ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2014)
76, *Part* 3, *pp.* 627–649

High dimensional thresholded regression and shrinkage effect

Zemin Zheng, Yingying Fan and Jinchi Lv

University of Southern California, Los Angeles, USA

[Received May 2012. Revised May 2013]

Summary. High dimensional sparse modelling via regularization provides a powerful tool for analysing large-scale data sets and obtaining meaningful interpretable models. The use of non-convex penalty functions shows advantage in selecting important features in high dimensions, but the global optimality of such methods still demands more understanding. We consider sparse regression with a hard thresholding penalty, which we show to give rise to thresholded regression. This approach is motivated by its close connection with L_0-regularization, which can be unrealistic to implement in practice but of appealing sampling properties, and its computational advantage. Under some mild regularity conditions allowing possibly exponentially growing dimensionality, we establish the oracle inequalities of the resulting regularized estimator, as the global minimizer, under various prediction and variable selection losses, as well as the oracle risk inequalities of the hard thresholded estimator followed by further L_2-regularization. The risk properties exhibit interesting shrinkage effects under both estimation and prediction losses. We identify the optimal choice of the ridge parameter, which is shown to have simultaneous advantages to both the L_2-loss and the prediction loss. These new results and phenomena are evidenced by simulation and real data examples.

Keywords: Global optimality; Hard thresholding; High dimensionality; Prediction and variable selection; Shrinkage effect; Thresholded regression

1. Introduction

The advances of information technologies in the past few decades have made it much easier than before to collect large amounts of data over a wide spectrum of dimensions in different fields. As a powerful tool of sparse modelling and variable selection, regularization methods have been widely used to analyse large-scale data sets and to produce meaningful interpretable models. Depending on the type of penalty functions that are used, the regularization methods can be grouped into two classes: convex and non-convex. A typical example of a convex penalty is the L_1-penalty which gives rise to the L_1-regularization methods such as the lasso (Tibshirani, 1996) and Dantzig selector (Candès and Tao, 2007). The convexity of these methods makes the implementation efficient and facilitates the theoretical analysis. In a seminal paper, Bickel *et al*. (2009) established the oracle inequalities of both the lasso and the Dantzig selector under various prediction and estimation losses and, in particular, proved their asymptotic equivalence under certain conditions.

Despite their convexity and popularity, it has become well known that convex regularization methods can suffer from the bias issue that is inherited from the convexity of the penalty function. This issue can deteriorate the power of identifying important covariates and the efficiency of

Address for correspondence: Yingying Fan, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA.
E-mail: fanyingy@marshall.usc.edu

 1369–7412/14/76627

estimating their effects in high dimensions. To attenuate this issue, Fan and Li (2001) initiated the general framework of non-concave penalized likelihood with non-convex penalties including the proposed smoothly clipped absolute deviation penalty and showed that oracle properties can hold for a wide class of non-convex regularization methods. Other non-convex regularization methods include bridge regression using the L_q-penalty for $0 < q < 1$ (Frank and Friedman, 1993), the minimax concave penalty (Zhang, 2010) and the smooth integration of counting and absolute deviation penalty SICA (Lv and Fan, 2009). The main message of these works is that non-convex regularization can be beneficial in selecting important covariates in high dimensions.

Although there is a growing literature on non-convex regularization methods, some important questions still remain. As an important step, most existing studies for these methods focus on some local minimizer with appealing properties due to their general non-convexity. Yet the properties of the global minimizer need more delicate analysis, and the global theory may depend on the specific form of regularization. A natural question is whether the oracle inequalities hold for non-convex regularization methods as for the L_1-regularization methods (Candès and Tao, 2007; Bickel *et al.*, 2009), and the logarithmic factor of dimensionality that appears commonly in the oracle inequalities is optimal. In the problem of Gaussian mean estimation, there is the well-known phenomenon of Stein's shrinkage effect (Stein, 1956; James and Stein, 1961) stating that the maximum likelihood estimator or least squares estimator may no longer be optimal in risk under quadratic loss in multiple dimensions. Thus another natural question is whether similar shrinkage effects hold for these methods under both estimation and prediction losses.

In this paper, we intend to provide some partial answers to the aforementioned questions, with a focus on one particular member of the non-convex family: the hard thresholding penalty. L_0-regularization, which amounts to best subset regression, motivated different forms of regularization. This method has been proved to enjoy oracle risk inequalities under the prediction loss in Barron *et al.* (1999). It is, however, unrealistic to implement in practice owing to its combinatorial computational complexity. As an alternative to the L_0-penalty, the hard thresholding penalty is continuous with a fixed, finite maximum concavity which controls the computational difficulty. We show that both approaches give rise to a thresholded regression. As is well known in the wavelets literature, hard thresholding regularization is equivalent to L_0-regularization in the case of an orthonormal design matrix. This connection motivates us to investigate the approach of hard thresholding regularization more fully.

The main contributions of this paper are twofold. First, we establish comprehensive global properties of hard thresholding regularization, including the oracle inequalities under various prediction and variable selection losses and the non-optimality of the logarithmic factor of dimensionality. Second, we show that hard thresholding regularization followed by a further L_2-regularization enjoys interesting Stein shrinkage effects in terms of risks under both estimation and prediction losses. The identified optimal choice of the ridge parameter is revealed to have simultaneous advantages to both the L_2-loss and the prediction loss, which result builds an interesting connection between model selection and prediction. These new results and phenomena provide further insights into the hard thresholding regularization method.

The rest of the paper is organized as follows. Section 2 presents thresholded regression with the hard thresholding penalty and L_0-penalty, and their hard thresholding property. We establish the global properties of thresholded regression under various prediction and variable selection losses and unveil Stein's shrinkage effects for both estimation and prediction losses, with optimal choices of the ridge parameter, in Section 3. Section 4 discusses briefly the implementation of thresholded regression. We provide several simulation and real data examples in Section 5. All technical details are relegated to Appendix A and on-line supplementary material.

2. Thresholded regression

To address the questions that were raised in Section 1, we focus our attention on the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is an n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ deterministic design matrix consisting of p covariate vectors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-dimensional regression coefficient vector and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is an n-dimensional error vector. The goal of variable selection is to recover the true underlying sparse model $\mathrm{supp}(\boldsymbol{\beta}_0) = \{j : \beta_{0,j} \neq 0, 1 \leqslant j \leqslant p\}$ consistently for the true regression coefficient vector $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^{\mathrm{T}}$ in model (1), and to estimate the $s = \|\boldsymbol{\beta}_0\|_0$ non-zero regression coefficients $\beta_{0,j}$.

To produce a sparse estimate of $\boldsymbol{\beta}_0$, we consider the approach of penalized least squares which minimizes

$$Q(\boldsymbol{\beta}) = (2n)^{-1}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \|p_\lambda(\boldsymbol{\beta})\|_1, \tag{2}$$

the penalized residual sum of squares with penalty function $p_\lambda(t)$. Here we use the compact notation $p_\lambda(\boldsymbol{\beta}) = p_\lambda(|\boldsymbol{\beta}|) = (p_\lambda(|\beta_1|), \ldots, p_\lambda(|\beta_p|))^{\mathrm{T}}$ with $|\boldsymbol{\beta}| = (|\beta_1|, \ldots, |\beta_p|)^{\mathrm{T}}$. The penalty function $p_\lambda(t)$, which is defined on $t \in [0, \infty)$ and indexed by $\lambda \geqslant 0$, is assumed to be increasing in both t and λ with $p_\lambda(0) = 0$, indicating that the amount of regularization increases with the magnitude of the parameter and the regularization parameter λ. To align all covariates to a common scale, we assume that each covariate vector $\mathbf{x}_j$ is rescaled to have L_2-norm $n^{1/2}$, matching that of the constant covariate $\mathbf{1}$ for the intercept. See, for example, the references mentioned in Section 1 for the specific forms of various penalty functions that have been proposed for sparse modelling.

