed C_2. We use the same ϵ in (20) and (21) without much loss of generality.

By requiring fixed $\{C_1, C_2\}$, we implicitly impose regularity conditions on the design $\boldsymbol{X}$ and the sparsity index s in (19). Existing oracle inequalities can be used to verify (20) for various regularized estimators of $\boldsymbol{\beta}$ under different sets of conditions on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ [CT07, ZH08, BRT09, vdGB09, Zha09, Zha10, YZ10, SZ11, ZZ11]. Although most existing results are derived for penalty/threshold levels depending on a known noise level σ and under the ℓ_0 sparsity condition on $\boldsymbol{\beta}$, their proofs can be combined or extended to obtain (20) once (21) becomes available. For the joint estimation of $\{\boldsymbol{\beta}, \sigma\}$ with (10) or (11), specific sets of sufficient conditions for both (20) and (21), based on [SZ11], are stated in Subsection 3.3. In fact, the probability of the union of the two events is smaller than ϵ in the specific case where $\lambda_0 = A\sqrt{(2/n)\log(p/\epsilon)}$ in (10) for a certain $A > 1$.

Theorem 1. *Let $\widehat{\beta}_j$ be the LDPE in (4) with an initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$. Let η_j and τ_j be the bias and noise factors in (7), $\sigma^* = \|\boldsymbol{\varepsilon}\|_2/\sqrt{n}$, $\max(\epsilon_n', \epsilon_n'') \to 0$, and $\eta^* > 0$. Suppose (20) holds with $\eta^* C_1 s \sqrt{(2/n)\log(p/\epsilon)} \le \epsilon_n'$. If $\eta_j \le \eta^*$, then*

$$(22) \qquad P\Big\{|\tau_j^{-1}(\widehat{\beta}_j - \beta_j) - \boldsymbol{z}_j^T \boldsymbol{\varepsilon}/\|\boldsymbol{z}_j\|_2| > \sigma^* \epsilon_n'\Big\} \le \epsilon.$$

If in addition (21) holds with $C_2 s(2/n)\log(p/\epsilon) \le \epsilon_n''$, then for all $t \ge (1 + \epsilon_n')/(1 - \epsilon_n'')$,

$$(23) \qquad P\Big\{|\widehat{\beta}_j - \beta_j| \ge \tau_j \widehat{\sigma} t\Big\} \le 2\Phi_{n-1}(-(1 - \epsilon_n'')t + \epsilon_n') + 2\epsilon,$$

where $\Phi_n(t)$ is the student-t distribution function with n degrees of freedom. Moreover, for the covariance matrix $\boldsymbol{V}$ in (17) and all fixed m,

$$(24) \qquad \lim_{n\to\infty} \inf_{\boldsymbol{a} \in \mathscr{A}_{n,p,m}} P\Big\{|\boldsymbol{a}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{a}^T\boldsymbol{\beta}| \le \widehat{\sigma}\Phi^{-1}(1 - \alpha/2)(\boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a})^{1/2}\Big\} = 1 - \alpha,$$

where $\Phi(t) = P\{N(0,1) \le t\}$ and $\mathscr{A}_{n,p,m} = \{\boldsymbol{a} : \|\boldsymbol{a}\|_0 \le m, \max_{j\le p}|a_j|\eta_j \le \eta^*\}$.

Since $(\boldsymbol{z}_j^T\boldsymbol{\varepsilon}/\|\boldsymbol{z}_j\|_2, j \le p)$ has a multivariate normal distribution with identical marginal distributions $N(0, \sigma^2)$, (22) establishes the joint asymptotic normality of the LDPE for finitely many $\widehat{\beta}_j$ under (20). This allows us to write the LDPE as an approximate Gaussian sequence

$$(25) \qquad \widehat{\beta}_j = \beta_j + N(0, \tau_j^2\sigma^2) + o_P(\tau_j\sigma).$$

Under the additional condition (21), (23) and (24) justify the approximate coverage probability of the resulting confidence intervals.

Remark 1. *In Theorem 1, all conditions on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are imposed through (20), (21), and the requirement of relatively small η_j to work with these conditions. The uniform signal strength condition,*

$$\text{(26)} \qquad\qquad \min_{\beta_j \neq 0} |\beta_j| \geq C\sigma\sqrt{(2/n)\log p}, \ C > 1/2,$$

required for variable selection consistency [Wai09a, Zha10]*, is not required for (20) and (21). This is the most important feature of the LDPE that sets it apart from variable selection approaches. More explicit sufficient conditions for (20) and (21) are given in Subsection 3.3 for the initial estimators (10) and (11).*

Remark 2. *Although Theorem 1 does not require τ_j to be small, the noise factor is proportional to the width of the confidence interval and thus its square is reciprocal to the efficiency of the LDPE. The bias factor η_j is required to be relatively small for (1) and (4), but no condition is imposed on $\{\eta_k, k \neq j\}$ for the inference of β_j. Since η_j and τ_j are computed in Table 1, one may apply Theorem 1 to a set of the easy-to-estimate β_j with small $\{\eta_j, \tau_j\}$ and leave out some hard-to-estimate regression coefficients.*

In our implementation in Table 1, $\boldsymbol{z}_j$ is the residual of the Lasso estimator in the regression model for $\boldsymbol{x}_j$ against $\boldsymbol{X}_{-j} = (\boldsymbol{x}_k, k \neq j)$. It follows from Proposition 1 that under proper conditions on the design matrix, $\eta_j \asymp \sqrt{\log p}$ and $\tau_j \leq 1/\|\boldsymbol{z}_j\|_2 \asymp n^{-1/2}$ for the algorithm in Table 1. Such rates are realized in the simulation experiments described in Section 4 and further verified for Gaussian designs in Subsection 3.4. Thus, the dimension constraint for the asymptotic normality and proper coverage probability in Theorem 1 is $s(\log p)/\sqrt{n} \to 0$.

3.2. **Simultaneous confidence intervals and the thresholded LDPE.** Here we provide theoretical justifications for simultaneous applications of the proposed LDPE confidence interval, with multiplicity adjustments, in the absence of a preconceived parameter of interest. In Theorem 1, (22) is uniform in $\epsilon \in [\alpha_0/p^2, 1]$ and (23) is uniform in the corresponding t. This uniformity allows Bonferroni adjustments to control familywise error rate in simultaneous interval estimation. This uniformity also applies to the approximation in (25), leading to sharp ℓ_2 and selection error bounds of a thresholded LDPE for the estimation of the entire vector $\boldsymbol{\beta}$. We present these consequences of Theorem 1 in the following two theorems.

Theorem 2. *Suppose (20) holds with $\eta^* C_1 s\sqrt{(2/n)\log(p/\epsilon)} \leq \epsilon_n'$. Then,*

$$\text{(27)} \qquad\qquad P\left\{ \max_{\eta_j \leq \eta^*} \left| \tau_j^{-1}(\widehat{\beta}_j - \beta_j) - \boldsymbol{z}_j^T \boldsymbol{\varepsilon} / \|\boldsymbol{z}_j\|_2 \right| > \sigma^* \epsilon_n' \right\} \leq \epsilon.$$

If (21) also holds with $C_2 s(2/n)\log(p/\epsilon) \leq \epsilon_n''$, then for all $j \leq p$ and $t \geq (1 + \epsilon_n')/(1 - \epsilon_n'')$,

$$\text{(28)} \quad P\left\{ \max_{\eta_j \leq \eta^*} |\widehat{\beta}_j - \beta_j|/(\tau_j \widehat{\sigma}) > t \right\} \leq 2\Phi_n(-(1 - \epsilon_n'')t + \epsilon_n')\#\{j : \eta_j \leq \eta^*\} + 2\epsilon.$$

If, in addition to (20) and (21), $\max_{j \leq p} \eta_j \leq \eta^*$ and $\max(\epsilon_n', \epsilon) \to 0$ as $\min(n, p) \to \infty$, then for fixed $\alpha \in (0, 1)$ and $c_0 > 0$,

$$\text{(29)} \qquad\qquad \liminf_{n \to \infty} P\left\{ \max_{j \leq p} \left| \frac{\widehat{\beta}_j - \beta_j}{\tau_j(\widehat{\sigma} \wedge \sigma)} \right| \leq c_0 + \sqrt{2\log(p/\alpha)} \right\} \geq 1 - \alpha.$$

The error bound (27) asserts that the $o_P(1)$ in (25) is uniform in j. This uniform central limit theorem and the simultaneous confidence intervals (28) and (29) are valid as long as (20) and (21) hold with $s \log p = o(n^{1/2})$. Since (20) and (21) are consequences of (19) and proper regularity conditions on $\boldsymbol{X}$, these results do not require the uniform signal strength condition (26).

It follows from (25) and Proposition 1 (ii) that for a fixed j, the estimation error of $\widehat{\beta}_j$ is of the order $\tau_j \sigma$ and $\tau_j \asymp n^{-1/2}$ under proper conditions. With penalty level $\lambda = \sigma \sqrt{(2/n) \log p}$, the Lasso may have a high probability of estimating β_j by zero when $\beta_j = \lambda/2$. Thus, in the worst case scenario, the Lasso inflates the error by a factor of order $\sqrt{\log p}$. Of course, the Lasso is super efficient when it estimates the actual zero β_j by zero.

The situation is different for the estimation of the entire vector $\boldsymbol{\beta}$. The raw LDPE has an ℓ_2 error of order $\sigma^2 p/n$, compared with $\sigma^2 s (\log p)/n$ for the Lasso. However, this is not what the LDPE is designed for. The thrust of the LDPE approach is to turn the regression problem (1) into a Gaussian sequence model (25) with uniformly small approximation error and a consistent estimator of the covariance structure. The raw LDPE is sufficient for statistical inference of a preconceived β_j. For the estimation of the entire $\boldsymbol{\beta}$ or variable selection, our recommendation is to use a thresholded LDPE. We may use either the hard or soft thresholding methods:

$$(30) \qquad \widehat{\beta}_j^{(thr)} = \begin{cases} \widehat{\beta}_j I\{|\widehat{\beta}_j| > \widehat{t}_j\}, & \text{(hard threshold)} \\ \text{sgn}(\widehat{\beta}_j)(|\widehat{\beta}_j| - \widehat{t}_j)^+, & \text{(soft threshold)}, \end{cases}$$

$$\widehat{S}^{(thr)} = \{j : |\widehat{\beta}_j| > \widehat{t}_j\},$$

where $\widehat{\beta}_j$ is as in Theorem 1 and $\widehat{t}_j \approx \widehat{\sigma} \tau_j \Phi^{-1}(1 - \alpha/(2p))$ with $\alpha > 0$. Although the theory is similar between the two [DJ94], our explicit analysis focuses on soft-thresholding.

Theorem 3. *Let $L_0 = \Phi^{-1}(1 - \alpha/(2p))$, $\widetilde{t}_j = \tau_j \sigma L_0$, and $\widehat{t}_j = (1 + c_n) \widehat{\sigma} \tau_j L_0$ with positive constants α and c_n. Suppose (20) holds with $\eta^* C_1 s/\sqrt{n} \leq \epsilon_n'$, $\max_{j \leq p} \eta_j \leq \eta^*$, and*

$$(31) \qquad P\left\{ \frac{(\widehat{\sigma}/\sigma) \vee (\sigma/\widehat{\sigma}) - 1 + \epsilon_n' \sigma^*/(\widehat{\sigma} \wedge \sigma)}{1 - (\widehat{\sigma}/\sigma - 1)_+} > c_n \right\} \leq 2\epsilon.$$

Let $\widehat{\boldsymbol{\beta}}^{(thr)} = (\widehat{\beta}_1^{(thr)}, \ldots, \widehat{\beta}_p^{(thr)})^T$ be the soft thresholded LDPE (30) with these $\widehat{t}_j$. Then, there exists an event Ω_n with $P\{\Omega_n^c\} \leq 3\epsilon$ such that

$$(32) \qquad E\|\widehat{\boldsymbol{\beta}}^{(thr)} - \boldsymbol{\beta}\|_2^2 I_{\Omega_n} \leq \sum_{j=1}^p \min\left\{ \beta_j^2, \tau_j^2 \sigma^2 (L_0^2(1 + 2c_n)^2 + 1) \right\} + (\epsilon L_n/p) \sigma^2 \sum_{j=1}^p \tau_j^2,$$

where $L_n = 4/L_0^3 + 4c_n/L_0 + 12 c_n^2 L_0$. Moreover, with at least probability $1 - \alpha - 3\epsilon$,

$$(33) \qquad \{j : |\beta_j| > (2 + 2c_n) \widetilde{t}_j\} \subseteq \widehat{S}^{(thr)} \subseteq \{j : \beta_j \neq 0\}.$$

Theorem 3 asserts that thresholding the LDPE provides similar error bounds to thresholding a Gaussian sequence $N(\beta_j, \tau_j^2 \sigma^2)$, $j \leq p$. Since $\max_{j \leq p} \eta_j \leq C\sqrt{\log p}$ can be achieved under mild conditions, the main requirement is $s\sqrt{(\log p)/n} \to 0$ for the estimation and selection error bounds in (32) and (33). This is a weaker requirement than $s(\log p)/\sqrt{n} \to 0$ for the asymptotic normality in (24). When $C_2 s(2L_0^2/n) \leq \epsilon_n''$, (31) follows from (21) and

$P\{(1 - \epsilon_n''')/(1 - \epsilon_n'') \leq \sigma^*/\sigma \leq (1 + \epsilon_n''')/(1 + \epsilon_n'')\} \leq \epsilon$ with $c_n \geq (\epsilon_n''' + \epsilon_n')/(1 - \epsilon_n''')^2$. The condition on σ^*/σ is easy to check since $(\sigma^*/\sigma)^2 \sim \chi_n^2/n$. In what follows, we always assume that proper small constants $c_n > 0$ are taken in (31) so that it is a consequence of (21).