As elucidated in Section 1, we focus on the hard thresholding penalty

$$p_{\mathrm{H},\lambda}(t) = \tfrac{1}{2}\{\lambda^2 - (\lambda - t)_+^2\}, \qquad t \geqslant 0, \tag{3}$$

which is closely related to the L_0-penalty $p_{\mathrm{H}_0,\lambda}(t) = 2^{-1}\lambda^2 \mathbf{1}_{\{t \neq 0\}}$, $t \geqslant 0$. It is well known that in the wavelets setting with the design matrix $\mathbf{X}$ multiplied by $n^{-1/2}$ being orthonormal, i.e. $n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = I_p$, the penalized least squares in equation (2) reduces to a componentwise minimization problem with $Q(\boldsymbol{\beta}) = 2^{-1}\|\hat{\boldsymbol{\beta}}_{\mathrm{ols}} - \boldsymbol{\beta}\|_2^2 + \|p_\lambda(\boldsymbol{\beta})\|_1$, where $\hat{\boldsymbol{\beta}}_{\mathrm{ols}} = n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ is the ordinary least squares estimator. In this setting, the use of hard thresholding penalty $p_{\mathrm{H},\lambda}(t)$ gives the componentwise hard thresholding, which is of the form $z\mathbf{1}_{\{|z| > \lambda\}}$, on the ordinary least squares estimator (Antoniadis, 1996). In contrast, the use of the L_1-penalty $p_\lambda(t) = \lambda t$ yields the soft thresholding which is of the form $\mathrm{sgn}(z)(|z| - \lambda)_+$. When the L_0-penalty $p_{\mathrm{H}_0,\lambda}(t)$ is used, an identical hard thresholding rule to that by hard thresholding penalty $p_{\mathrm{H},\lambda}(t)$ is obtained. We see that, in the case of an orthonormal design matrix, both approaches of hard thresholding regularization and L_0-regularization are equivalent. This simple connection suggests that they may have a more general connection, which motivates our study.

Moreover, the hard thresholding penalty in equation (3) is continuous and has fixed, finite maximum concavity

$$\kappa(p_{\mathrm{H},\lambda}) = \sup_{0 < t_1 < t_2 < \infty} -\frac{p'_{\mathrm{H},\lambda}(t_2) - p'_{\mathrm{H},\lambda}(t_1)}{t_2 - t_1} = 1, \tag{4}$$

which is related to the computational difficulty of the regularization method and gives rise to its computational advantage. The computationally attractive method of hard thresholding

regularization indeed shares some similarity with L_0-regularization in the general case, as shown in the following lemma on the hard thresholding property.

Lemma 1. For both hard thresholding penalty $p_{\mathrm{H},\lambda}(t)$ and L_0-penalty $p_{\mathrm{H}_0,\lambda}(t)$, minimizing $Q(\boldsymbol{\beta})$ in equation (2) along the jth co-ordinate with $1 \leqslant j \leqslant p$, at any p-vector $\boldsymbol{\beta}_j$ with jth component 0, gives the univariate global minimizer for that co-ordinate of the same form $\hat{\beta}(z) = z\mathbf{1}_{\{|z|>\lambda\}}$, with $z = n^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_j)^{\mathrm{T}}\mathbf{x}_j$.

The simple observation in lemma 1 facilitates our technical analysis and enables us to derive parallel results for both methods. Since the global minimizer is necessarily the global minimizer along each co-ordinate, the characterization of each co-ordinate in lemma 1 shows that the regularized estimators that are given by both methods are natural generalizations of the univariate hard thresholded estimator. In this sense, we refer to both methods as thresholded regression using hard thresholding. There is, however, no guarantee that both estimators are identical when $p > 1$. We show in theorem 1 in Section 3.2 that the two methods can have similar oracle inequalities under various prediction and variable selection losses, which justifies a further connection between them.

3. Global properties and shrinkage effects of thresholded regression

3.1. Technical conditions

It is well known that high collinearity is commonly associated with large-scale data sets. High collinearity can lead to unstable estimation or even loss of model identifiability in regression problems. More specifically, there may be another p-vector $\boldsymbol{\beta}_1$ that is different from $\boldsymbol{\beta}_0$ such that $\mathbf{X}\boldsymbol{\beta}_1$ is (nearly) identical to $\mathbf{X}\boldsymbol{\beta}_0$, when the dimensionality p is large compared with the sample size n. Thus, to ensure model identifiability and to reduce model instability, it is necessary to control the size of sparse models, since it is clear from the geometric point of view that the collinearity between the covariates increases with the dimensionality. This idea was exploited in Donoho and Elad (2003) for the problem of sparse recovery, i.e. the noiseless case of model (1). To ensure the identifiability of $\boldsymbol{\beta}_0$, they introduced the concept of spark, denoted as $\mathrm{spark}(\mathbf{X})$, for a design matrix $\mathbf{X}$, which is defined as the smallest number τ such that there exists a linearly dependent subgroup of τ columns from $\mathbf{X}$. In particular, $\boldsymbol{\beta}_0$ is uniquely defined as long as $s < \mathrm{spark}(\mathbf{X})/2$, which provides a basic condition for model identifiability.

Since we are interested in variable selection in the presence of noise, we extend their concept of spark to the robust case as follows.

Definition 1. The robust spark $M = \mathrm{rspark}_c(\mathbf{X})$ of an $n \times p$ design matrix $\mathbf{X}$ with bound c is defined as the smallest number τ such that there exists a subgroup of τ columns from $n^{-1/2}\mathbf{X}$ such that the corresponding submatrix has a singular value less than the given positive constant c.

An equivalent representation of the robust spark $M = \mathrm{rspark}_c(\mathbf{X})$ in definition 1 is the largest number τ such that the following inequality holds:

$$\min_{\|\boldsymbol{\delta}\|_0 < \tau, \|\boldsymbol{\delta}\|_2 = 1} n^{-1/2}\|\mathbf{X}\boldsymbol{\delta}\|_2 \geqslant c. \tag{5}$$

This inequality provides a natural constraint on the collinearity for sparse models. In view of inequality (5), our robust spark condition of $s < M/2$, which is to be introduced in condition 2, is in a similar spirit to the restricted eigenvalue condition in Bickel *et al.* (2009). The restricted eigenvalue condition assumes inequality (5) with the L_0-norm constraint $\|\boldsymbol{\delta}\|_0 < \tau$ replaced by the L_1-norm constraint of $\|\boldsymbol{\delta}_{J_0^c}\|_1 \leqslant c_0\|\boldsymbol{\delta}_{J_0}\|_1$ for some positive constant c_0, where $J_0 \subset$

$\{1, \ldots, p\}$ with $|J_0| \leqslant s'$, J_0^c is the complement of J_0 and $\boldsymbol{\delta}_A$ denotes a subvector of $\boldsymbol{\delta}$ consisting of components with indices in a given set A. The robust spark condition of $s < M/2$ requires that inequality (5) holds for $\tau = 2s + 1$. Since such an L_0-norm constraint generally defines a smaller subset than the above L_1-norm constraint for $s' = 2s$, the robust spark condition can be weaker than the restricted eigenvalue condition. It is easy to show that the robust spark $\text{rspark}_c(\mathbf{X})$ increases as c decreases and approaches the spark $\text{spark}(\mathbf{X})$ as $c \to 0+$. Thus M can generally be any positive integer no larger than $n + 1$.

To ensure model identifiability and to reduce the instability in the estimated model, we consider the regularized estimator $\hat{\boldsymbol{\beta}}$ on the union of co-ordinate subspaces $\mathbb{S}_{M/2} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 < M/2\}$, as exploited in Fan and Lv (2011) to characterize the global optimality of non-concave penalized likelihood estimators. Thus, throughout the paper, the regularized estimator $\hat{\boldsymbol{\beta}}$ is defined as the global minimizer

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{S}_{M/2}} Q(\boldsymbol{\beta}), \tag{6}$$

where $Q(\boldsymbol{\beta})$ is defined in equation (2). When the size of sparse models exceeds $M/2$, i.e. $\boldsymbol{\beta}$ falls outside the space $\mathbb{S}_{M/2}$, there is generally no guarantee for model identifiability.

To facilitate our technical analysis, we make the following three regularity conditions.

Condition 1. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ for some positive σ.

Condition 2. It holds that $s < M/2$, $s = o(n)$, and $b_0 = \min_{j \in \text{supp}(\boldsymbol{\beta}_0)} |\beta_{0,j}| > \{\sqrt{(16/c^2)} \vee 1\} c^{-1} c_2 \sqrt{\{(2s+1)\log(\tilde{p})/n\}}$, where M is the robust spark of $\mathbf{X}$ with bound c given in definition 1, $c_2 \geqslant \sigma\sqrt{10}$ is some positive constant and $\tilde{p} = n \vee p$.