Remark 3. *The major difference between (33) and the existing variable selection consistency theory is again in the signal requirement. Variable selection consistency requires the uniform signal strength condition (26) as discussed in Remark 1, and existing variable selection methods are not guaranteed to select correctly variables with large $|\beta_j|$ or $\beta_j = 0$ in the presence of small $|\beta_j| \neq 0$. In comparison, Theorem 3 makes no assumption of (26). Under the regularity conditions for (33), large $|\beta_j|$ are selected by the thresholded LDPE and $\beta_j = 0$ are not selected, in the presence of possibly many small nonzero $|\beta_j|$.*

The analytical difference between the thresholded LDPE and existing regularized estimators lies in the quantities thresholded. For the LDPE, the effect of thresholding to the approximate Gaussian sequence (25) is explicit and requires only univariate analysis to understand. In comparison, for the Lasso and some other regularized estimators, thresholding is applied to the gradient $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n$ via the Karush-Kuhn-Tucker type condition, leading to more complicated nonlinear multivariate analysis.

For the estimation of $\boldsymbol{\beta}$, the order of the ℓ_2 error bound in (32), $\sum_{j=1}^p \min(\beta_j^2, \sigma^2 \lambda_{univ}^2)$, is slightly sharper than the typical order of $\|\boldsymbol{\beta}\|_0 \sigma^2 \lambda_{univ}^2$ or $\sigma \lambda_{univ} \sum_{j=1}^p \min\{|\beta_j|, \sigma \lambda_{univ}\}$ in the literature, where $\lambda_{univ} = \sqrt{(2/n)\log p}$. However, since the Lasso and other regularized estimators are proven to be rate optimal in the ℓ_2 estimation loss for many classes of sparse $\boldsymbol{\beta}$, the main advantage of the thresholded LDPE seems to be the clarity of the effect of thresholding to the individual $\widehat{\beta}_j$ in the approximate Gaussian sequence (25).

3.3. **Checking conditions by oracle inequalities.** Our main theoretical results, stated in Theorems 1, 2, and 3 in the above two subsections, provide justifications for the LDPE-based confidence interval of a single preconceived linear parameter of $\boldsymbol{\beta}$, simultaneous confidence intervals for all β_j, and the estimation and selection error bounds for the thresholded LDPE for the vector $\boldsymbol{\beta}$. These results are based on conditions (20) and (21). We have mentioned that for proper $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$, these two generic conditions can be verified in many ways under condition (19) based on existing results. The purpose of this subsection is to describe a specific way of verifying these two conditions and thus provide a more definitive and complete version of the theory.

In regularized linear regression, oracle inequalities have been established for different regularized estimators and loss functions. We confine our discussion here to the scaled Lasso (10) and the scaled Lasso-LSE (11) as specific choices of the initial estimator, since the confidence interval in Theorem 1 is based on the joint estimation of regression coefficients and the noise level. We further confine our discussion to bounds for the ℓ_1 error of $\widehat{\boldsymbol{\beta}}^{(init)}$ and the relative error of $\widehat{\sigma}$ involved in (20) and (21).

We use the results in [SZ11] where properties of estimators (10) and (11) were established based on a compatibility factor [vdGB09] and sparse eigenvalues. Let $\xi \geq 1$, $S = \{j : |\beta_j| > $

$\sigma\lambda_{univ}\}$, and $\mathscr{C}(\xi, S) = \{\boldsymbol{u} : \|\boldsymbol{u}_{S^c}\|_1 \le \xi\|\boldsymbol{u}_S\|_1\}$. The compatibility factor is defined as

$$(34) \qquad \kappa(\xi, S) = \inf\{\|\boldsymbol{X}\boldsymbol{u}\|_2 |S|^{1/2}/(n^{1/2}\|\boldsymbol{u}_S\|_1) : 0 \ne \boldsymbol{u} \in \mathscr{C}(\xi, S)\}.$$

Let $\phi_{\min}$ and $\phi_{\max}$ denote the smallest and largest eigenvalues of matrices respectively. For positive integers m, define sparse eigenvalues as

$$\phi_-(m, S) = \min_{B \supset S, |B \setminus S| \le m} \phi_{\min}(\boldsymbol{X}_B^T \boldsymbol{X}_B/n),$$
$$(35) \qquad \phi_+(m, S) = \min_{B \cap S = \emptyset, |B| \le m} \phi_{\max}(\boldsymbol{X}_B^T \boldsymbol{X}_B/n).$$

The following theorem is a consequence of checking the conditions of Theorem 1 by Theorems 2 and 3 in [SZ11].

Theorem 4. *Let $\{A, \xi, c_0\}$ be fixed positive constants with $\xi > 1$ and $A > (\xi + 1)/(\xi - 1)$. Let $\lambda_0 = A\sqrt{(2/n)\log(p/\epsilon)}$. Suppose $\boldsymbol{\beta}$ is sparse in the sense of (19), $\kappa^2(\xi, S) \ge c_0$, and $(s \vee 1)(2/n)\log(p/\epsilon) \le \mu_*$ for a certain $\mu^* > 0$.*

(i) Let $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$ be the scaled Lasso estimator in (10). Then, conditions (20) and (21) hold for certain constants $\{\mu_*, C_1, C_2\}$ depending on $\{A, \xi, c_0\}$ only. Consequently, all conclusions of Theorems 1, 2, and 3 hold with $C_1\eta^*(s\lambda_0/A) \le \epsilon'_n$ and $C_2 s(\lambda_0/A)^2 \le \epsilon''_n$.

(ii) Let $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$ be the scaled Lasso-LSE in (11). Suppose $\xi^2/\kappa^2(\xi, S) \le K/\phi_+(m, S)$ and $\phi_-(m, S) \ge c_1 > 0$ for certain $K > 0$ and integer $m - 1 < K|S| \le m$. Then, (20) and (21) hold for certain constants $\{\mu^*, C_1, C_2\}$ depending on $\{A, \xi, c_0, c_1, K\}$ only. Consequently, all conclusions of Theorems 1, 2, and 3 hold with $C_1\eta^*(s\lambda_0/A) \le \epsilon'_n$ and $C_2 s(\lambda_0/A)^2 \le \epsilon''_n$.

Remark 4. *Let $A > (\xi + 1)/(\xi - 1)$ as in Theorem 4 (i). Then, there exist constants $\{\tau_0, \nu_0\} \subset (0, 1)$ satisfying the condition $(1 - \tau_0^2)A = (\xi + 1)/\{\xi - (1 + \nu_0)/(1 - \nu_0)\}$. For these $\{\tau_0, \nu_0\}$, $n \ge 3$, and $p \ge 7$, we may take*

$$(36) \quad \mu_* = \min\left\{\frac{2c_0\tau_0^2}{A^2(\xi + 1)}, \frac{\tau_0^2/(1/\nu_0 - 1)}{2A(\xi + 1)}, \log(4/e)\right\}, \ C_2 = \frac{\tau_0^2}{\mu_*}, \ C_1 = \frac{C_2}{A(1 - \tau_0^2)}.$$

The main conditions of Theorem 4 are

$$(37) \qquad \kappa^2(\xi, S) \ge c_0, \ \xi^2/\kappa^2(\xi, S) \le K/\phi_+(m, S), \ \phi_-(m, S) \ge c_1,$$

where m is the smallest integer upper bound of $K|S|$. While Theorem 4 (i) requires only the first inequality in (37), Theorem 4 (ii) requires all three. Let

$$RE_2(\xi, S) = \inf\{\|\boldsymbol{X}\boldsymbol{u}\|_2/(n^{1/2}\|\boldsymbol{u}\|_2) : \boldsymbol{u} \in \mathscr{C}(\xi, S), u_j\boldsymbol{x}_j^T\boldsymbol{X}\boldsymbol{u} \le 0, j \notin S\},$$
$$F_1(\xi, S) = \inf\{\|\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u}\|_\infty |S|/(n\|\boldsymbol{u}_S\|_1) : \boldsymbol{u} \in \mathscr{C}(\xi, S), u_j\boldsymbol{x}_j^T\boldsymbol{X}\boldsymbol{u} \le 0, j \notin S\},$$

be respectively the restricted eigenvalue and sign restricted cone invertibility factor for the Gram matrix. It is worthwhile to note that

$$(38) \qquad F_1(\xi, S) \ge \kappa^2(\xi, S) \ge RE_2^2(\xi, S)$$

always holds and lower bounds of these quantities can be expressed in terms of sparse eigenvalues [YZ10]. By [SZ11], one may replace $\kappa^2(\xi, S)$ throughout Theorem 4 with $F_1(\xi, S)$. In view of (38), this will actually weaken the condition. However, since more explicit proofs

are given in terms of $\kappa(\xi, S)$ in [SZ11], the compatibility factor is used in Theorem 4 to facilitate a direct matching of proofs between the two papers.By [ZH08, Zha10, HZ12], (37) can be replaced by the sparse Riesz condition,

$$(39) \qquad s \leq d^*/\{\phi_+(d^*, \emptyset)/\phi_-(d^*, \emptyset) + 1/2\}.$$

Proposition 2 below provides a way of checking (37) for a given design in (1).

Proposition 2. *Let* $\{\xi, M_0, c_*, c^*\}$ *be fixed positive constants,* $\lambda_1 = M_0\sqrt{(\log p)/n}$, *and*

$$\widehat{\boldsymbol{\Sigma}} = \left((\boldsymbol{x}_j^T\boldsymbol{x}_k/n)I\{|\boldsymbol{x}_j^T\boldsymbol{x}_k/n| \geq \lambda_1\}\right)_{p \times p}$$

be the thresholded Gram matrix. Suppose $\phi_{\min}(\widehat{\boldsymbol{\Sigma}}) \geq c_*$ *and* $s\lambda_1(1 + \xi)^2 \leq c_*/2$. *Then, for all* $|S| \leq s$, $\kappa^2(\xi, S) \geq c_*/2$. *Let* $K = 2\xi^2(c^*/c_* + 1/2)$. *If in addition,* $\phi_{\max}(\widehat{\boldsymbol{\Sigma}}) \leq c^*$ *and* $s\lambda_1(1 + K) + \lambda_1 \leq c_*/2$, *then* $\phi_-(m, S) \geq c_*/2$ *and (37) holds with* $c_0 = c_*/2$.

The main condition of Proposition 2 is a small $s\sqrt{(\log p)/n}$. This is not restrictive since Theorem 1 requires the stronger condition of a small $s(\log p)/\sqrt{n}$. It follows from [BL08] that after hard thresholding at a level of order λ_1, sample covariance matrices converge to a population covariance matrix in the spectrum norm under mild sparsity conditions on the population covariance matrix. Since convergence in the spectrum norm implies convergence of the minimum and maximum eigenvalues, $\phi_{\min}(\widehat{\boldsymbol{\Sigma}}) \geq c_*$ and $\phi_{\max}(\widehat{\boldsymbol{\Sigma}}) \leq c^*$ are reasonable conditions. This and other applications of random matrix theory are discussed in the next subsection.

3.4. **Checking conditions by random matrix theory.** The most basic conditions for our main theoretical results in Subsections 3.1 and 3.2 are (20), (21), and the existence of $\boldsymbol{z}_j$ with small η_j and τ_j. For deterministic design matrices, sufficient conditions for (20) and (21) are given in Theorem 4 in the form of (37), and sufficient conditions for the existence of $\eta_j \leq C\sqrt{\log p}$ and $\tau_j \asymp n^{-1/2}$ are given in Proposition 1. These sufficient conditions are all analytical ones on the design matrix. In this subsection, we use random matrix theory to check these conditions with more explicit constant factors.

The conditions of Theorems 1 and 4 hold in the following classes of design matrices:

$$\begin{aligned}
\mathscr{X}_{s,n,p} &= \mathscr{X}_{s,n,p}(c_*, \delta, \xi, K) \\
&= \Big\{\boldsymbol{X} : \max_{j \leq p} \eta_j \leq 3\sqrt{\log p}, \ \max_{j \leq p} \tau_j^2 \sigma_j^2 \leq 2/n, \ \min_{|S| \leq s} \kappa^2(\xi, S) \geq c_*(1 - \delta)/4,
\end{aligned}$$
$$(40) \qquad \qquad \max_{|S| \leq s} \phi_+(m, S)\xi^2/\kappa^2(\xi, S) \leq K, \ \min_{|S| \leq s} \phi_-(m, S) \geq c_*(1 - \delta)\Big\},$$

for certain positive $\{s, c_*, \delta, \xi, K\}$, where $\{\eta_j, \tau_j\}$ are computed from $\boldsymbol{X}$ by the algorithm in Table 1 with $\kappa_0 \leq 1/4$ and $3/(1 + \kappa_1) > \sqrt{8}$, and $\kappa(\xi, S)$ and $\phi_\pm(m, S)$ are the compatibility factor and sparse eigenvalues of $\boldsymbol{X}$ given in (34) and (35), with $m - 1 < Ks \leq m$. We note that $1/\sigma_j^2 \leq 1/c_*$ by (44), so that $\max_{j \leq p} \tau_j^2 \leq 2/(nc_*)$ in $\mathscr{X}_{s,n,p}(c_*, \delta, \xi, K)$.