Condition 3. $\|\boldsymbol{\beta}_0\|_2$ is bounded from below by some positive constant and $\max_{\|\boldsymbol{\delta}\|_0 < M/2, \|\boldsymbol{\delta}\|_2 = 1} n^{-1/2} \|\mathbf{X}\boldsymbol{\delta}\|_2 \leqslant c_3$ for some positive constant c_3.

Condition 1 is standard in the linear regression model. The Gaussian error distribution is assumed to simplify the technical arguments. The theoretical results continue to hold for other error distributions with possibly different probability bound in theorem 1; see, for example, Fan and Lv (2011) for more technical details. In particular, some numerical results for the t error distribution are presented in Section 5.1.2. The heavy-tailedness of the error distribution typically leads to lower probability for the prediction and variable selection bounds to hold.

The first part $s < M/2$ of condition 2 puts a sparsity constraint on the true model size s that involves the robust spark given in definition 1. As explained above, such a robust spark condition is needed to ensure model identifiability. We typically assume a diverging ratio of the sample size n to the number of true covariates s, i.e. $s = o(n)$, to obtain consistent estimation of $\boldsymbol{\beta}_0$. The third part of condition 2 gives a lower bound on the minimum signal strength for model selection consistency.

Condition 3, which is needed only in theorem 2, facilitates the derivation of the oracle risk properties of the regularized estimator, which are stronger than the oracle inequalities in theorem 1. In particular, the first part of condition 3 assumes that the L_2-norm of $\boldsymbol{\beta}_0$ is bounded from below, which is mild and sensible. The second part of condition 3 is a restricted-eigenvalue-type assumption and requires that the maximum singular value of each submatrix of $n^{-1/2}\mathbf{X}$ by taking out less than $M/2$ columns is bounded from above.

3.2. Global properties and shrinkage effects

In view of lemma 1, the regularization parameter λ determines the threshold level for both hard

thresholding penalty $p_{\mathrm{H},\lambda}(t)$ and L_0-penalty $p_{\mathrm{H}_0,\lambda}(t)$. So a natural idea for ensuring the model selection consistency is to choose an appropriately large regularization parameter λ to suppress all noise covariates and to retain important covariates. This approach is shown to be effective in the following theorem on the model selection consistency and oracle inequalities of thresholded regression.

Theorem 1. Assume that conditions 1 and 2 hold and $c^{-1}c_2\sqrt{\{(2s+1)\log(\tilde{p})/n\}} < \lambda < b_0\{1 \wedge \sqrt{(c^2/2)}\}$. Then, for both hard thresholding penalty $p_{\mathrm{H},\lambda}(t)$ and L_0-penalty $p_{\mathrm{H}_0,\lambda}(t)$, the regularized estimator $\hat{\boldsymbol{\beta}}$ in equation (6) satisfies that with probability at least $1-(2/\pi)^{1/2}c_2^{-1}\sigma \times \log(\tilde{p})^{-1/2}\tilde{p}^{1-c_2^2/(2\sigma^2)} - (2/\pi)^{1/2}c_2'^{-1}\sigma s\log(n)^{-1/2}\, n^{-c_2'^2/(2\sigma^2)}$ for some positive constant $c_2' \geqslant \sigma\sqrt{2}$, it holds simultaneously that

(a) (model selection consistency) $\mathrm{supp}(\hat{\boldsymbol{\beta}}) = \mathrm{supp}(\boldsymbol{\beta}_0)$,
(b) (prediction loss) $n^{-1/2}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2 \leqslant 2c_2'c^{-1}\sqrt{\{s\log(n)/n\}}$ and
(c) (estimation losses) $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q \leqslant 2c^{-2}c_2's^{1/q}\sqrt{\{\log(n)/n\}}$ for $q \in [1,2]$ and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty$ is bounded by the same upper bound as for $q=2$.

With the above choice of the regularization parameter λ, the prediction loss of the regularized estimator is within a logarithmic factor $\log(n)^{1/2}$ of that of the oracle estimator, which is referred to as the least squares estimator on the true underlying sparse model. Theorem 1 also establishes the oracle inequalities of the regularized estimator under the L_q-estimation losses with $q \in [1,2] \cup \{\infty\}$. These results hold simultaneously with significant probability that converges to 1 polynomially with sample size n, since $\tilde{p} = n \vee p$. The dimensionality p is allowed to grow up to exponentially fast with the sample size n, in view of the range for λ.

The key to deriving these rates is establishing the model selection consistency of the hard thresholded estimator, i.e. the exact recovery of the true underlying sparse model. Such a property enables us to construct a key event with significant probability, on which we can conduct delicate analysis. A suitable range of the regularization parameter is critical in this theorem since the lower bound on λ is needed for suppressing all noise covariates and the upper bound on λ is needed for retaining all true covariates, although this range is unknown to us in practice.

Theorem 1 builds on lemma 1, both of which share a common feature that the technical arguments apply equally to both the hard thresholding penalty and the L_0-penalty. Thus, under conditions of theorem 1, the regularized estimators given by both hard thresholding penalty $p_{\mathrm{H},\lambda}(t)$ and L_0-penalty $p_{\mathrm{H}_0,\lambda}(t)$ are approximately asymptotically equivalent, i.e. have the same convergence rates in the oracle inequalities under various prediction and variable selection losses. This formally justifies the motivation and advantage of studying hard thresholding regularization. In fact, their approximate asymptotic equivalence extends to the oracle risk inequalities under different prediction and variable selection losses. These results complement those on the oracle risk inequalities under the prediction loss in Barron *et al.* (1999). Since it enjoys the same appealing properties as L_0-regularization, hard thresholding regularization provides an attractive alternative to L_0-regularization thanks to its computational advantage, as discussed in Section 2.

As mentioned in Section 1, many studies have contributed to the oracle inequalities for L_1-regularization methods. For instance, Candès and Tao (2007) proved that the Dantzig selector can achieve a loss within a logarithmic factor of the dimensionality compared with that for the oracle estimator. Bunea *et al.* (2007) established sparsity oracle inequalities for the lasso estimator. Bickel *et al.* (2009) derived parallel oracle inequalities for the lasso estimator and Dantzig selector under the prediction loss and L_q-estimation losses with $q \in [1,2]$. A common feature of these results is the appearance of some power of the logarithmic factor $\log(p)$ of the dimensionality p. In contrast, such a factor is replaced by the logarithmic factor $\log(n)$ of

the sample size n in our setting. This suggests the general non-optimality of the logarithmic factor of dimensionality in the oracle inequalities when p grows non-polynomially with n. Our results are also related to other work on non-convex regularization methods. Antoniadis and Fan (2001) obtained comprehensive oracle inequalities and universal thresholding parameters for a wide class of general penalty functions, in the wavelets setting. Zhang (2010) proved that the minimax concave penalty estimator can attain certain minimax convergence rates for the estimation of regression coefficients in L_q-balls.

Although providing bounds on different estimation and prediction losses on an event with large probability, the oracle inequalities of the thresholded regression presented in theorem 1 do not take into account its performance over the full sample space. Thus it is of interest to investigate a stronger property of oracle risk inequalities for thresholded regression, where the risk under a loss is its expectation over all realizations. As shown in the proof of theorem 1, the hard thresholded estimator $\hat{\boldsymbol{\beta}}$ in equation (6) on its support $\mathrm{supp}(\hat{\boldsymbol{\beta}})$ is exactly the ordinary least squares estimator constructed by using covariates in $\mathrm{supp}(\hat{\boldsymbol{\beta}})$. Motivated by such a representation, we consider a refitted estimator constructed by applying a further L_2-regularization to the thresholded regression

$$\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} = (\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1 + \lambda_1 I_{s_1})^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y}, \tag{7}$$

where $\mathbf{X}_1$ is a submatrix of the design matrix $\mathbf{X}$ consisting of columns in $\mathrm{supp}(\hat{\boldsymbol{\beta}})$, $s_1 = \|\hat{\boldsymbol{\beta}}\|_0$ and $\lambda_1 \geqslant 0$ is the ridge parameter. In the special case of $\lambda_1 = 0$, the above refitted estimator $\hat{\boldsymbol{\beta}}_{\mathrm{refitted}}$ becomes the original hard thresholded estimator $\hat{\boldsymbol{\beta}}$ in equation (6).

Let $\mathbf{X}_0$ be a submatrix of the design matrix $\mathbf{X}$ consisting of columns in $\mathrm{supp}(\boldsymbol{\beta}_0)$ and $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 = P^{\mathrm{T}}DP$ an eigendecomposition with P an orthogonal matrix and $D = \mathrm{diag}(d_1, \ldots, d_s)$. We show that Stein's shrinkage effects (Stein, 1956; James and Stein, 1961) also hold for thresholded regression followed by L_2-regularization in terms of risks under both estimation and prediction losses. These results are presented in the following theorem on the oracle risk inequalities of the L_2-regularized thresholded regression.

Theorem 2. Assume that the conditions of theorem 1 and condition 3 hold. Then the L_2-regularized refitted estimator $\hat{\boldsymbol{\beta}}_{\mathrm{refitted}}$ in equation (7) satisfies that

(a) (L_2-risk) the minimum L_2-risk $E\|\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0\|_2^2$ is attained at the optimal ridge parameter

$$\lambda_1 = \lambda_{1,\mathrm{opt}} = O(s\|\boldsymbol{\beta}_0\|_2^{-2}) + O(s^2 n^{-1}\|\boldsymbol{\beta}_0\|_2^{-4}),$$

with the leading term $O(s\|\boldsymbol{\beta}_0\|_2^{-2})$ sandwiched between $s\sigma^2\|\boldsymbol{\beta}_0\|_2^{-2}(\lambda_{\min}/\lambda_{\max})^2$ and $s\sigma^2\|\boldsymbol{\beta}_0\|_2^{-2}(\lambda_{\max}/\lambda_{\min})^2$, and equals $O(s/n) + O(s^2 n^{-2}\|\boldsymbol{\beta}_0\|_2^{-2})$ with the leading term $O(s/n)$ being $\sum_{j=1}^{s}(\lambda_{1,\mathrm{opt}}^2 b_j^2 + d_j\sigma^2)/(d_j + \lambda_{1,\mathrm{opt}})^2$;

(b) (L_q-risk) the minimum L_q-risk $E\|\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0\|_q^q$ equals

$$O(s/n^{q/2}) + O(s^2\|\boldsymbol{\beta}_0\|_2^{-2}/n^{q/2+1})$$

for $q \in [1, 2]$, and the minimum L_∞-risk $E\|\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0\|_\infty$ equals

$$O(s^{1/2}/n^{1/2}) + O(s^{3/2}\|\boldsymbol{\beta}_0\|_2^{-2}/n^{3/2});$$

(c) (prediction risk) the minimum prediction risk $n^{-1}E\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2$ is attained at the optimal ridge parameter $\lambda_1 = \lambda'_{1,\mathrm{opt}} = O(s\|\boldsymbol{\beta}_0\|_2^{-2}) + O(s^2 n^{-1}\|\boldsymbol{\beta}_0\|_2^{-4})$, with the leading term $O(s\|\boldsymbol{\beta}_0\|_2^{-2})$ sandwiched between $s\sigma^2\|\boldsymbol{\beta}_0\|_2^{-2}\lambda_{\min}/\lambda_{\max}$ and $s\sigma^2\|\boldsymbol{\beta}_0\|_2^{-2} \times \lambda_{\max}/\lambda_{\min}$, and equals $O(s/n) + O(s^2 n^{-2}\|\boldsymbol{\beta}_0\|_2^{-2})$ with the leading term $O(s/n)$ being $n^{-1}\sum_{j=1}^{s}\{(\lambda'_{1,\mathrm{opt}})^2 b_j^2 d_j + d_j^2\sigma^2\}/(d_j + \lambda'_{1,\mathrm{opt}})^2$,

where $(b_1, \ldots, b_s)^{\mathrm{T}} = P\boldsymbol{\beta}_{0,1}$ with $\boldsymbol{\beta}_{0,1}$ a subvector of $\boldsymbol{\beta}_0$ consisting of all non-zero components, and $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$ respectively.

Although it has the well-known bias issue, ridge regression applied after thresholded regression is shown in theorem 2 to be capable of improving both estimation and prediction, since the original hard thresholded estimator is simply the refitted estimator with $\lambda_1 = 0$ and the minimum risks under the losses are attained at non-zero ridge parameters λ_1. Intuitively, the bias that is incurred by an appropriately small amount of L_2-regularization can be offset by the reduction in estimation variability, leading to improvement in the overall risks of the regularized estimator. This can be clearly seen in the representative L_2-risk and prediction risk curves as a function of the ridge parameter λ_1 in Section 5. The risks drop as λ_1 increases from 0 and start to rise after the minimum risks have been attained.

The model selection consistency of thresholded regression plays a key role in deriving the risk properties of the L_2-regularized refitted estimator. The optimal risks are attained at non-trivial ridge parameter λ_1 for both the L_q-loss and the prediction loss. Since $s = o(n)$ by condition 2 and $\|\boldsymbol{\beta}_0\|_2$ is bounded from below by some positive constant by condition 3, we see that both optimal ridge parameters $\lambda_{1,\mathrm{opt}}$ and $\lambda'_{1,\mathrm{opt}}$ for the L_2-risk and prediction risk respectively are of the same leading order $O(s\|\boldsymbol{\beta}_0\|_2^{-2})$. In particular, the leading term $O(s\|\boldsymbol{\beta}_0\|_2^{-2})$ of the optimal ridge parameter for L_2-risk has a similar range to that of the optimal ridge parameter for prediction risk, differing by a factor of only $\lambda_{\max}/\lambda_{\min}$. Such a factor is the condition number of the Gram matrix $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$ resulting from the true design matrix $\mathbf{X}_0$. In view of inequality (5) and condition 3, its condition number $\lambda_{\max}/\lambda_{\min}$ is sandwiched between 1 and c_3^2/c^2.

It is interesting to observe that the optimal choices of the ridge parameter for both L_2-loss and prediction loss are of the same order $O(s\|\boldsymbol{\beta}_0\|_2^{-2})$, which is proportional to the true model size s and has an inverse relationship with $\|\boldsymbol{\beta}_0\|_2$. This indicates that a stronger signal leads to smaller optimal L_2-shrinkage. Thus the optimal ridge parameter has a simultaneous benefit on both the L_2-loss and the prediction loss. Furthermore, the minimum L_2-risk and minimum prediction risk share the same order of $O(s/n)$, when the risks are minimized by the optimal ridge parameters. These risk properties demonstrate that Stein's shrinkage effects extend to the thresholded regression followed by the L_2-regularization under both estimation and prediction losses.

The idea of refitting has also been investigated in van de Geer *et al.* (2011), who established bounds on the prediction loss and L_q-loss with $q \in [1, 2]$ for the thresholded lasso estimator. The thresholded lasso is a three-step procedure with the lasso followed by hard thresholding and an ordinary least squares refitting, whereas the above L_2-regularized refitting is a two-step procedure with hard thresholding and ordinary least squares refitting automatic in thresholded regression. The main difference is that our study focuses on the risk properties and identifying optimal ridge parameters for minimizing the risks. These new risk properties reveal interesting Stein shrinkage effects in thresholded regression, which were lacking before.

4. Implementation

Efficient algorithms for the implementation of regularization methods include the LQA (Fan and Li, 2001), LARS (Efron *et al.*, 2004) and LLA (Zou and Li, 2008) algorithms. As an alternative to these algorithms, co-ordinate optimization has become popular owing to its scalability for large-scale problems; see, for example, Friedman *et al.* (2007), Wu and Lange (2008) and Fan and Lv (2011). In this paper, we apply the ICA algorithm (Fan and Lv, 2011) to implement the regularization methods. See section V in Fan and Lv (2011) for a detailed description of

this algorithm. An analysis of convergence properties of this algorithm has been presented in Lin and Lv (2013). In particular, the univariate global minimizer for each co-ordinate admits a closed form as given in lemma 1, for both hard thresholding penalty $p_{\mathrm{H},\lambda}(t)$ and L_0-penalty $p_{\mathrm{H}_0,\lambda}(t)$. We point out that the algorithm is not guaranteed to find the global minimizer.

Although our theory relies on the union of co-ordinate subspaces $\mathbb{S}_{M/2}$ that are associated with the robust spark of the design matrix, the implementation via the ICA algorithm does not require knowledge of such a space. It is a path following algorithm, based on a decreasing grid of regularization parameter λ, that produces a sequence of most sparse solutions to less sparse solutions, with the solution given by the previous λ as an initial value for the next λ. The collinearity of sparse models can be tracked easily by calculating the smallest singular value of the subdesign matrix given by the support of each sparse solution produced.