Let $P_{\boldsymbol{\Sigma}}$ be probability measures under which

$$(41) \qquad \qquad \widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_p) \in \mathbb{R}^{n \times p} \ \text{ has iid } N(0, \boldsymbol{\Sigma}) \text{ rows.}$$

The column standardized version of $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_p)$ is

$$(42) \qquad \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p), \quad \boldsymbol{x}_j = \widetilde{\boldsymbol{x}}_j \sqrt{n}/\|\widetilde{\boldsymbol{x}}_j\|_2.$$

Since our discussion is confined to column standardized design matrices for simplicity, we assume without loss of generality that the diagonal elements of $\boldsymbol{\Sigma}$ all equal to 1. Under $P_{\boldsymbol{\Sigma}}$, $\boldsymbol{X}$ does not have independent rows but $\boldsymbol{x}_j$ is still related to $\boldsymbol{X}_{-j}$ through

$$(43) \qquad \boldsymbol{x}_j = \boldsymbol{X}_{-j}\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j \sqrt{n}/\|\widetilde{\boldsymbol{x}}_j\|_2, \ \boldsymbol{\varepsilon}_j \sim N(0, \sigma_j^2 I_{n \times n}),$$

where $\boldsymbol{\varepsilon}_j$ is independent of $\boldsymbol{X}_{-j}$. Let Θ_{jk} be the elements of $\boldsymbol{\Sigma}^{-1}$. Since the linear regression of $\widetilde{\boldsymbol{x}}_j$ against $(\widetilde{\boldsymbol{x}}_k, k \neq j)$ has coefficients $-\Theta_{jk}/\Theta_{jj}$ and noise level $1/\Theta_{jj}$, we have

$$(44) \qquad \boldsymbol{\gamma}_j = \left(-\sigma_j^2 \Theta_{jk}\|\widetilde{\boldsymbol{x}}_k\|_2/\|\widetilde{\boldsymbol{x}}_j\|_2, k \neq j \right)^T, \ \sigma_j^2 = 1/\Theta_{jj}.$$

The aim of this subsection is to prove that $P_{\boldsymbol{\Sigma}}(\mathscr{X}_{s,n,p})$ is uniformly large for a general collection of $P_{\boldsymbol{\Sigma}}$. This result has two interpretations. The first interpretation is that when $\boldsymbol{X}$ is indeed generated in accordance with (41) and (42), the regularity conditions have a high probability to hold. The second interpretation is that $\mathscr{X}_{s,n,p}$, a deterministic subset of $\mathbb{R}^{n \times p}$, is sufficiently large as measured by $P_{\boldsymbol{\Sigma}}$ in the collection. Since $\mathscr{X}_{s,n,p}$ does not depend on $\boldsymbol{\Sigma}$ and the probability measures $P_{\boldsymbol{\Sigma}}$ are nearly orthogonal for different $\boldsymbol{\Sigma}$, the use of $P_{\boldsymbol{\Sigma}}$ does not add the random design assumption to our results.

The following theorem specifies $\{c_*, c^*, \delta, \xi, K\}$ in (40) for which $P_{\boldsymbol{\Sigma}}\{\mathscr{X}_{s,n,p}(c_*, \delta, \xi, K)\}$ is large when $s(\log p)/n$ is small. This works with the LDPE theory since $s(\log p)/\sqrt{n} \to 0$ is required anyway in Theorem 1. Define a class of coefficient vectors with small ℓ_q tail as

$$\mathscr{B}_q(s, \lambda) = \left\{ \boldsymbol{b} \in \mathbb{R}^p : \sum_{j=1}^p \min(|b_j|^q/\lambda^q, 1) \leq s \right\}.$$

We note that $\mathscr{B}_q(s, \sigma\lambda_{univ})$ is the collection of all $\boldsymbol{\beta}$ satisfying the capped-ℓ_1 sparsity condition (19).

Theorem 5. *Suppose $diag(\boldsymbol{\Sigma}) = \boldsymbol{I}_{p \times p}$, eigenvalues$(\boldsymbol{\Sigma}) \subset [c_*, c^*]$, and all rows of $\boldsymbol{\Sigma}^{-1}$ are in $\mathscr{B}_1(s, \lambda_{univ})$. Then, there exist positive numerical constants $\{\delta_0, \delta_1, \delta_2\}$ and K depending only on $\{\delta_1, \xi, c_*, c^*\}$ such that*

$$\inf_{(K+1)(s+1) \leq \delta_0 n/\log p} P_{\boldsymbol{\Sigma}}\{\boldsymbol{X} \in \mathscr{X}_{s,n,p}(c_*, \delta_1, \xi, K)\} \geq 1 - e^{-\delta_2 n}.$$

Consequently, when the $\boldsymbol{X}$ in (1) is indeed generated from (41) and (42), all conclusions of Theorems 1, 2, and 3 hold for both (10) and (11) with an adjustment of a probability smaller than $2e^{-\delta_2 n}$, provided that $\boldsymbol{\beta} \in \mathscr{B}_1(s, \sigma\lambda_{univ})$ and $\lambda_0 = A\sqrt{(2/n)\log(p/\epsilon)}$ in (10) with a fixed $A > (\xi + 1)/(\xi - 1)$.

Remark 5. *It follows from Theorem II.13 of [DS01] that for certain positive $\{\delta_0, \delta_1, \delta_2\}$,*

$$\mathscr{X}'_{n,p} = \left\{ \boldsymbol{X} : \min_{|S|+m \leq \delta_0 n/\log p} \phi_-(m, S) \geq c_*(1 - \delta_1), \max_{|S|+m \leq \delta_0 n/\log p} \phi_+(m, S) \leq c^*(1 + \delta_1) \right\}$$

satisfies $P_{\boldsymbol{\Sigma}}\{\mathscr{X}'_{n,p}\} \geq 1 - e^{-\delta_2 n}$ for all $\boldsymbol{\Sigma}$ in Theorem 5 [CT05, ZH08]. Let $K = 4\xi^2(c^*/c_*)(1 + \delta_1)/(1 - \delta_1)$ and $\{k, \ell\}$ be positive integers satisfying $4\ell/k \geq K$ and $\max\{k + \ell, 4\ell\} \leq \delta_0 n/\log p$. For $\boldsymbol{X} \in \mathscr{X}'_{n,p}$, the conditions

$$\kappa(\xi, S) \geq \{c_*(1 - \delta_1)\}^{1/2}/2, \ \xi^2 \phi_+(m, S)/\kappa^2(\xi, S) \leq K,$$

hold for all $|S| \leq k$, where m is smallest integer upper bound of $K|S|$.

The $P_{\boldsymbol{\Sigma}}$-induced regression model (43) provides a motivation for the use of the Lasso in (12) and Table 1 to generate score vectors $\boldsymbol{z}_j$. However, the goal of the procedure is to find $\boldsymbol{z}_j$ with small η_j and τ_j for controlling the variance and bias of the LDPE (4) as in Theorem 1. This is quite different from the usual applications of the Lasso for prediction, estimation of regression coefficients, or model selection.

4. Simulation Results

We set $n = 200$, $p = 3000$, and run several simulation experiments with 100 replications in each setting. In each replication, we generate an independent copy of $(\widetilde{\boldsymbol{X}}, \boldsymbol{X}, \boldsymbol{y})$, where, given a particular $\rho \in (-1, 1)$, $\widetilde{\boldsymbol{X}} = (\widetilde{x}_{ij})_{n \times p}$ has iid $N(0, \boldsymbol{\Sigma})$ rows with $\boldsymbol{\Sigma} = (\rho^{|j-k|})_{p \times p}$, $\boldsymbol{x}_j = \widetilde{\boldsymbol{x}}_j \sqrt{n}/|\widetilde{\boldsymbol{x}}_j|_2$, and $(\boldsymbol{X}, \boldsymbol{y})$ is as in (1) with $\sigma = 1$. Given a particular $\alpha \geq 1$, $\beta_j = 3\lambda_{univ}$ for $j = 1500, 1800, 2100, \ldots, 3000$, and $\beta_j = 3\lambda_{univ}/j^{\alpha}$ for all other j, where $\lambda_{univ} = \sqrt{(2/n)\log p}$. Our simulation design is set to test the performance of the LDPE methods beyond the assumptions of the theorems in Section 3; this setup gives $(s, s*(\log p)/n^{1/2}) = (8.93, 5.05)$ and $(29.24, 16.55)$ respectively for $\alpha = 2$ and 1, while the theorems require $s(\log p)/\sqrt{n} \to 0$, where $s = \sum_j \min(|\beta_j|/\lambda_{univ}, 1)$. This simulation example includes four cases, labeled (A), (B), (C), and (D), respectively: $(\alpha, \rho) = (2, 1/5), (1, 1/5), (2, 4/5)$, and $(1, 4/5)$, with case (D) being the most difficult one.

		Estimator					
		Lasso	scLasso	scLasso-LSE	oracle	LDPE	R-LDPE
(A)	bias	-0.2965	-0.4605	-0.0064	-0.0045	-0.0038	-0.0028
	sd	0.0936	0.1360	0.1004	0.0730	0.0860	0.0960
	median abs error	0.2948	0.4519	0.0549	0.0507	0.0531	0.0627
(B)	bias	-0.2998	-0.5341	-0.0476	0.0049	-0.0160	-0.0167
	sd	0.1082	0.1590	0.2032	0.0722	0.1111	0.1213
	median abs error	0.2994	0.5150	0.0693	0.0500	0.0705	0.0799
(C)	bias	-0.3007	-0.4423	-0.0266	-0.0049	-0.0194	-0.0181
	sd	0.1207	0.1520	0.1338	0.1485	0.1358	0.1750
	median abs error	0.3000	0.4356	0.0657	0.0994	0.0902	0.1150
(D)	bias	-0.3258	-0.5548	-0.1074	-0.0007	-0.0510	-0.0405
	sd	0.1367	0.1844	0.2442	0.1455	0.1768	0.2198
	median abs error	0.3319	0.5620	0.0857	0.0955	0.1112	0.1411

Table 2. Summary statistics for various estimates of the maximal $\beta_j = |\boldsymbol{\beta}|_{\infty}$: the Lasso, the scaled Lasso (scLasso), the scaled Lasso-LSE (scLasso-LSE), the oracle estimator, the LDPE, and the R-LDPE.

In addition to the Lasso with penalty level λ_{univ}, the scaled Lasso (10) with penalty level $\lambda_0 = \lambda_{univ}$, and the scaled Lasso-LSE (11), we consider an oracle estimator along with the
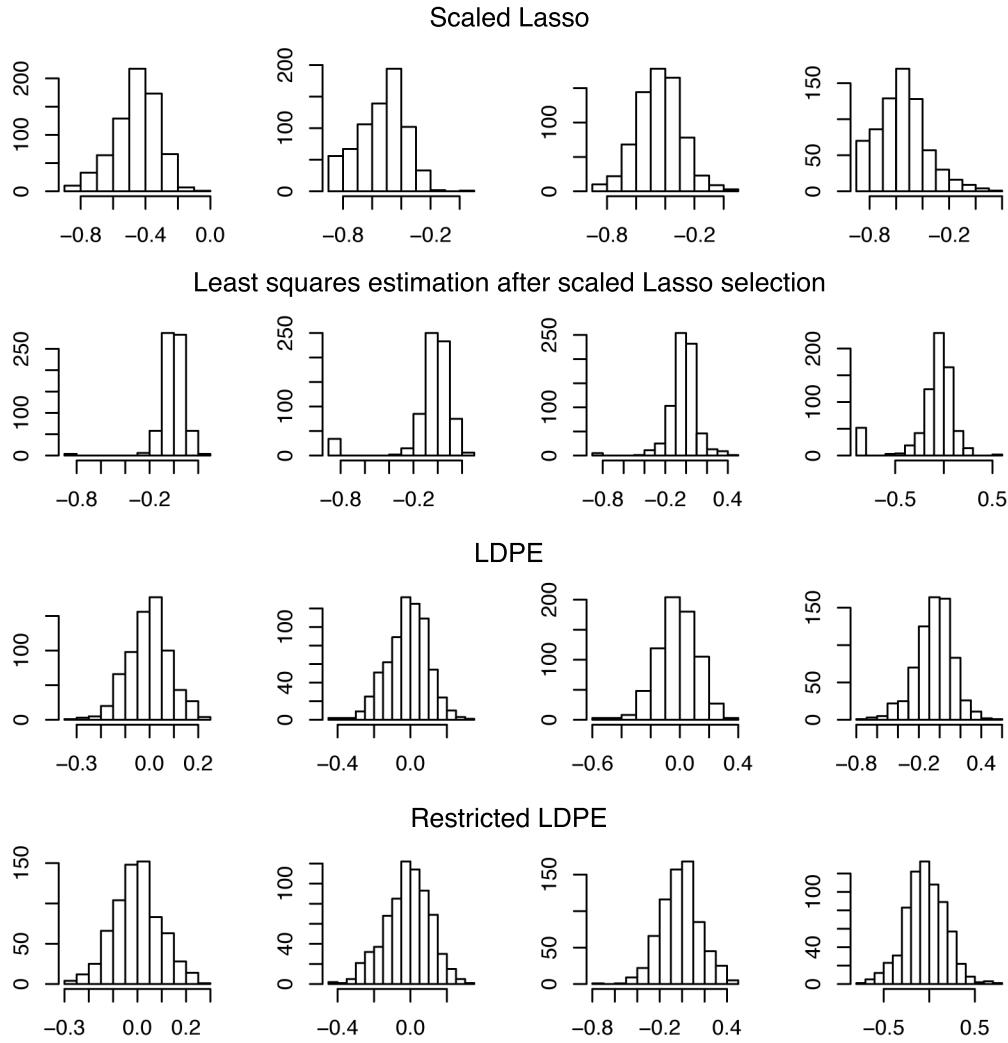
Scaled Lasso

Least squares estimation after scaled Lasso selection

LDPE

Restricted LDPE

FIGURE 1. Histogram of errors when estimating maximal β_j using the scaled Lasso, the scaled Lasso-LSE, the LDPE, and the R-LDPE. From left to right, plots correspond to simulation settings (A), (B), (C), and (D).

LDPE (4) and its restricted version derived from (13), the R-LDPE. The oracle estimator is the the least squares estimator of β_j when the β_k are given for all $k \neq j$ except for those k with $|k - j|$ among the smallest three. It can be written as

$$(45) \qquad \widehat{\beta}_j^{(o)} = \frac{(\boldsymbol{z}_j^{(o)})^T}{\|\boldsymbol{z}_j^{(o)}\|_2^2}\Big(\boldsymbol{y} - \sum_{k \notin K_j} \boldsymbol{x}_k \beta_k\Big), \ \widehat{\sigma}^{(o)} = \|\boldsymbol{P}_{K_j}^{\perp} \boldsymbol{\varepsilon}\|_2/\sqrt{n},$$

where $K_j = \{j - 1, j, j + 1\}$ for $1 < j < p$, $K_1 = \{1, 2, 3\}$, $K_p = \{p - 2, p - 1, p\}$, and $\boldsymbol{z}_j^{(o)} = \boldsymbol{P}_{K_j \setminus \{j\}}^{\perp} \boldsymbol{x}_j$. Here, $\boldsymbol{P}_K^{\perp}$ is the orthogonal projection to the space of n-vectors

orthogonal to $\{\boldsymbol{x}_k, k \in K\}$. Note that the oracular knowledge reduces the complexity of the problem from $(n, p) = (200, 3000)$ to $(n, p) = (200, 3)$, and that the variables $\{\boldsymbol{x}_k, k \in K_j\}$ also have the highest correlation to $\boldsymbol{x}_j$. For both the LDPE and the R-LDPE, the scaled Lasso-LSE (11) is used to generate $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$, while the algorithm in Table 1 is used to generate $\boldsymbol{z}_j$, with $\kappa_0 = 1/4$. The default $\eta_j^* = \sqrt{2 \log p}$ passed the test in Step 1 of Table 1 without adjustment in all instances in the simulation study. This guarantees $\eta_j \leq \sqrt{2 \log p}$ for the bias factor. For the R-LDPE, $m = 4$ is used in (13).

The asymptotic normality of the LDPE holds well in our simulation experiments. Table 2 and Figure 1 demonstrate the behavior of the LDPE and R-LDPE for the largest β_j, compared with that of the other estimation methods. The scaled Lasso has more bias and a larger variance than the Lasso, but is entirely data-driven. The bias can be significantly reduced though the scaled Lasso-LSE; however, error resulting from failure to select some maximal β_j remains. This is clearest in the histograms corresponding the distribution of errors for the scaled Lasso-LSE in settings (B) and (D), where $\alpha = 1$ and the β_j decay at a slower rate. For a small increase in variance, the LDPE and R-LDPE further reduce the bias of the scaled Lasso-LSE. This is also the case when $\widehat{\boldsymbol{\beta}}^{(init)}$ is a heavily biased estimator such as the Lasso or scaled Lasso, and the improvement is most dramatic when estimating large β_j. Although the asymptotic normality of the LDPE holds even better for small β_j in the simulation study, a parallel comparison for small β_j is not meaningful; the Lasso typically estimates small β_j by zero, while the raw LDPE is not designed to be sparse.

		(A)	(B)	(C)	(D)
all β_j	LDPE	0.9597	0.9845	0.9556	0.9855
	R-LDPE	0.9595	0.9848	0.9557	0.9885
maximal β_j	LDPE	0.9571	0.9814	0.9029	0.9443
	R-LDPE	0.9614	0.9786	0.9414	0.9786

TABLE 3. Mean coverage probability of LDPE and R-LDPE.

The overall coverage probability of the LDPE-based confidence interval matches relatively well to the preassigned level, as expected from our theoretical results. The LDPE and R-LDPE create confidence intervals $\widehat{\beta}_j \pm 1.96 \widehat{\sigma} \tau_j$ with approximately 95% coverage in settings (A) and (C) and somewhat higher coverage probability in (B) and (D). Refer to Table 3 for precise values. Since the coverage probabilities for each individual β_j are calculated based on a sample of 100 replications, the empirical distribution of the simulated relative coverage frequencies exhibits some randomness, which matches that of the binomial$(n, \widetilde{p})$ distribution, with $n = 100$ and $\widetilde{p}$ equal to the simulated mean coverage, as shown in Figure 2.

Two separate issues may lead to some variability in the coverage. As is the case with settings (B) and (D), overall coverage may exceed the stated confidence level when presence of many small signals in β is interpreted as noise, increasing $\widehat{\sigma}$ and hence the width of the confidence intervals, along with the coverage; however, this phenomenon will not result in under-coverage. In addition, compared with the overall coverage probability, the coverage probability is somewhat smaller when large values of β_j are associated with highly
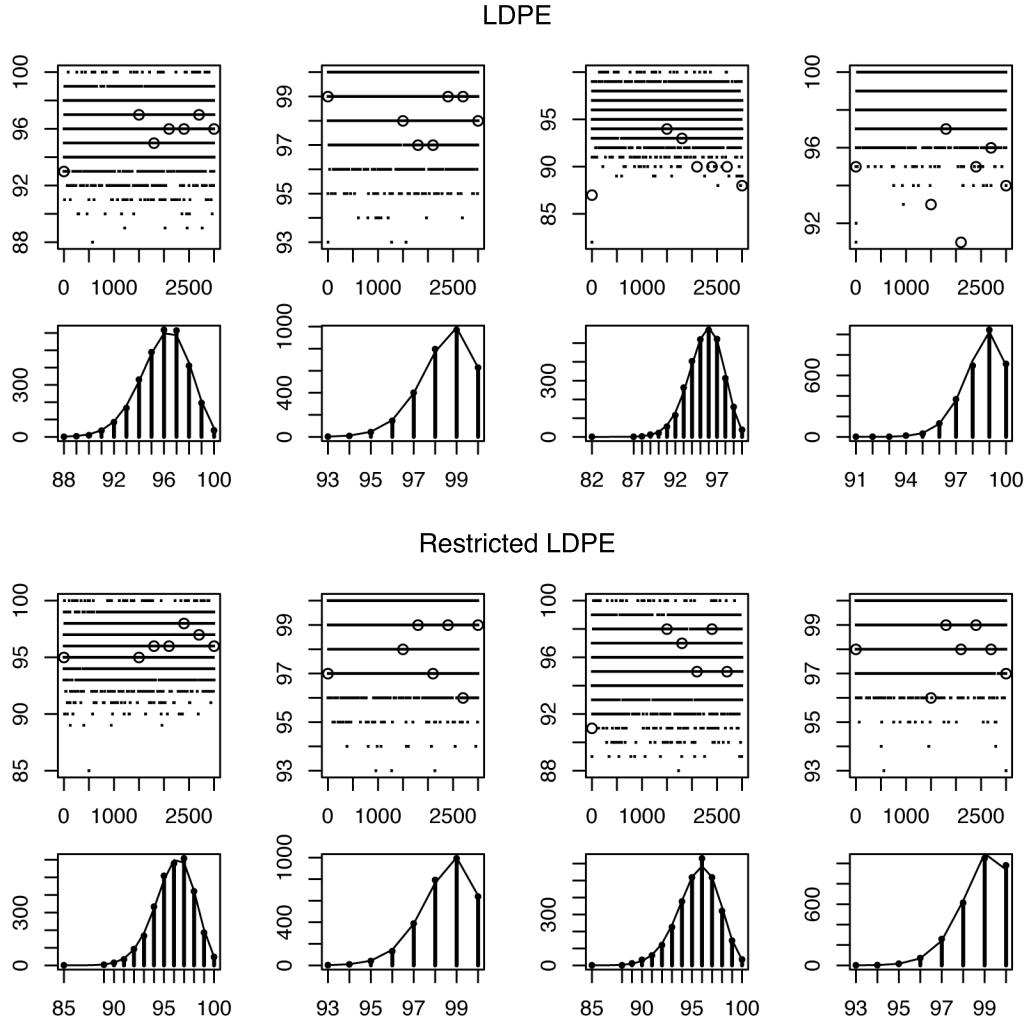
FIGURE 2. Rows 1 and 3: Coverage frequencies versus the index of β_j. Points corresponding to maximal β_j are plotted as large circles. Rows 2 and 4: The number of variables for given values of the relative coverage frequency, superimposed on the binomial$(100, \tilde{p})$ probability mass function, where $\tilde{p}$ is the simulated mean coverage. Figures depict results from simulations (A), (B), (C), and (D), from left to right.

correlated columns of $\boldsymbol{X}$. This is most apparent when plotting coverage versus index in (C) and (D), the two settings with higher correlation between adjacent columns of $\boldsymbol{X}$. For additional clarity, the points corresponding to maximal values of β_j in Figure 2 are emphasized by larger circles, and the coverage of the LDPE and R-LDPE for maximal β_j are listed separately from the overall coverage in the last two rows of Table 3. It can be seen from these details that the R-LDPE (13) further eliminates the bias caused by relatively
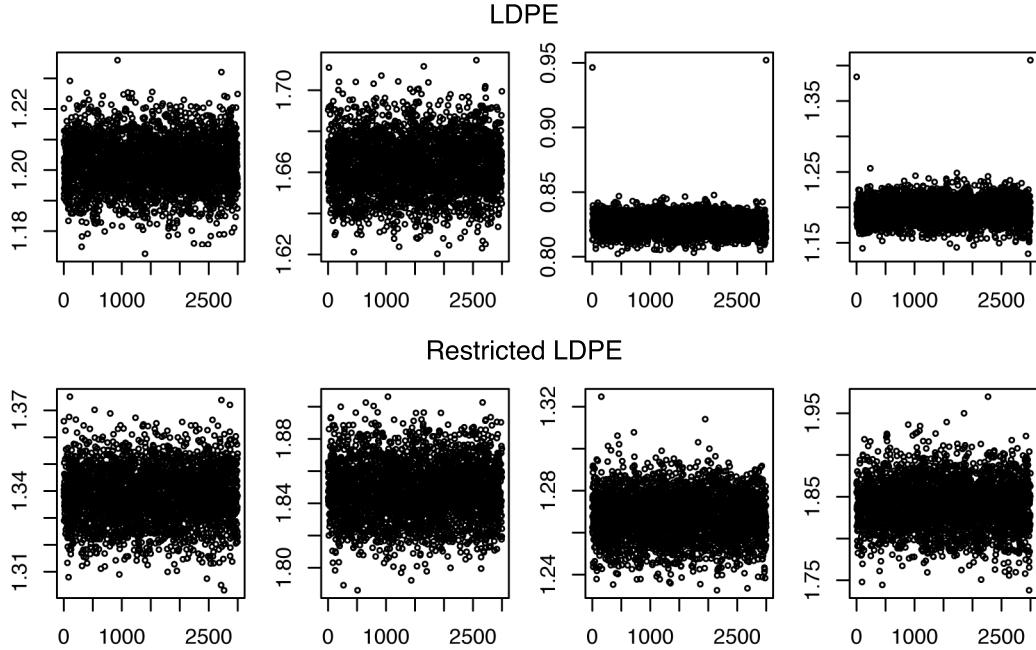
LDPE

Restricted LDPE

FIGURE 3. Median ratio of width of the LDPE and R-LDPE confidence intervals versus the oracle confidence interval for each β_j.

large values of β_j associated with highly correlated columns of $\boldsymbol{X}$ and improves coverage probabilities. The bias correction effect can be also seen in the histograms in Figure 1 in setting (D), but not in (C).

	(A)	(B)	(C)	(D)
LDPE	1.2020	1.6400	0.8209	1.1758
R-LDPE	1.3359	1.8238	1.2678	1.8150

TABLE 4. Median of the width ratio medians in Figure 3.

The LDPE and R-LDPE confidence intervals are of reasonable width, comparable to that of the confidence intervals derived from the oracle estimator. Consider the median ratio between the width of the LDPE (and restrictd LDPE) confidence intervals and the oracle confidence intervals, shown in Figure 3. The distribution of the median ratio associated with each β_j is uniform over the different $j = 1, \ldots, 3000$ in settings (A) and (B). The anomalies at $j = 1$ and $j = 3000$ in settings (C) and (D) are a result of the structure of $\boldsymbol{X}$. When the correlation between nearby columns of $\boldsymbol{X}$ is high, the fact that the first and last columns of $\boldsymbol{X}$ have fewer highly-correlated neighbors gives the oracle a relatively greater advantage. Since the medians of the ratios are uniformly distributed over j, it is reasonable to summarize the ratios in each simulation setting with the median value over

every replication of every β_j, as listed in Table 4. Note that the LDPE is more efficient than the oracle estimator in the high-correlation settings (C) and (D). This is probably due to the benefit of relaxing the orthogonality constraint of $\boldsymbol{x}_j^{\perp}$ when the correlation of the design is high and the error of the initial estimator is relatively small. The median ratio between the widths for the LDPE estimator reaches its highest value of 1.6400 in setting (B), where the coverage of the LDPE intervals is high and the benefit of relaxing the orthogonality constraint is small, if any, relative to the oracle.