To illustrate our theoretical results better and to make a fair comparison of all methods, we select the tuning parameters by minimizing the prediction error calculated by using an independent validation set, with size equal to the sample size in the simulation study. We use SICA (Lv and Fan, 2009) with penalty $p_\lambda(t;a) = \lambda(a+1)t/(a+t)$, with a small shape parameter a such as 10^{-4} or 10^{-2}, as a proxy of the L_0-regularization method. Following Lin and Lv (2013), some pilot solutions with larger values of a are computed to stabilize the solution. See also Lin and Lv (2013) for the closed form solution of the univariate SICA estimator.

5. Numerical studies

In this section, we investigate the finite sample performance of regularization methods with hard thresholding and SICA penalties, with comparison with the lasso and oracle procedure which knew the true model in advance. We consider both cases of light-tailed and heavy-tailed errors, with a Gaussian distribution for the former and a t-distribution for the latter.

5.1. Simulation examples

5.1.1. Simulation example 1

We first consider the linear regression model (1) with Gaussian error $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$. We generated 100 data sets from this model with true regression coefficient vector $\boldsymbol{\beta}_0 = (\mathbf{v}^{\mathrm{T}}, \ldots, \mathbf{v}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ with the pattern $\mathbf{v} = (\boldsymbol{\beta}_{\mathrm{strong}}^{\mathrm{T}}, \boldsymbol{\beta}_{\mathrm{weak}}^{\mathrm{T}})^{\mathrm{T}}$ repeated q times, where $\boldsymbol{\beta}_{\mathrm{strong}} = (0.6, 0, 0, -0.6, 0, 0)^{\mathrm{T}}$ and $\boldsymbol{\beta}_{\mathrm{weak}} = (0.05, 0, 0, -0.05, 0, 0)^{\mathrm{T}}$ or $(0.1, 0, 0, -0.1, 0, 0)^{\mathrm{T}}$. The coefficient subvectors $\boldsymbol{\beta}_{\mathrm{strong}}$ and $\boldsymbol{\beta}_{\mathrm{weak}}$ denote the strong signals and weak signals in $\boldsymbol{\beta}_0$ respectively. The two choices of $\boldsymbol{\beta}_{\mathrm{weak}}$ showed the performance of four methods under different levels of weak signals. We set $q = 3$ so that there are six strong signals (with magnitude 0.6) and six weak signals (with magnitude 0.05 or 0.1) in the true coefficient vector. The sample size n was chosen to be 100 and two settings of $(p, \sigma) = (1000, 0.4)$ and $(p, \sigma) = (5000, 0.3)$ were considered. For each data set, all the rows of the $n \times p$ design matrix $\mathbf{X}$ were sampled as independent and identically distributed copies from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{1 \leqslant i,j \leqslant p}$. This allows for correlation between the covariates at the population level. The sample collinearity between the covariates can be at an even higher level owing to high dimensionality. We applied the lasso, hard thresholding and SICA to produce a sequence of sparse models and selected the tuning parameters as discussed in Section 4.

The overall signal-to-noise ratios in the settings of $(p, |\beta_{\mathrm{weak}}|) = (1000, 0.05)$, (1000, 0.1), (5000, 0.05), (5000, 0.1) are 11.70, 11.77, 20.80 and 20.92 respectively. These overall measures, however, do not reflect the individual signal strength for each strong or weak signal, which measures the difficulty of the variable selection problem. In the case of $p = 1000$, the indi-

vidual signal-to-noise ratio is $0.6^2/\sigma^2 = 2.25$ for each strong signal, and $0.05^2/\sigma^2 = 0.0156$ or $0.1^2/\sigma^2 = 0.0625$ for each weak signal with level 0.05 or 0.1. In the case $p = 5000$, the individual signal-to-noise ratio is $0.6^2/\sigma^2 = 4$ for each strong signal, and $0.05^2/\sigma^2 = 0.0278$ or $0.1^2/\sigma^2 = 0.1111$ for each weak signal with level 0.05 or 0.1. We see that the six weak covariates have very low signal strength. Their signal strength is even lower when the high dimensionality is taken into account, owing to the well-known phenomenon of noise accumulation in high dimensions.

To compare the three regularization methods with the oracle procedure, we consider several performance measures. The first measure is the prediction error PE defined as $E(Y - \mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2$ with $\hat{\boldsymbol{\beta}}$ an estimate and $(\mathbf{x}^{\mathrm{T}}, Y)$ an independent observation of the covariates and response. To calculate the expectation, we generated an independent test sample of size 10000. The second, third and fourth measures are the L_q-estimation losses $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q$ with $q = 2, 1, \infty$ respectively. The fifth, sixth and seventh measures are the number of false positive (FP) selections and numbers of false negative (FN) selections for strong signals and FN selections for weak signals for variable selection, where an FP selection means a falsely selected noise covariate in the model and an FN

Table 1. Means and standard deviations (in parentheses) of various performance measures by all methods over 100 simulations in Section 5.1.1

Setting	*Measure*	*Results for the following methods:*			
		Lasso	*Hard thresholding*	*SICA*	*Oracle*
$p = 1000$,	PE	0.3025 (0.0479)	0.1862 (0.0086)	0.1862 (0.0103)	0.1829 (0.0100)
$\lvert\beta_{\mathrm{weak}}\rvert = 0.05$	L_2-loss	0.4007 (0.0653)	0.1679 (0.0238)	0.1678 (0.0276)	0.1505 (0.0324)
	L_1-loss	1.7660 (0.2942)	0.5274 (0.0769)	0.5276 (0.0921)	0.4277 (0.0979)
	L_∞-loss	0.2012 (0.0418)	0.0804 (0.0258)	0.0790 (0.0255)	0.0854 (0.0207)
	FP	33.7900 (7.0457)	0.0800 (0.2727)	0.0900 (0.4044)	0 (0)
	FN (strong)	0 (0)	0 (0)	0 (0)	0 (0)
	FN (weak)	5.6000 (0.6513)	5.9900 (0.1000)	5.9900 (0.1000)	0 (0)
	$\hat{\sigma}$	0.4295 (0.0473)	0.4158 (0.0328)	0.4155 (0.0351)	0.4000 (0.0347)
$p = 1000$,	PE	0.3643 (0.0584)	0.2272 (0.0115)	0.2283 (0.0124)	0.1829 (0.0100)
$\lvert\beta_{\mathrm{weak}}\rvert = 0.1$	L_2-loss	0.4882 (0.0674)	0.2749 (0.0223)	0.2769 (0.0224)	0.1505 (0.0324)
	L_1-loss	2.2134 (0.3202)	0.8466 (0.1018)	0.8553 (0.1052)	0.4277 (0.0979)
	L_∞-loss	0.2225 (0.0453)	0.1068 (0.0177)	0.1077 (0.0177)	0.0854 (0.0207)
	FP	34.4300 (6.9866)	0.0900 (0.3208)	0.1600 (0.5453)	0 (0)
	FN (strong)	0 (0)	0 (0)	0 (0)	0 (0)
	FN (weak)	4.9200 (0.9711)	5.8200 (0.6257)	5.8000 (0.5125)	0 (0)
	$\hat{\sigma}$	0.4676 (0.0541)	0.4559 (0.0377)	0.4540 (0.0425)	0.4000 (0.0347)
$p = 5000$,	PE	0.2634 (0.0744)	0.1097 (0.0058)	0.1088 (0.0039)	0.1027 (0.0062)
$\lvert\beta_{\mathrm{weak}}\rvert = 0.05$	L_2-loss	0.4419 (0.0904)	0.1476 (0.0185)	0.1450 (0.0127)	0.1122 (0.0260)
	L_1-loss	1.8507 (0.3387)	0.4593 (0.0602)	0.4528 (0.0464)	0.3166 (0.0775)
	L_∞-loss	0.2188 (0.0507)	0.0621 (0.0206)	0.0592 (0.0152)	0.0663 (0.0188)
	FP	37.3900 (4.9826)	0.0600 (0.2778)	0.0100 (0.1000)	0 (0)
	FN (strong)	0 (0)	0 (0)	0 (0)	0 (0)
	FN (weak)	5.8600 (0.3487)	5.9900 (0.1000)	6.0000 (0)	0 (0)
	$\hat{\sigma}$	0.3822 (0.0452)	0.3173 (0.0239)	0.3187 (0.0231)	0.2976 (0.0242)
$p = 5000$,	PE	0.3603 (0.1089)	0.1838 (0.2401)	0.1489 (0.0070)	0.1027 (0.0062)
$\lvert\beta_{\mathrm{weak}}\rvert = 0.1$	L_2-loss	0.5594 (0.1054)	0.2830 (0.1654)	0.2581 (0.0136)	0.1122 (0.0260)
	L_1-loss	2.4396 (0.3980)	0.8361 (0.4546)	0.7685 (0.1077)	0.3166 (0.0775)
	L_∞-loss	0.2584 (0.0618)	0.1117 (0.0704)	0.1016 (0.0066)	0.0663 (0.0188)
	FP	38.6000 (4.2593)	0.0700 (0.4324)	0.2200 (1.8123)	0 (0)
	FN (strong)	0 (0)	0.1100 (0.7771)	0 (0)	0 (0)
	FN (weak)	5.5300 (0.6269)	5.7700 (0.5096)	5.7100 (0.6403)	0 (0)
	$\hat{\sigma}$	0.4417 (0.0557)	0.3826 (0.1214)	0.3629 (0.0429)	0.2976 (0.0242)