Recall that the R-LDPE improves the coverage probability for large β_j at the cost of an increase in the variance of the estimator; thus, the R-LDPE confidence intervals are somewhat wider than the LDPE confidence intervals. Although the improvement in coverage probability is focused on the larger values of β_j, all β_j are affected by the increase in variance and confidence interval width.
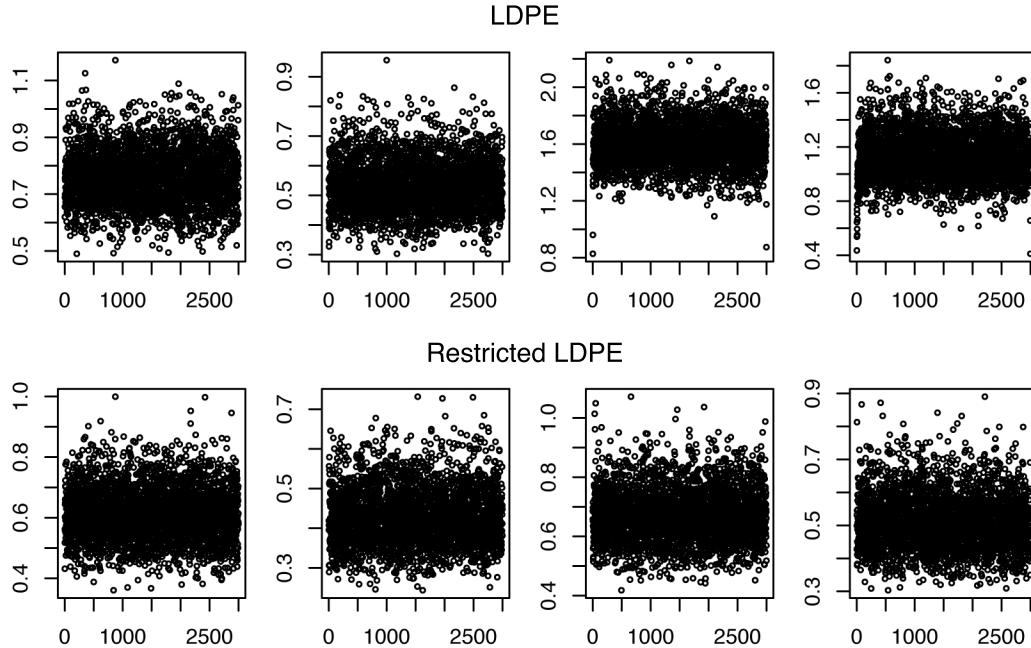

FIGURE 4. Efficiency (the ratio of the MSE's) of the LDPE and R-LDPE estimators versus the oracle estimator for each β_j.

	(A)	(B)	(C)	(D)
LDPE	0.7551	0.5232	1.5950	1.1169
R-LDPE	0.6086	0.4232	0.6656	0.5049

TABLE 5. Medians of the MSE ratios in Figure 4.

We may also consider the performance of LDPE as a point estimator. Table 5 and Figure 4 compare the MSEs of the LDPE and R-LDPE estimators β_j to that of the oracle estimator of β_j. This comparison is consistent with the comparison of the median width of confidence intervals in Table 4 and Figure 3 discussed earlier.

The Lasso and scaled Lasso estimators have larger biases for bigger values of β_j but perform very well for smaller values. On the other hand, the LDPE and the oracle estimator are not designed to be sparse and has very stable errors over the β_j. For the estimation of the entire vector $\boldsymbol{\beta}$ or its support, it is appropriate to compare a thresholded LDPE with the Lasso, the scaled Lasso, the scaled Lasso-LSE, and a matching thresholded oracle estimator. Hard thresholding was implemented: $\widehat{\beta}_j I\{|\widehat{\beta}_j| \leq \widehat{t}_j\}$ for the thresholded LDPE with $\widehat{t}_j = \widehat{\sigma}\tau_j \Phi^{-1}(1 - 1/(2p))$ and $\widehat{\beta}_j^{(o)} I\{|\widehat{\beta}_j^{(o)}| \leq \widehat{t}_j^{(o)}\}$ for the thresholded oracle with $\widehat{t}_j^{(o)} = \widehat{\sigma}^{(o)}\|\boldsymbol{z}_j^{(o)}\|_2^{-1}\Phi^{-1}(1 - 1/(2p))$, where $\{\widehat{\beta}_j^{(o)}, \widehat{\sigma}^{(o)}, \boldsymbol{z}_j^{(o)}\}$ are as in (45). Since $\beta_j \neq 0$ for all j, the comparison is confined to the ℓ_2 estimation error. Table 6 lists the mean, standard deviation, and median of the ℓ_2 loss of these five estimators over 100 replications. Of the five estimators, only the scaled Lasso, the scaled Lasso-LSE, and the thresholded LDPE are purely data-driven. The performance of the scaled Lasso-LSE, thresholded LDPE, and thresholded oracle are comparable and they always outperform the scaled Lasso. They also outperform the Lasso in cases (A), (B), and (C). In the hardest case, (D), which has both a high correlation between adjacent columns of $\boldsymbol{X}$ and a slower decay in β_j, the thresholded oracle slightly outperforms the Lasso and the Lasso sightly outperforms the scaled Lasso-LSE and thresholded LDPE. Generally, the ℓ_2 loss of the thresholded LDPE remains slightly above that of the scaled Lasso-LSE, which improves upon the scaled Lasso by reducing its bias. Note that our goal is not to find a better estimator for the entire vector $\boldsymbol{\beta}$ since quite a few versions of estimation optimality of regularized estimators have already been established. What we demonstrate here is that the cost of removing the bias with the LDPE, and thus giving up shrinkage, is small.

5. Discussion

We have developed the LDPE method of constructing $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ for the individual regression coefficients and estimators for their finite dimensional covariance structure. Under proper conditions on $\boldsymbol{X}$ and $\boldsymbol{\beta}$, we have proven the asymptotic unbiasedness and normality of the finite-dimensional distribution functions of these estimators and the consistency of their estimated covariances. Thus, LDPE yields an approximate Gaussian sequence as in (25), also called raw LDPE, which allows one to assess the level of significance of each unknown coefficient β_j without the uniform signal strength assumption (26), compared with the existing variable selection approach. The proposed method applies to making inference about a preconceived low-dimensional parameter, an interesting practical problem and a primary goal of this paper. It also applies to making inference about all regression coefficients via simultaneous interval estimation and correct selection of large and zero coefficients in the presence of many small coefficients.

The raw LDPE estimator is not sparse, but it can be thresholded to take advantage of the sparsity of $\boldsymbol{\beta}$, and the sampling distribution of the thresholded LDPE can still be bounded

		Estimator				
		Lasso	scLasso	scaled Lasso-LSE	T-oracle	T-LDPE
(A)	mean	0.8470	1.2706	0.3288	0.3624	0.3621
	sd	0.1076	0.2393	0.1465	0.0908	0.1884
	median	0.8252	1.2131	0.3042	0.3577	0.3312
(B)	mean	0.9937	1.5837	0.7586	0.5658	0.7969
	sd	0.1214	0.2624	0.2976	0.0615	0.3873
	median	0.9820	1.5560	0.6219	0.5675	0.6983
(C)	mean	0.8836	1.2411	0.4817	0.6803	0.5337
	sd	0.1402	0.2208	0.2083	0.2843	0.2164
	median	0.8702	1.2295	0.4343	0.6338	0.4642
(D)	mean	1.0775	1.6303	1.0102	0.9274	1.2627
	sd	0.1437	0.2381	0.3572	0.2342	0.5576
	median	1.0570	1.6389	0.9216	0.8716	1.1011

TABLE 6. Summary statistics for the ℓ_2 loss of five estimators of $\boldsymbol{\beta}$: the Lasso, the scaled Lasso, the scaled Lasso-LSE, the thresholded oracle estimator (T-oracle), and the thresholded LDPE (T-LDPE)

.

based on the approximate distribution of the raw LDPE. A thresholded LDPE is proven to attain ℓ_2 rate optimality for the estimation of an entire sparse $\boldsymbol{\beta}$.

The focus of this paper is interval estimation and hypothesis testing without the uniform signal strength condition. Another important problem is prediction. Since prediction at a design point $\boldsymbol{a}$ is equivalent to the estimation of the "contrast" $\boldsymbol{a}^T\boldsymbol{\beta}$, with possibly large $\|\boldsymbol{a}\|_0$, the implication of LDPE on prediction is an interesting future research direction.

We use the Lasso to provide a relaxation of the projection of $\boldsymbol{x}_j$ to $\boldsymbol{x}_j^\perp$. This choice is primarily due to our familiarity with the computation of the Lasso and the readily available scaled Lasso method of choosing a penalty level. We have also considered some other methods of relaxing the projection. Among these other methods, a particularly interesting one is the following constrained minimization of the variance of the noise term in (5):

$$(46) \qquad \boldsymbol{z}_j = \arg\min_{\boldsymbol{z}} \left\{ \|\boldsymbol{z}\|_2^2 : |\boldsymbol{z}_j^T \boldsymbol{x}_j| = n, \max_{k \neq j} |\boldsymbol{z}_j^T \boldsymbol{x}_k / n| \leq \lambda_j' \right\}.$$

Similar to the Lasso in (9), (46) is a quadratic programme. The Lasso solution (9) is feasible in (46) with $\lambda_j n / |\boldsymbol{z}_j^T \boldsymbol{x}_j| = \lambda_j'$. Our results on these and other extensions of our ideas and methods will be presented in a forthcoming paper.

6. APPENDIX

Proof of Proposition 1. (i) For $\widehat{\boldsymbol{\gamma}}_j(\lambda) = 0$, $\|\boldsymbol{z}_j(\lambda)\|_2 = \|\boldsymbol{x}_j\|_2 = \sqrt{n}$ and $\eta_j(\lambda) = \max_{k \neq j} |\boldsymbol{x}_k^T \boldsymbol{x}_j| / \sqrt{n}$ do not depend on λ. Consider $\widehat{\boldsymbol{\gamma}}_j(\lambda) \neq 0$. Since $\widehat{\boldsymbol{\gamma}}_j(\lambda)$ is continuous and piecewise linear in λ, it suffices to consider a fixed open interval $\lambda \in I_0$ in which

$s = \operatorname{sgn}(\widehat{\gamma}_j(\lambda))$ do not change with λ. Let $A = \{k \neq j : s_k \neq 0\}$, and let $\boldsymbol{Q}_A$ be the projection operator $\boldsymbol{b} \to \boldsymbol{b}_A$. It follows from the Karush-Kuhn-Tucker conditions for the Lasso that

$$\boldsymbol{X}_A^T \boldsymbol{z}_j(\lambda) = \boldsymbol{X}_A^T\{\boldsymbol{x}_j - \boldsymbol{X}_A \boldsymbol{Q}_A \widehat{\gamma}_j(\lambda)\}/n = \boldsymbol{X}_A^T\{\boldsymbol{x}_j - \boldsymbol{X}_{-j}\widehat{\gamma}_j(\lambda)\}/n = \lambda \boldsymbol{s}_A.$$

This gives $(\partial/\partial\lambda)\boldsymbol{Q}_A\widehat{\gamma}_j(\lambda) = -(\boldsymbol{X}_A^T \boldsymbol{X}_A/n)^{-1}\boldsymbol{s}_A$ for all $\lambda \in I_0$. It follows that

$$
\begin{aligned}
(\partial/\partial\lambda)\|\boldsymbol{z}_j(\lambda)\|_2^2 &= (\partial/\partial\lambda)\|\boldsymbol{x}_j - \boldsymbol{X}_A\boldsymbol{Q}_A\widehat{\gamma}_j(\lambda)\|_2^2 \\
&= -2\{(\partial/\partial\lambda)\boldsymbol{Q}_A\widehat{\gamma}_j(\lambda)\}^T\boldsymbol{X}_A^T(\boldsymbol{x}_j - \boldsymbol{X}_A\boldsymbol{Q}_A\widehat{\gamma}_j(\lambda)) \\
&= 2\{(\boldsymbol{X}_A^T\boldsymbol{X}_A/n)^{-1}\boldsymbol{s}_A\}^T\boldsymbol{X}_A^T\boldsymbol{z}_j(\lambda) = (2/\lambda)\|\boldsymbol{P}_A\boldsymbol{z}_j(\lambda)\|_2^2,
\end{aligned}
$$

where $\boldsymbol{P}_A = \boldsymbol{X}_A(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^T$ is the projection to the column space of $\boldsymbol{X}_A$. Thus, $\|\boldsymbol{z}_j(\lambda)\|_2$ is nondecreasing in λ. Since $\|\boldsymbol{s}\|_\infty = 1$, $\eta_j(\lambda) = n\lambda/\|\boldsymbol{z}_j(\lambda)\|_2$, so that

$$(\lambda^3/2)(\partial/\partial\lambda)\{\eta_j(\lambda)/n\}^{-2} = (\lambda^3/2)(\partial/\partial\lambda)\{\lambda^{-2}\|\boldsymbol{z}_j(\lambda)\|_2^2\} = \|\boldsymbol{P}_A\boldsymbol{z}_j(\lambda)\|_2^2 - \|\boldsymbol{z}_j(\lambda)\|_2^2 \leq 0.$$

Thus, $\eta_j(\lambda)$ is nondecreasing in λ. Since $\widehat{\sigma}_j(\lambda)$ is the solution of $\|\boldsymbol{z}_j(\lambda\sigma)\|_2 = \sigma\sqrt{n}$, it is also a solution of $\eta_j(\sigma\lambda) = \lambda\sqrt{n}$. For $\sigma < \widehat{\sigma}_j(\lambda)$, $\eta_j(\sigma\lambda) \leq \lambda\sqrt{n}$, so $\|\boldsymbol{z}_j(\sigma\lambda)\|_2 \geq \sigma\sqrt{n}$. Thus, since smaller λ gives smaller $\|\boldsymbol{z}_j(\lambda\sigma)\|_2$, $\widehat{\sigma}_j(\lambda)$ is also nondecreasing in λ. Since $\boldsymbol{x}_j^T\boldsymbol{z}_j(\lambda) = \|\boldsymbol{z}_j(\lambda)\|_2^2 + \{\boldsymbol{X}_{-j}\widehat{\gamma}_j(\lambda)\}^T\boldsymbol{z}_j(\lambda) = \|\boldsymbol{z}_j(\lambda)\|_2^2 + \lambda\|\widehat{\gamma}_j(\lambda)\|_1$, we also have $\tau_j(\lambda) \leq 1/\|\boldsymbol{z}_j(\lambda)\|_2$.