Table 2. Means and standard deviations (in parentheses) of various performance measures by all methods followed by L_2-regularization over 100 simulations in Section 5.1.1

Setting	*Measure*	*Results for the following methods:*			
		Lasso–L_2	*Hard thresholding–L_2*	*SICA–L_2*	*Oracle–L_2*
$p=1000$, $\lvert\beta_{\text{weak}}\rvert=0.05$	PE	0.3501 (0.0538)	0.1851 (0.0083)	0.1852 (0.0101)	0.1812 (0.0098)
	L_2-loss	0.4464 (0.0635)	0.1658 (0.0237)	0.1657 (0.0277)	0.1437 (0.0329)
	L_1-loss	2.5473 (0.3986)	0.5169 (0.0764)	0.5177 (0.0928)	0.4061 (0.1001)
	L_∞-loss	0.1698 (0.0412)	0.0752 (0.0250)	0.0741 (0.0248)	0.0772 (0.0206)
$p=1000$, $\lvert\beta_{\text{weak}}\rvert=0.1$	PE	0.4168 (0.0675)	0.2257 (0.0109)	0.2270 (0.0117)	0.1812 (0.0098)
	L_2-loss	0.5272 (0.0688)	0.2734 (0.0221)	0.2755 (0.0218)	0.1435 (0.0328)
	L_1-loss	3.0270 (0.4430)	0.8366 (0.1016)	0.8450 (0.1045)	0.4053 (0.0990)
	L_∞-loss	0.1905 (0.0472)	0.1053 (0.0150)	0.1060 (0.0142)	0.0770 (0.0204)
$p=5000$, $\lvert\beta_{\text{weak}}\rvert=0.05$	PE	0.2642 (0.0532)	0.1090 (0.0055)	0.1082 (0.0037)	0.1020 (0.0060)
	L_2-loss	0.4358 (0.0675)	0.1462 (0.0181)	0.1437 (0.0127)	0.1082 (0.0263)
	L_1-loss	2.4131 (0.3343)	0.4508 (0.0586)	0.4448 (0.0462)	0.3019 (0.0758)
	L_∞-loss	0.1896 (0.0454)	0.0597 (0.0195)	0.0569 (0.0146)	0.0610 (0.0197)
$p=5000$, $\lvert\beta_{\text{weak}}\rvert=0.1$	PE	0.3594 (0.0816)	0.1830 (0.2403)	0.1481 (0.0071)	0.1020 (0.0060)
	L_2-loss	0.5492 (0.0833)	0.2823 (0.1656)	0.2574 (0.0141)	0.1082 (0.0264)
	L_1-loss	3.0841 (0.4236)	0.8280 (0.4560)	0.7614 (0.1090)	0.3017 (0.0760)
	L_∞-loss	0.2242 (0.0562)	0.1116 (0.0704)	0.1013 (0.0058)	0.0610 (0.0198)

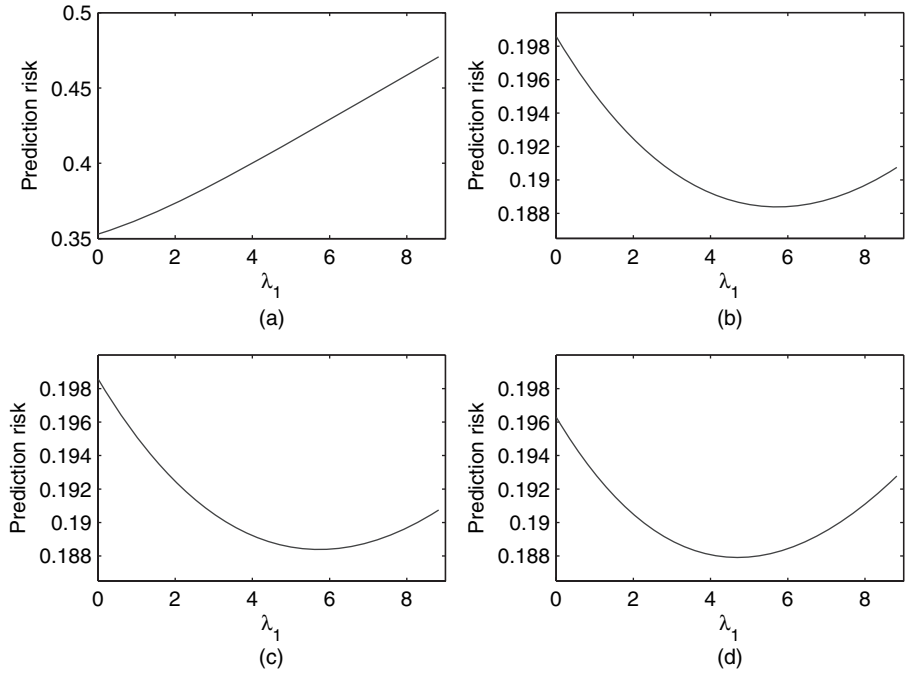

Fig. 1. Representative prediction risk curves as a function of the ridge parameter λ_1 by all methods in Section 5.1.1 for the case of $(p, |\beta_{\text{weak}}|) = (1000, 0.05)$: (a) lasso–$L_2$-regularization; (b) hard thresholding–L_2-regularization; (c) SICA–L_2-regularization; (d) oracle–L_2-regularization

selection means a missed true covariate. We also compare the estimated error standard deviation $\hat{\sigma}$ by all methods.

Table 1 summarizes the comparison results by all methods. As seen in the measure of FN selections for weak signals, the weak covariates tended to be excluded by each regularization method since they have very low signal strength. At the weak signal level of 0.05, thanks to their concavity both hard thresholding and SICA followed very closely the oracle procedure in terms of all other measures, whereas the lasso produced a much larger model with lower prediction and variable selection accuracy owing to its well-known bias issue. When the weak signal level increases to 0.1, the performance of each method deteriorated because of the difficulty of recovering weak covariates. We also considered the case of no weak signals with $\boldsymbol{\beta}_{\text{weak}} = \mathbf{0}$. In such a case, all methods performed better and their relative performance was the same as in the case with the weak signal level of 0.05, with both hard thresholding and SICA having almost identical performance to that of the oracle procedure. For brevity these additional simulation results are not included here but are available on request.