(ii) Since the Lasso path $\widehat{\gamma}_j(\lambda)$ is continuos in λ and $\eta_j(\lambda)$ is nondecreasing, the range of $\eta_j(\lambda)$ is an interval. Within the interior of this interval, $\widehat{\gamma}_j(\lambda) \neq 0$ and $\eta_j(\lambda) = n\lambda/\|\boldsymbol{z}_j(\lambda)\|_2$. We have shown in the proof of (i) that $\widehat{\sigma}_j(t)$ is a solution of $\eta_j(\sigma t) = t\sqrt{n}$ when $t\sqrt{n}$ is in the range of $\eta_j(\lambda)$. If $\eta_j(\lambda) < t\sqrt{n}$ for all λ, then $\widehat{\sigma}_j(t) = \|\boldsymbol{x}_j\|_2/\sqrt{n} = 1$ is attained at $\widehat{\gamma}_j(\infty) = 0$. If $\eta_j(\lambda) > t\sqrt{n}$ for all λ, then $\widehat{\sigma}_j(t) = 0$. This gives (15). The upper bounds follow for η_j^* and τ_j.

It remains to verify the last assertion of part (ii) for $\boldsymbol{z}_j(0) = 0$. Consider vectors $\boldsymbol{b}$ with $\|\boldsymbol{b} - \widehat{\gamma}_j(0+)\|_1 \leq \epsilon$ and the loss function in (14) at $\{t\boldsymbol{b}, \sigma\}$ with $0 \leq t \leq 1$. Since $\boldsymbol{X}_{-j}\widehat{\gamma}(0+) = \boldsymbol{x}_j$ and $\|\boldsymbol{x}_j\|_2 = \sqrt{n}$, the minimum of the loss function over σ is approximately

$$\min_\sigma\left\{\|\boldsymbol{x}_j - t\boldsymbol{X}_{-j}\boldsymbol{b}\|_2^2/(2n\sigma) + \sigma/2 + t\lambda\|\boldsymbol{b}\|_1\right\} = \{1 - t + O(\epsilon)\} + t\lambda\{\|\widehat{\gamma}(0+)\|_1 + O(\epsilon)\}.$$

When $\lambda\|\widehat{\gamma}_j(0+)\|_1 > 1$, the minimum of the above expression is attained at $t \approx 0$ for sufficiently small ϵ. This gives $\widehat{\sigma}_j(\lambda) > 0$. Conversely, when $\lambda\|\widehat{\gamma}_j(0+)\|_1 < 1$, the optimal t for $\widehat{\gamma}(\lambda)$ with very small λ is $t \approx 1$, so that by the joint convexity of the loss function, $\widehat{\sigma}_j(\lambda) = 0$. Since, by (15), $\eta_j(0+) = \inf\{\lambda\sqrt{n} : \widehat{\sigma}_j(\lambda) > 0\}$, the relationship between $\eta_j(0+)$ and the ℓ_1 minimization problem follows.

(iii) Let $\boldsymbol{\gamma}_j$ be the solution of $\boldsymbol{x}_j = \boldsymbol{X}_{-j}\boldsymbol{\gamma}_j$ with the shortest $\|\boldsymbol{\gamma}_j\|_1$, and $\lambda = 1/\|\boldsymbol{\gamma}_j\|_1$. Let $\boldsymbol{\beta}_{-j} = s\lambda_{univ}\lambda\boldsymbol{\gamma}_j$, and $\beta_j = -s\lambda_{univ}\lambda$. Then, $\boldsymbol{X}\boldsymbol{\beta} = 0$ and $\sum_{j=1}^p \min(|\beta_j|/\lambda_{univ}, 1) \leq s + 1$. It follows that for the optimal $\boldsymbol{\delta}$,

$$4C_0 s\lambda_{univ}^2 \geq 2\|\boldsymbol{\beta} - \boldsymbol{\delta}\|_2^2 + 2\|\boldsymbol{\delta}\|_2^2 \geq \|\boldsymbol{\beta}\|_2^2 \geq |\beta_j|^2 = (s\lambda_{univ}\lambda)^2.$$

Taking $s = a_0 n/(\log p)$ gives $\lambda^2 \leq (4C_0/a_0)(\log p)/n$. Thus, by part (ii), $\max_j \eta_j^2(0+) \leq \lambda^2 n \leq (4C_0/a_0)\log p$. This implies the upper bound for $\max_{j \leq p} \eta_j^*$ by Step 1. $\square$

Proof of Theorem 1. The error decomposition in (5) and (6) implies

$$\left|\tau_j^{-1}(\widehat{\beta}_j - \beta_j) - \boldsymbol{z}_j^T \boldsymbol{\varepsilon}/\|\boldsymbol{z}_j\|_2\right| \leq \left(\max_{k \neq j} |\boldsymbol{z}_j^T \boldsymbol{x}_k|/\|\boldsymbol{z}_j\|_2\right)\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 = \eta_j \|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1.$$

This and (20) yield (22). When $\left|\tau_j^{-1}(\widehat{\beta}_j - \beta_j) - \boldsymbol{z}_j^T \boldsymbol{\varepsilon}/\|\boldsymbol{z}_j\|_2\right| \leq \sigma^* \epsilon_n'$ and $|\widehat{\sigma}/\sigma^* - 1| \leq \epsilon_n''$, $\tau_j^{-1}|\widehat{\beta}_j - \beta_j| \geq \widehat{\sigma} t$ implies $|\boldsymbol{z}_j^T \boldsymbol{\varepsilon}|/\|\boldsymbol{z}_j\|_2 \geq \widehat{\sigma} t - \sigma^* \epsilon_n' \geq \sigma^*\{(1 - \epsilon_n'')t - \epsilon_n'\}$. Since $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{z}_j$ depends on $\boldsymbol{X}$ only, $\boldsymbol{z}_j^T \boldsymbol{\varepsilon}/(\|\boldsymbol{z}_j\|_2 \sigma^*) \sim \sqrt{n}\varepsilon_1/\|\boldsymbol{\varepsilon}\|_2$. Thus, for $x \geq 1$,

$$P\{|\boldsymbol{z}_j^T \boldsymbol{\varepsilon}|/\|\boldsymbol{z}_j\|_2 \geq \sigma^* x\} = P\{(n - x^2)\varepsilon_1^2 \geq x^2(\varepsilon_2^2 + \cdots + \varepsilon_n^2)\} \leq 2\Phi_n(-x).$$

The same argument also implies (24) with fixed m, since $\max(\epsilon_n', \epsilon_n'') \to 0$ and $\boldsymbol{V}$ in (17) is the approximate covariance between $\widehat{\beta}_j$ and $\widehat{\beta}_k$. $\square$

Proof of Theorem 2. Since (22) is uniform in $\epsilon \in [\alpha_0/p^2, 1]$, (27) and (28) follow directly. By Lemma 1 of [SZ11], $2\Phi_n(-\sqrt{n\{\exp(2t^2/(n-1)) - 1\}}) \leq (\pi^{-1/2} + o(1))e^{-t^2}/t$ as $\min(n, t) \to \infty$. When $s \log p = o(n^{1/2})$, $\epsilon_n' = o(1)$ and $\epsilon_n'' = o(n^{-1/2})$. Let $t = \sqrt{2 \log(p/\alpha)} + c_0$. Since $\log(p/\alpha) \ll n$ and α is fixed,

$$-(1 - \epsilon_n'')t + \epsilon_n' = -t + o(1) = -\sqrt{n\{\exp(2t^2/(n-1)) - 1\}} - c_0 + o(1).$$

Thus, the right-hand side of (28) is no greater than α in the limit. This and a similar inequality with the true σ yields (29). $\square$

The following lemma, needed in the proof of Theorem 3, controls the loss of a perturbed soft threshold estimator. It extends Lemma 8.3 of [Joh98] and Lemma 6.2 of [Zha05].

Lemma 1. *Let $s_t(x) = \text{sgn}(x)(|x| - t)_+$, $z = \mu + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$. Suppose that for certain constants t and Δ, $|\widehat{z} - z| + |\widehat{t} - t| \leq \Delta \leq t$ and $\widehat{t} > t + |\widehat{z} - z|$ in an event Ω. Then,*

$$\begin{aligned}
&E\{s_{\widehat{t}}(\widehat{z}) - \mu\}^2 I_\Omega \\
\leq\ & \min\Big\{2E(\varepsilon - t)_+^2 + \mu^2, \sigma^2 + (t + \Delta)^2\Big\} + \Delta\Big\{2E(\varepsilon - t)_+ + 3\Delta P(\varepsilon > t)\Big\} \\
\leq\ & \min\Big\{\mu^2, \sigma^2 + (t + \Delta)^2\Big\} + \varphi(t/\sigma)\Big\{4\sigma^5/t^3 + 2\Delta\sigma^3/t^2 + 3\Delta^2\sigma/t\Big\},
\end{aligned}$$

where $\varphi(x)$ and $\Phi(x)$ are the $N(0, 1)$ density and distribution functions.

Proof. Assume without loss of generality that $\mu > 0$. Let

$$f_{t,\Delta}(z, \mu) = |-(z + t)_- - \mu|I_{\{z < 0\}} + |(z - t - \Delta)_+ - \mu|I_{\{z > 0\}}.$$

By assumption $z + t \leq \widehat{z} + \widehat{t}$ and $z - t - \Delta \leq \widehat{z} - \widehat{t} \leq z - t$. Since $s_t(z) = (z - t)_+ - (z + t)_-$,

$$\begin{aligned}
|s_{\widehat{t}}(\widehat{z}) - \mu|^2 I_\Omega &\leq |-(z + t)_- - \mu|^2 I_{\{z < 0\}} + \{|(z - t - \Delta)_+ - \mu| + \Delta I_{\{z - t > \mu\}}\}^2 I_{\{z > 0\}} \\
&\leq f_{t,\Delta}^2(z, \mu) + \Delta\{2(z - t - \mu + \Delta) + \Delta\}I_{\{z > t + \mu\}}.
\end{aligned}$$

Since $(\partial/\partial\mu)f_{t,\Delta}^2(\varepsilon + \mu, \mu) = 2\mu I_{\{-t < \varepsilon + \mu < t + \Delta\}}$, $E f_{t,\Delta}^2(\varepsilon + \mu, \mu) \leq E f_{t,\Delta}^2(\varepsilon, 0) + \int_0^\mu 2x\,dx$ and $E f_{t,\Delta}^2(\varepsilon + \mu, \mu) \uparrow \sigma^2 + (t + \Delta)^2$. Since $E f_{t,\Delta}^2(\varepsilon, 0) \leq E f_{t,0}^2(\varepsilon, 0) = 2E(\varepsilon - t)_+^2$, we have

$$E f_{t,\Delta}^2(\varepsilon + \mu, \mu) \leq \min\Big\{2E(\varepsilon - t)_+^2 + \mu^2, \sigma^2 + (t + \Delta)^2\Big\}.$$

Thus, the first inequality follows from $E\{2(\varepsilon - t + \Delta) + \Delta\}I_{\{\varepsilon > t\}} = 2E(\varepsilon - t)_+ + 3\Delta P(\varepsilon > t)$, and the second from $E(\varepsilon - t)_+^k \le \sigma^k \varphi(t/\sigma) \int_0^\infty x^k e^{-xt/\sigma} dx = k!\sigma^{2k+1}\varphi(t/\sigma)/t^{k+1}$. $\square$

Proof of Theorem 3. We first prove the equivalence of the following two statements:

(47) $(\widehat{\sigma}/\sigma) \vee (\sigma/\widehat{\sigma}) - 1 + \epsilon_n' \sigma^*/(\widehat{\sigma} \wedge \sigma) \le \{1 - (\widehat{\sigma}/\sigma - 1)_+\}c_n;$

(48) $\widetilde{t}_j + \epsilon_n'(\sigma^*/\sigma)\widetilde{t}_j \le \widehat{t}_j = (1 + c_n)(\widehat{\sigma}/\sigma)\widetilde{t}_j,\ \widehat{t}_j - \widetilde{t}_j + \epsilon_n'(\sigma^*/\sigma)\widetilde{t}_j \le 2c_n\widetilde{t}_j.$

For $\widehat{\sigma} \le \sigma$, (47) is equivalent to $\sigma/\widehat{\sigma} - 1 + \epsilon_n'\sigma^*/\widehat{\sigma} \le c_n$, and (48) to $\widetilde{t}_j + \epsilon_n'(\sigma^*/\sigma)\widetilde{t}_j \le (1 + c_n)(\widehat{\sigma}/\sigma)\widetilde{t}_j$. For $\widehat{\sigma} > \sigma$, (47) is equivalent to $\widehat{\sigma}/\sigma - 1 + \epsilon_n'\sigma^*/\sigma \le (2 - \widehat{\sigma}/\sigma)c_n$, and (48) to $(1 + c_n)(\widehat{\sigma}/\sigma - 1)\widetilde{t}_j + \epsilon_n'(\sigma^*/\sigma)\widetilde{t}_j \le c_n\widetilde{t}_j$. After canceling $\widetilde{t}_j$ and some algebra, we observe that (47) and (48) are equivalent in both cases.