We also investigate the risk properties and shrinkage effects of the L_2-regularized refitted estimators $\hat{\boldsymbol{\beta}}_{\text{refitted}}$ defined in equation (7) for all methods. Table 2 presents the performance of these shrinkage estimators in the above two settings with the ridge parameter λ_1 selected to minimize the corresponding risks. A comparison of risks under different losses in Tables 1 and 2 shows an improvement of the L_2-regularized refitted estimators over the estimators given by hard thresholding, SICA and the oracle procedure. These numerical results are in line with the theoretical results in theorem 2. The results of the L_2-regularized refitted estimator for the lasso show no improvement in risks. This is because the bias issue of the lasso gives rise to a large

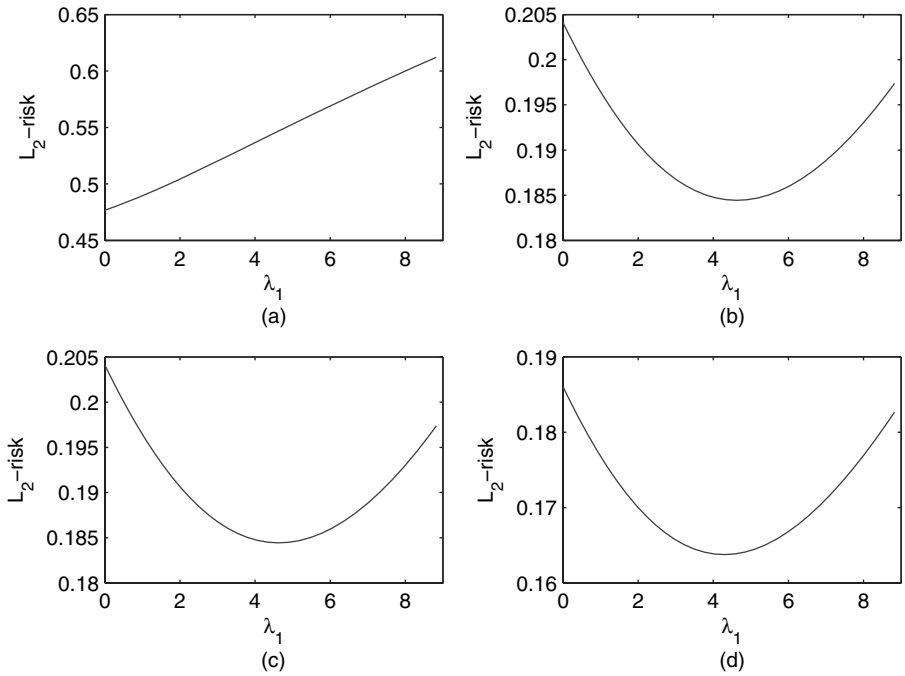

Fig. 2. Representative L_2-risk curves as a function of the ridge parameter λ_1 by all methods in Section 5.1.1 for the case of $(p, |\beta_{\text{weak}}|) = (1000, 0.05)$: (a) lasso–$L_2$-regularization; (b) hard thresholding–L_2-regularization; (c) SICA–L_2-regularization; (d) oracle–L_2-regularization

model. Figs 1 and 2 depict some representative risk curves as a function of the ridge parameter λ_1 by all methods for the prediction loss and L_2-loss respectively. These plots demonstrate Stein's shrinkage effects for the thresholded regression followed by L_2-regularization under both estimation and prediction risks.

5.1.2. Simulation example 2

A natural question is whether the results and phenomena for light-tailed errors hold for heavy-tailed errors or not. We now turn our attention to such a case for the linear regression model (1) with t error distribution. The setting of this simulation example is the same as that in Section 5.1.1 except that the error vector is $\varepsilon = \sigma\eta$, where the components of the n-dimensional random vector η are independent and follow the t-distribution with $\mathrm{df} = 10$ degrees of freedom. We compared the lasso, hard thresholding and SICA with the oracle procedure in the same two

Table 3. Means and standard deviations (in parentheses) of various performance measures by all methods over 100 simulations in Section 5.1.2†

Setting	*Measure*	*Results for the following methods:*			
		Lasso	*Hard thresholding*	*SICA*	*Oracle*
$p = 1000$,	PE	0.3845 (0.0705)	0.2277 (0.0137)	0.2285 (0.0168)	0.2276 (0.0151)
$\lvert\beta_{\mathrm{weak}}\rvert = 0.05$	L_2-loss	0.4547 (0.0801)	0.1718 (0.0316)	0.1734 (0.0361)	0.1655 (0.0415)
	L_1-loss	1.9683 (0.3523)	0.5335 (0.0932)	0.5386 (0.1124)	0.4682 (0.1202)
	L_∞-loss	0.2306 (0.0554)	0.0858 (0.0347)	0.0870 (0.0360)	0.0937 (0.0283)
	FP	32.7800 (8.6311)	0.0600 (0.2778)	0.1000 (0.4606)	0 (0)
	FN (strong)	0 (0)	0 (0)	0 (0)	0 (0)
	FN (weak)	5.6200 (0.6321)	6.0000 (0)	6.0000 (0)	0 (0)
	Error SD	0.4867 (0.0599)	0.4652 (0.0412)	0.4645 (0.0417)	0.4517 (0.0368)
$p = 1000$,	PE	0.4462 (0.0777)	0.2693 (0.0149)	0.2702 (0.0172)	0.2276 (0.0151)
$\lvert\beta_{\mathrm{weak}}\rvert = 0.1$	L_2-loss	0.5331 (0.0773)	0.2787 (0.0245)	0.2797 (0.0272)	0.1655 (0.0415)
	L_1-loss	2.4177 (0.3695)	0.8557 (0.1003)	0.8628 (0.1188)	0.4682 (0.1202)
	L_∞-loss	0.2491 (0.0558)	0.1123 (0.0250)	0.1131 (0.0259)	0.0937 (0.0283)
	FP	34.1400 (8.1996)	0.0600 (0.2387)	0.1300 (0.6139)	0 (0)
	FN (strong)	0 (0)	0 (0)	0 (0)	0 (0)
	FN (weak)	5.0200 (0.9209)	5.9200 (0.2727)	5.8600 (0.4499)	0 (0)
	Error SD	0.5152 (0.0633)	0.5033 (0.0445)	0.5004 (0.0512)	0.4517 (0.0368)
$p = 5000$,	PE	0.3295 (0.1096)	0.1343 (0.0058)	0.1343 (0.0060)	0.1277 (0.0068)
$\lvert\beta_{\mathrm{weak}}\rvert = 0.05$	L_2-loss	0.4897 (0.1151)	0.1539 (0.0168)	0.1541 (0.0169)	0.1226 (0.0270)
	L_1-loss	2.0497 (0.4196)	0.4831 (0.0588)	0.4833 (0.0586)	0.3411 (0.0796)
	L_∞-loss	0.2439 (0.0625)	0.0684 (0.0209)	0.0688 (0.0212)	0.0717 (0.0191)
	FP	38.4600 (5.4558)	0.0300 (0.1714)	0.0300 (0.1714)	0 (0)
	FN (strong)	0 (0)	0 (0)	0 (0)	0 (0)
	FN (weak)	5.8800 (0.3266)	5.9800 (0.1407)	5.9800 (0.1407)	0 (0)
	Error SD	0.4254 (0.0569)	0.3560 (0.0319)	0.3560 (0.0319)	0.3366 (0.0321)
$p = 5000$,	PE	0.4307 (0.1419)	0.1761 (0.0080)	0.1767 (0.0132)	0.1277 (0.0068)
$\lvert\beta_{\mathrm{weak}}\rvert = 0.1$	L_2-loss	0.6030 (0.1245)	0.2671 (0.0146)	0.2680 (0.0219)	0.1226 (0.0270)
	L_1-loss	2.6203 (0.4894)	0.8068 (0.0701)	0.8150 (0.1238)	0.3411 (0.0796)
	L_∞-loss	0.2845 (0.0722)	0.1047 (0.0143)	0.1035 (0.0105)	0.0717 (0.0191)
	FP	38.3900 (5.0510)	0.0500 (0.2190)	0.1800 (1.2092)	0 (0)
	FN (strong)	0 (0)	0 (0)	0 (0)	0 (0)
	FN (weak)	5.5900 (0.5702)	5.8500 (0.3860)	5.8100 (0.4648)	0 (0)
	Error SD	0.4828 (0.0625)	0.4039 (0.0354)	0.4005 (0.0458)	0.3366 (0.0321)

†The population error standard deviation $\mathrm{SD} = \sigma\sqrt{\{\mathrm{df}/(\mathrm{df}-2)\}}$ is 0.4472 in the case of $p = 1000$, and 0.3354 in the case of $p = 5000$.