Let $\widetilde{\varepsilon}_j = \tau_j \boldsymbol{z}_j^T \boldsymbol{\varepsilon}/\|\boldsymbol{z}_j\|_2 \sim N(0, \tau_j^2\sigma^2)$, $\widetilde{\beta}_j = \beta_j + \widetilde{\varepsilon}_j$, and

$$\Omega_n = \{|\widetilde{\beta}_j - \widehat{\beta}_j| \le \epsilon_n'(\sigma^*/\sigma)\widetilde{t}_j,\ (48) \text{ holds},\ \forall j \le p\}.$$

As in the proof of Theorem 1, $|\widetilde{\beta}_j - \widehat{\beta}_j| \le \tau_j\eta_j\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1$. Since $\max_{j \le p} \eta_j C_1 s/\sqrt{n} \le \epsilon_n'$, we have $|\widetilde{\beta}_j - \widehat{\beta}_j| \le \epsilon_n'(\sigma^*/\sigma)\widetilde{t}_j$ when $\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 \le C_1 s\sigma^*L_0/\sqrt{n}$. Thus, $P\{\Omega_n\} \ge 1 - 3\epsilon$ by (20) and (31). Consider the event Ω_n in the rest of the proof, so that (48) gives

$$\widehat{t}_j \ge \widetilde{t}_j + |\widehat{\beta}_j - \widetilde{\beta}_j|,\ |\widehat{\beta}_j - \widetilde{\beta}_j| + |\widehat{t}_j - \widetilde{t}_j| \le 2c_n\widetilde{t}_j.$$

Since $\widetilde{\varepsilon}_j \sim N(0, \tau_j^2\sigma^2)$ and $\widetilde{t}_j/(\tau_j\sigma) = L_0$, it follows from Lemma 1 with $\Delta = 2c_n\widetilde{t}_j$ that

$$E\|\widehat{\boldsymbol{\beta}}^{(thr)} - \boldsymbol{\beta}\|_2^2 I_{\Omega_n} \le \sum_{j=1}^p \left[\min\left\{\beta_j^2, \tau_j^2\sigma^2 + \widetilde{t}_j^2(1 + 2c_n)^2\right\} \right.$$
$$\left. + \varphi(L_0)\left\{4\tau_j^2\sigma^2/L_0^3 + 4c_n\tau_j^2\sigma^2/L_0 + 12c_n^2\tau_j^2\sigma^2 L_0\right\}\right].$$

This gives (32) since $\varphi(L_0) = \epsilon/p$.

Since $\widehat{t}_j \ge \widetilde{t}_j + |\widehat{\beta}_j - \widetilde{\beta}_j|$, $|\widehat{\beta}_j| > \widehat{t}_j$ implies $|\widetilde{\varepsilon}_j| > \widetilde{t}_j$ for $\beta_j = 0$. Since $|\widehat{\beta}_j - \widetilde{\beta}_j| + |\widehat{t}_j - \widetilde{t}_j| \le 2c_n\widetilde{t}_j$, $|\widehat{\beta}_j| \le \widehat{t}_j$ implies $|\widetilde{\varepsilon}_j| > \widetilde{t}_j$ for $|\beta_j| > (2 + 2c_n)\widetilde{t}_j$. Thus,

$$P\left(\{j : |\beta_j| > (2 + 2c_n)\widetilde{t}_j\} \subseteq \widehat{S}^{(thr)} \subseteq \{j : \beta_j \ne 0\}\right) \ge P\{\Omega_n^c\} + pP\{|\widetilde{\varepsilon}_j| > \widetilde{t}_j\}.$$

Hence, (33) follows from $P\{|\widetilde{\varepsilon}_j| > \widetilde{t}_j\} = 2\Phi(-L_0) \le \alpha/p$. $\square$

Proof of Theorem 4. Due to the scale invariance of (10) and (11), we assume $\sigma = 1$ without loss of generality. Let $\boldsymbol{h} = \widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}$ and $z^* = \|\boldsymbol{X}^T\boldsymbol{\varepsilon}/n\|_\infty/\sigma^*$. By (19), we have $\|\boldsymbol{\beta}_{S^c}\|_1 \le \lambda_{univ}s$ and $|S| \le s$. Let $\{\mu_*, C_1, C_2\}$ be as in (36) and define

$$\xi' = (1 - \nu_0)(\xi + 1) - 1,\ \tau_*^2 = (\lambda_0/\sigma^*)(\xi' + 1)\max\left\{\frac{\lambda_{univ}s}{\nu_0}, \frac{\sigma^*\lambda_0 s}{2(1 - \nu_0)\kappa^2(\xi, S)}\right\}.$$

In the event $z^* \le (1 - \tau_*^2)\lambda_0(\xi' - 1)/(\xi' + 1)$, Theorem 2 of [SZ11] gives

(49) $\max\{1 - \widehat{\sigma}/\sigma^*, 1 - \sigma^*/\widehat{\sigma}\} \le \tau_*^2,\ \|\boldsymbol{h}\|_1 \le (\sigma^*/\lambda_0)\tau_*^2/(1 - \tau_*^2)$

due to $\xi = (\xi' + \nu_0)/(1 - \nu_0)$. Since $\kappa^2(\xi, S) \geq c_0$, in the event $\sigma^* > 1/2$,

$$\frac{\tau_*^2}{\xi + 1} \leq \max\left\{ \frac{2\lambda_0\lambda_{univ}s}{\nu_0/(1 - \nu_0)}, \frac{\lambda_0^2 s}{2c_0} \right\} \leq \max\left\{ \frac{(2/A)\lambda_0^2 s}{\nu_0/(1 - \nu_0)}, \frac{\lambda_0^2 s}{2c_0} \right\} = \frac{\tau_0^2\lambda_0^2 s}{A^2(1 + \xi)\mu_*}.$$

Since $(2s/n)\log(p/\epsilon) \leq \mu_*$, this gives $\tau_*^2 \leq (\tau_0^2/\mu_*)(2s/n)\log(p/\epsilon) = C_2(2s/n)\log(p/\epsilon) \leq \tau_0^2$ in (49). In addition, (49) gives

$$\|\boldsymbol{h}\|_1 \leq \frac{\sigma^*\tau_*^2}{\lambda_0(1 - \tau_*^2)} \leq \frac{\sigma^* C_2 s\lambda_0^2}{A^2\lambda_0(1 - \tau_0^2)} \leq \sigma^* C_1 s\sqrt{(2/n)\log(p/\epsilon)}.$$

Thus, the union of the events in (20) and (21) has at most probability

$$p_n = P\{z^* \geq (1 - \tau_*^2)\lambda_0(\xi' - 1)/(\xi' + 1) \text{ or } \sigma^* < 1/2\}.$$

We prove below $p_n \leq \epsilon$. Since $\xi' + 1 = (1 - \nu_0)(\xi + 1)$, we have

$$(1 - \tau_*^2)\lambda_0(\xi' - 1)/(\xi' + 1) \geq (1 - \tau_0^2)\lambda_0\{\xi - (1 + \nu_0)/(1 - \nu_0)\}/(\xi + 1) = \lambda_0/A.$$

Thus, with $p_n' = P\{\sigma^* < 1/2\} = P\{\chi_n^2 < n/4\}$, Theorem 2 of [SZ11] gives

$$p_n \leq (1 + \epsilon_{n-1})\epsilon/\{\pi\log(p/\epsilon)\}^{1/2} + p_n'$$

with $\epsilon_m = \{2/(m-1)\}^{1/2}\Gamma((m+1)/2)/\Gamma(m/2)$. Since $p_n' = \int_0^{n/4} t^{n/2-1}e^{-t/2}dt/\{2^{n/2}\Gamma(n/2)\}$ and $(n/2)\log(4/e) \geq \log(p/\epsilon)$, the Stirling formula gives

$$p_n' \leq \frac{(n/8)^{n/2}}{\Gamma(n/2 + 1)} \leq \frac{(n/8)^{n/2}}{e^{-n/2}(n/2)^{n/2}\sqrt{2\pi}} \leq (e/4)^{n/2}/\sqrt{2\pi} \leq \epsilon/(p\sqrt{2\pi}).$$

Since $\epsilon_{n-1} \leq \sqrt{\pi/2}$ for $n \geq 3$, $p_n \leq \epsilon(1 + \sqrt{\pi/2})/\sqrt{\pi\log p} + \epsilon/(p\sqrt{2\pi}) \leq \epsilon$ for $p \geq 7$. This proves Theorem 4 (i) for the $\{\mu^*, C_1, C_2\}$ in (36). The proof of Theorem 4 (ii) follows from Theorem 3 of [SZ11] in the same way with somewhat different constants. We omit the details. $\square$

Proof of Proposition 2. For any $\boldsymbol{u} \in \mathscr{C}(\xi, S)$,

$$\frac{\|\boldsymbol{X}\boldsymbol{u}\|_2^2|S|}{n\|\boldsymbol{u}_S\|_1^2} \geq \frac{\boldsymbol{u}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{u}}{\|\boldsymbol{u}\|_2^2} - \frac{\boldsymbol{u}^T(\boldsymbol{X}^T\boldsymbol{X}/n - \widehat{\boldsymbol{\Sigma}})\boldsymbol{u}}{\|\boldsymbol{u}_S\|_1^2/|S|} \geq c_* - \frac{\lambda_1\|\boldsymbol{u}\|_1^2}{\|\boldsymbol{u}_S\|_1^2/|S|}$$

which is no smaller than $c_* - |S|\lambda_1(1 + \xi)^2 \geq c_*/2$.

Now assume the additional condition that $s\lambda_1(1 + K) \leq c_*/2$. Consider $|S| \geq 1$ since the case of empty S is trivial. Since $s\lambda_1(1 + K) + \lambda_1 \leq c_*/2$, we have $\phi_-(m, S) \geq \phi_{\min}(\widehat{\boldsymbol{\Sigma}}) - \lambda_1(m + S) \geq c_* - \{|S|\lambda_1(1 + K) + \lambda_1\} \geq c_*/2$. Similarly, $\phi_+(m, S) \leq c_* + c_*/2$. Thus, $\phi_+(m, S)\xi^2/\kappa^2(\xi, S) \leq \xi^2(c^* + c_*/2)/(c_*/2) = K$. $\square$

Proof of Theorem 5. We first prove the bounds for $\kappa(\xi, S)$ and $\xi^2\phi_+(m, S)/\kappa^2(\xi, S)$ in Remark 5. Let $\{\delta_2, \delta_0, \delta_1, \mathscr{X}'_{n,p}, K, k, \ell\}$ be as in Remark 5. Suppose $\mathscr{X}'_{n,p}$ happens. Let S be a subset of $\{1, \ldots, p\}$ with $|S| = k$, $\boldsymbol{u}$ a vector in $\mathscr{C}(\xi, S)$ with $\|\boldsymbol{u}_S\|_1 = 1$, A the union of S and the set of the indices of the ℓ largest $|u_j|$ with $j \notin S$, and $\boldsymbol{w}$ a unit vector in $\mathbb{R}^n$ with $\boldsymbol{w}^T\boldsymbol{X}_A\boldsymbol{u}_A = \|\boldsymbol{X}_A\boldsymbol{u}_A\|_2$. We pick a $\boldsymbol{u}$ satisfying

$$\kappa(\xi, S) = (k/n)^{1/2}\|\boldsymbol{X}\boldsymbol{u}\|_2 \geq (k/n)^{1/2}\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{u} = (k/n)^{1/2}\big(\|\boldsymbol{X}_A\boldsymbol{u}_A\|_2 + \boldsymbol{w}^T\boldsymbol{X}_{A^c}\boldsymbol{u}_{A^c}\big).$$

Let $u_* = \|\boldsymbol{u}_{A^c}\|_\infty$. Since $\boldsymbol{u} \in \mathscr{C}(\xi, S)$, $\|\boldsymbol{u}_{A^c}\|_1 \leq \|\boldsymbol{u}_{S^c}\|_1 - \ell u_* \leq \xi - \ell u_*$. Let $\boldsymbol{u}_{A^c}$ be the minimizer of $\boldsymbol{w}^T \boldsymbol{X}_{A^c} \boldsymbol{u}_{A^c}$ subject to $\|\boldsymbol{u}_{A^c}\|_1 \leq \xi - \ell u_*$ and $B_0 = \{j \notin A : u_j \neq 0\}$. Then, B_0 is the index set of certain $|B_0|$ largest $|\boldsymbol{w}^T \boldsymbol{x}_j|$ with $j \in A^c$ and $|u_j| = u_*$ for $j \in B_0$ with one possible exception. Since $1 = \|\boldsymbol{u}_S\|_1 \leq k^{1/2}\|\boldsymbol{u}_A\|_2$ and $\|\boldsymbol{w}\|_2 = 1$, (35) gives $(k/n)^{1/2}\|\boldsymbol{X}_A \boldsymbol{u}_A\|_2 \geq \sqrt{c_*(1-\delta_1)}$ and $\|\boldsymbol{w}^T \boldsymbol{X}_B/n^{1/2}\|_2 \leq \sqrt{c^*(1+\delta_1)}$ for all $B \subseteq B_0$ with $|B| \leq 4\ell$. For $|B_0| \geq 4\ell$, let B_1 be the index set of certain 4ℓ largest $|\boldsymbol{w}^T \boldsymbol{x}_j|$ with $j \in B_0$, so that $|\boldsymbol{w}^T \boldsymbol{x}_j|^2/n \leq c^*(1+\delta_1)/(4\ell)$ for $j \in B_0 \setminus B_1$. For $|B_0| \leq 4\ell$, $(\boldsymbol{w}^T \boldsymbol{X}_{A^c} \boldsymbol{u}_{A^c})^2/n \leq c^*(1+\delta_1)\|\boldsymbol{u}_{A^c}\|_2^2 \leq c^*(1+\delta_1)u_*(\xi - u_*\ell) \leq c^*(1+\delta_1)\xi^2/(4\ell)$. For $|B_0| > 4\ell$, $|\boldsymbol{w}^T \boldsymbol{X}_{A^c} \boldsymbol{u}_{A^c}/n^{1/2}| \leq \sqrt{c^*(1+\delta_1)u_*^2 4\ell} + \|\boldsymbol{u}_{B_0 \setminus B_1}\|_1 \sqrt{c^*(1+\delta_1)/(4\ell)} = \sqrt{c^*(1+\delta_1)/(4\ell)}\|\boldsymbol{u}_{A^c}\|_1$. In either cases,

$$
\begin{aligned}
\kappa(\xi, S) &\geq \{c_*(1-\delta_1)\}^{1/2} - \{kc^*(1+\delta_1)/(4\ell)\}^{1/2}\xi \\
&= \{c_*(1-\delta_1)\}^{1/2}\big(1 - \{kK/(16\ell)\}^{1/2}\big) \geq \{c_*(1-\delta_1)\}^{1/2}/2.
\end{aligned}
$$

Since $m - 1 < K|S| \leq m$ implies $m \leq 4\ell$, we also have $\xi^2 \phi_+(m, S)/\kappa^2(\xi, S) \leq K$. Thus, the conditions on $\kappa(\xi, S)$ and $\phi_\pm(m, S)$ of $\mathscr{X}_{s,n,p}(c_*, \delta_1, \xi, K)$ hold in $\mathscr{X}'_{n,p}$.

By Proposition 1 (ii), the conditions on η_j and τ_j of $\mathscr{X}_{s,n,p}(c_*, \delta_1, \xi, K)$ hold when

$$
(50) \qquad \min_{j \leq p} \widehat{\sigma}_j^2(\lambda_0)/\sigma_j^2 \geq (1 + \kappa_0)^2/2, \quad \lambda_0 = (1 + \kappa_1)^{-1} 3\sqrt{(\log p)/n}.
$$

Let $\widetilde{\sigma}_j(\lambda)$ be the scaled Lasso estimator of the noise level in the regression model

$$
(51) \qquad \widetilde{\boldsymbol{x}}_j = \sum_{k \neq j} \widetilde{\gamma}_{jk} \boldsymbol{x}_k + \boldsymbol{\varepsilon}_j, \quad \widetilde{\gamma}_{jk} = -\sigma_j^2 \Theta_{jk} \|\widetilde{\boldsymbol{x}}_k\|_2/\sqrt{n}.
$$

Since the scaled Lasso is scale invariant, $\widehat{\sigma}_j(\lambda_0) = \widetilde{\sigma}(\lambda_0)\sqrt{n}/\|\widetilde{\boldsymbol{x}}_j\|_2$ by (43). Since $c_* \leq \sigma_j^2 = 1/\Theta_{jj} \leq 1$ by (44), (50) is a question about the consistency of the scaled Lasso estimator $\widetilde{\sigma}(\lambda_0)$ in the regression model (51).

Let $\delta_3 \in (0, 1)$ with $(1 - \delta_3)(1 + \delta_3)^{-1} = (1 + \kappa_0)^{1/2} 2^{-1/4}$ and

$$
\mathscr{X}''_{n,p} = \{\max_j |1 - \|\widetilde{\boldsymbol{x}}_j\|_2/\sqrt{n}| \leq \delta_3, \max_j |1 - \|\boldsymbol{\varepsilon}_j\|_2/(\sigma_j \sqrt{n})| \leq \delta_3\} \cap \mathscr{X}'_{n,p}.
$$

We have $P\{\mathscr{X}''_{n,p}\} \geq 1 - 2e^{-n\delta_2}$, taking a smaller δ_2 if necessary. Consider the event $\mathscr{X}''_{n,p}$. Since $(\Theta_{jk}, k \neq j)^T \in \mathscr{B}_1(s, \lambda_{univ})$ for all j, the coefficients $\widetilde{\gamma}_{jk}$ in (51) satisfy

$$
\sum_{k \neq j} \min\{|\widetilde{\gamma}_{jk}|/(\sigma_j \lambda_{univ}), 1\} \leq \sum_k \min\{(1 + \delta_3)|\Theta_{jk}|/\lambda_{univ}, 1\} \leq (1 + \delta_3)s.
$$

We treat λ_0 as $(1 + \kappa_1)^{-1} 3\sqrt{(\log p)/n} = A\sqrt{(2/n)\log(p^4)}$ with $A = (1 + \kappa_1)^{-1} 3/\sqrt{8} > 1$. By checking regularity conditions as in Remarks 4 and 5, the scaled Lasso error bound for noise estimation gives

$$
P\{\widetilde{\sigma}_j(\lambda_0)\sqrt{n}/\|\boldsymbol{\varepsilon}_j\|_2 \geq (1 - \delta_3)/(1 + \delta_3), \mathscr{X}''_{n,p}\} \leq 1/p^3.
$$

In the same event, $\widehat{\sigma}_j(\lambda_0)/\sigma_j = (\widetilde{\sigma}_j(\lambda_0)/\sigma_j)\sqrt{n}/\|\widetilde{\boldsymbol{x}}_j\|_2 \geq (\|\boldsymbol{\varepsilon}_j\|_2/\sigma)\|\widetilde{\boldsymbol{x}}_j\|_2^{-1}(1-\delta_3)/(1+\delta_3) \geq (1-\delta_3)^2/(1+\delta_3)^2 = (1 + \kappa_0)/\sqrt{2}$. This gives (50) in the intersection of these events and completes the proof. $\qquad\square$

REFERENCES

[Ant10]   A. Antoniadis, *Comments on: ℓ_1-penalization for mixture regression models*, Test **19** (2010), no. 2, 257–258.

[BBZ10]   R. Berk, L.B. Brown, and L. Zhao, *Statistical inference after model selection*, Journal of Quantitative Criminology **26** (2010), 217–236.

[BCW11]   Alexandre Belloni, Victor Chernozhukov, and Lie Wang, *Square-root lasso: Pivotal recovery of sparse signals via conic programming*, Biometrika **98** (2011), no. 4, 791–806.

[BL08]   Peter J. Bickel and Elizaveta Levina, *Regularized estimation of large covariance matrices*, Annals of Statistics **36** (2008), no. 1, 199–227.

[BRT09]   Peter Bickel, Yaacov Ritov, and Alexandre Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Annals of Statistics **37** (2009), no. 4, 1705–1732.

[BvdG11] Peter Bühlmann and Sara van de Geer, *Statistics for high-dimensional data: Methods, theory and applications*, Springer, New York, 2011.

[CDS01]   Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Review **43** (2001), 129–159.

[CT05]   Emmanuel J. Candes and Terence Tao, *Decoding by linear programming*, IEEE Trans. on Information Theory **51** (2005), 4203–4215.

[CT07]   E. Candes and T. Tao, *The dantzig selector: statistical estimation when p is much larger than n (with discussion)*, Annals of Statistics **35** (2007), 2313–2404.

[DJ94]   D. L. Donoho and I. Johnstone, *Minimax risk over ℓ_p–balls for ℓ_q–error*, Probability Theory and Related Fields **99** (1994), 277–303.

[DS01]   K. Davidson and S. Szarek, *Local operator theory, random matrices and banach spaces*, Handbook on the Geometry of Banach Spaces, vol. 1, 2001.

[FF93]   I.E. Frank and J.H. Friedman, *A statistical view of some chemometrics regression tools (with discussion)*, Technometrics **35** (1993), 109–148.

[FL01]   Jianqing Fan and Runze Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association **96** (2001), 1348–1360.

[FL08]   Jianqing Fan and Jinchi Lv, *Sure independence screening for ultrahigh dimensional feature space (with discussion)*, J. R. Statist. Soc. **B, 70** (2008), 849–911.

[FL10]   _____, *A selective overview of variable selection in high dimensional feature space*, Statistica Sinica **20** (2010), 101–148.

[FP04]   J. Fan and H. Peng, *On non-concave penalized likelihood with diverging number of parameters*, Annals of Statistics **32** (2004), 928–961.

[GR04]   E. Greenshtein and Y. Ritov, *Persistence in high–dimensional linear predictor selection and the virtue of overparametrization*, Bernoulli **10** (2004), 971–988.

[Gre06]   E. Greenshtein, *Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint*, Annals of Statistics **34** (2006), 2367–2386.

[HMZ08]   J. Huang, S. Ma, and C.-H. Zhang, *Adaptive lasso for sparse high-dimensional regression models*, Statistica Sinica **18** (2008), 1603–1618.

[HZ12]   Jian Huang and Cun-Hui Zhang, *Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications*, Journal of Machine Learning Research **13** (2012), 1809–1834.

[Joh98]   Iain Johnstone, *Gaussian estimation: Sequence and wavelet models*, 1998.

[KCO08]   Yongdai Kim, Hosik Choi, and Hee-Seok Oh, *Smoothly clipped absolute deviation on high dimensions*, Journal of American Statistical Association **103** (2008), 1665–1673.

[KLT11]   V. Koltchinskii, K. Lounici, and A. B. Tsybakov, *Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion*, The Annals of Statistics **39** (2011), 2302–2329.

[Kol09]   V. Koltchinskii, *The dantzig selector and sparsity oracle inequalities*, Bernoulli **15** (2009), 799–828.

[LM11]   E. Laber and S.A. Murphy, *Adaptive confidence intervals for the test error in classification (with discussion)*, Journal of the American Statistical Association **106** (2011), 904–913.

[LP06]   Hannes Leeb and Benedikt M. Potscher, *Can one estimate the conditional distribution of post-model-selection estimators?*, The Annals of Statistics **34** (2006), 2554–2591.

[MB06]   Nicolai Meinshausen and Peter Bühlmann, *High-dimensional graphs and variable selection with the lasso*, Annals of Statistics **34** (2006), 1436–1462.

[MB10]   ———, *Stability selection (with discussion)*, Journal of the Royal Statistical Society, B **72** (2010), 417–473.

[MY09]   N. Meinshausen and B. Yu, *Lasso-type recovery of sparse representations for high-dimensional data*, Annals of Statistics **37** (2009), 246–270.

[SBvdG10] N. Städler, P. Bühlmann, and S. van de Geer, *ℓ_1-penalization for mixture regression models (with discussion)*, Test **19** (2010), no. 2, 209–285.

[SZ10]   Tingni Sun and Cun-Hui Zhang, *Comments on: ℓ_1-penalization for mixture regression models*, Test **19** (2010), no. 2, 270–275.

[SZ11]   Tungni Sun and Cun-Hui Zhang, *Scaled sparse linear regression*, Tech. Report arXiv:1104.4595, arXiv, 2011.

[Tib96]   R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **58** (1996), 267–288.

[Tro06]   J. A. Tropp, *Just relax: convex programming methods for identifying sparse signals in noise*, IEEE Transactions on Information Theory **52** (2006), 1030–1051.

[vdGB09]   S. van de Geer and P. Bühlmann, *On the conditions used to prove oracle results for the lasso*, Electronic Journal of Statistics **3** (2009), 1360–1392.

[Wai09a]   M. J. Wainwright, *Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting*, IEEE Transactions on Information Theory **55** (2009), 5728–5741.

[Wai09b]   ———, *Sharp thresholds for noisy and high–dimensional recovery of sparsity using ℓ_1–constrained quadratic programming (lasso)*, IEEE Transactions on Information Theory **55** (2009), 2183–2202.

[YZ10]   Fei Ye and Cun-Hui Zhang, *Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls*, Journal of Machine Learning Research **11** (2010), 3481–3502.

[ZH08]   Cun-Hui Zhang and Jian Huang, *The sparsity and bias of the Lasso selection in high-dimensional linear regression*, Annals of Statistics **36** (2008), no. 4, 1567–1594.

[Zha05]   Cun-Hui Zhang, *General empirical bayes wavelet methods and exactly adaptive minimax estimation1*, The Annals of Statistics **33** (2005), 54–100.

[Zha09]   Tong Zhang, *Some sharp performance bounds for least squares regression with L_1 regularization*, Ann. Statist. **37** (2009), no. 5A, 2109–2144.

[Zha10]   Cun-Hui Zhang, *Nearly unbiased variable selection under minimax concave penalty*, The Annals of Statistics **38** (2010), 894–942.

[Zha11a]   Tong Zhang, *Adaptive forward-backward greedy algorithm for learning sparse representations*, IEEE Transactions on Information Theory **57** (2011), 4689–4708.

[Zha11b]   ———, *Multi-stage convex relaxation for feature selection*, Tech. Report arXiv:1106.0565, arXiv, 2011.

[ZL08]   Hui Zou and Runze Li, *One-step sparse estimates in nonconcave penalized likelihood models*, Annals of Statistics **36** (2008), no. 4, 1509–1533.

[Zou06]   Hui Zou, *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association **101** (2006), 1418–1429.

[ZY06]   Peng Zhao and Bin Yu, *On model selection consistency of Lasso*, Journal of Machine Learning Research **7** (2006), 2541–2567.

[ZZ11]   Cun-Hui Zhang and Tong Zhang, *A general theory of concave regularization for high dimensional sparse estimation problems*, Tech. Report arXiv:1108.4988, arXiv, 2011.

DEPARTMENT OF STATISTICS AND BIOSTATISTICS, HILL CENTER, BUSCH CAMPUS, RUTGERS UNIVERSITY, PISCATAWAY, NJ 08854, USA
  E-mail address: czhang@stat.rutgers.edu

DEPARTMENT OF STATISTICS, COLUMBIA UNIVERSITY, NEW YORK, NY 10027
  E-mail address: sszhang@stat.columbia.edu