Table 4. Means and standard deviations (in parentheses) of various performance measures by all methods followed by L_2-regularization over 100 simulations in Section 5.1.2

Setting	*Measure*	*Results for the following methods:*			
		Lasso–L_2	*Hard thresholding–L_2*	*SICA–L_2*	*Oracle–L_2*
$p=1000$, $\lvert\beta_{\text{weak}}\rvert=0.05$	PE	0.4447 (0.0747)	0.2263 (0.0135)	0.2270 (0.0146)	0.2256 (0.0148)
	L_2-loss	0.5059 (0.0767)	0.1686 (0.0321)	0.1701 (0.0342)	0.1588 (0.0411)
	L_1-loss	2.8356 (0.5052)	0.5191 (0.0929)	0.5238 (0.1013)	0.4426 (0.1157)
	L_∞-loss	0.1976 (0.0541)	0.0796 (0.0328)	0.0808 (0.0335)	0.0858 (0.0280)
$p=1000$, $\lvert\beta_{\text{weak}}\rvert=0.1$	PE	0.5170 (0.0835)	0.2676 (0.0149)	0.2684 (0.0172)	0.2256 (0.0148)
	L_2-loss	0.5828 (0.0767)	0.2770 (0.0250)	0.2780 (0.0277)	0.1588 (0.0412)
	L_1-loss	3.3344 (0.5248)	0.8426 (0.1047)	0.8491 (0.1230)	0.4427 (0.1156)
	L_∞-loss	0.2180 (0.0567)	0.1099 (0.0218)	0.1107 (0.0227)	0.0858 (0.0281)
$p=5000$, $\lvert\beta_{\text{weak}}\rvert=0.05$	PE	0.3312 (0.0815)	0.1335 (0.0055)	0.1335 (0.0056)	0.1268 (0.0066)
	L_2-loss	0.4858 (0.0877)	0.1520 (0.0165)	0.1522 (0.0165)	0.1180 (0.0275)
	L_1-loss	2.6985 (0.4105)	0.4719 (0.0587)	0.4724 (0.0581)	0.3256 (0.0807)
	L_∞-loss	0.2090 (0.0523)	0.0645 (0.0182)	0.0649 (0.0187)	0.0653 (0.0182)
$p=5000$, $\lvert\beta_{\text{weak}}\rvert=0.1$	PE	0.4332 (0.1092)	0.1750 (0.0075)	0.1757 (0.0132)	0.1268 (0.0066)
	L_2-loss	0.5954 (0.0995)	0.2659 (0.0144)	0.2668 (0.0221)	0.1178 (0.0275)
	L_1-loss	3.3325 (0.5086)	0.7961 (0.0715)	0.8053 (0.1260)	0.3254 (0.0808)
	L_∞-loss	0.2467 (0.0623)	0.1036 (0.0113)	0.1029 (0.0100)	0.0649 (0.0180)

settings of $(p, \sigma) = (1000, 0.4)$ and $(p, \sigma) = (5000, 0.3)$. The same performance measures as in Section 5.1.1 are employed for comparison.

The means and standard deviations of the various performance measures by all methods are listed in Table 3. Table 4 details the performance of the L_2-regularized refitted estimators, as described in Section 5.1.1, with the ridge parameter λ_1 selected to minimize the corresponding risks. The conclusions are similar to those in Section 5.1.1. By comparing the results in this simulation example with those in Section 5.1.1 for Gaussian error, we see that the performance of all methods deteriorated when the error distribution becomes heavy tailed. Both hard thresholding and SICA still followed closely the oracle procedure at the weak signal level of 0.05. We also observe the phenomenon of Stein's shrinkage effects for the thresholded regression followed by L_2-regularization under both estimation and prediction risks in this case of a heavy-tailed error distribution.

5.2. Real data example

We apply the lasso, hard thresholding and SICA, as well as these methods followed by the L_2-regularization, to the diabetes data set studied in Efron *et al.* (2004). This data set consists of measurements for $n = 442$ diabetes patients on the response variable, a quantitative measure of disease progression 1 year after baseline, and 10 baseline variables: sex, age, body mass index bmi, average blood pressure bp, and six blood serum measurements tc, ldl, hdl, tch, ltg and glu. Efron *et al.* (2004) considered the quadratic model with interactions, by adding the squares of all baseline variables except the dummy variable sex, and all interactions between each pair of the 10 baseline variables. This results in a linear regression model with $p = 64$ predictors. We adopt this model to analyse the diabetes data set.

We randomly split the full data set 100 times into a training set of 400 samples and a validation set of 42 samples. For each splitting of the data set, we applied each regularization method

to the training set with the quadratic model and calculated the prediction error, as defined in Section 5.1.1, on the validation set. Minimizing the prediction error gives the best model for each regularization method. The means (with standard deviations in parentheses) of these minimum prediction errors over 100 random splittings were 2894.5 (655.5) for the lasso, 2802.5 (635.5) for hard thresholding and 2800.6 (615.2) for SICA. We see that both hard thresholding and SICA improved over the lasso in prediction accuracy. The relatively large standard deviations indicate the difficulty of the prediction problem for this data set. On the basis of the estimated model by each method, we also investigated the L_2-regularized refitted estimator with ridge parameter λ_1 selected by the validation set. The means (with standard deviations in parentheses) of their prediction errors over 100 random splittings were 2957.9 (671.3) for the lasso–L_2 regularization, 2770.3 (630.9) for hard thresholding–L_2 regularization and 2770.2 (614.9) for SICA–L_2 regularization. We observe in Fig. 3 shrinkage effects for both hard thresholding and SICA followed by L_2-regularization, whereas the refitting with L_2-regularization did not generally improve the performance of the lasso, as also shown in the simulation studies.

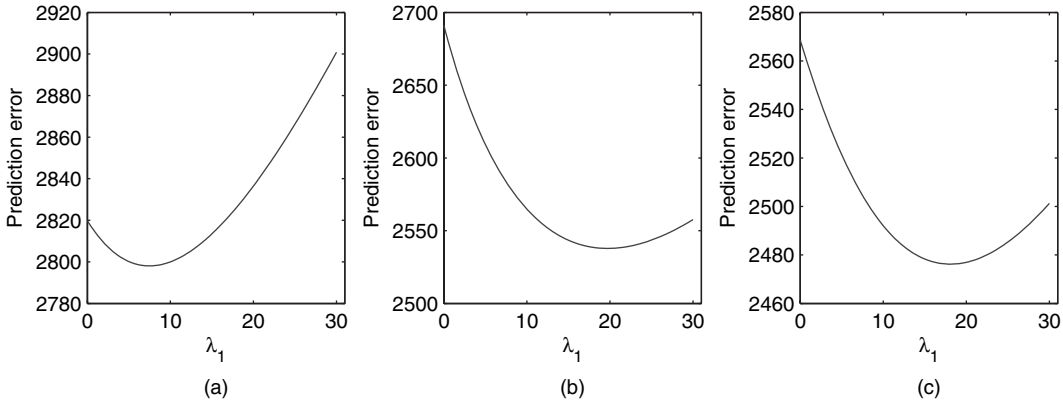

Fig. 3. Representative prediction error curves as a function of the ridge parameter λ_1 by all methods on the diabetes data set in Section 5.2: (a) lasso–L_2-regularization; (b) hard thresholding–L_2-regularization; (c) SICA–L_2-regularization

Table 5. Selection probabilities (*t*-statistics, with magnitude above 2 in italics) of the most frequently selected predictors with number up to median model size by each method across 100 random splittings of the diabetes data set in Section 5.2

Predictor	*Results for the following methods:*			*Predictor*	*Results for the following methods:*		
	Lasso	*Hard thresholding*	*SICA*		*Lasso*	*Hard thresholding*	*SICA*
sex	0.94 (*−2.03*)	0.83 (*−2.16*)	0.82 (*−2.07*)	bp^2	0.54 (0.42)	—	—
bmi	1.00 (*17.24*)	0.99 (*6.25*)	1.00 (*8.65*)	glu^2	1.00 (*3.95*)	0.50 (0.94)	0.57 (1.10)
bp	1.00 (*5.82*)	0.87 (*2.51*)	0.91 (*2.99*)	sex∗age	0.98 (*3.09*)	0.87 (*2.46*)	0.81 (*2.00*)
tc	0.43 (−0.67)	—	—	sex∗bp	0.73 (0.99)	—	—
hdl	1.00 (*−3.63*)	0.80 (−1.86)	0.79 (−1.83)	age∗bp	0.87 (1.27)	—	—
ltg	1.00 (*9.27*)	1.00 (*7.27*)	1.00 (*8.22*)	age∗ltg	0.74 (0.82)	—	—
glu	0.85 (1.21)	—	—	age∗glu	0.59 (0.82)	—	—
age^2	0.94 (1.91)	—	—	bmi∗bp	0.99 (*2.25*)	0.81 (1.94)	0.76 (1.69)
bmi^2	0.98 (*2.49*)	—	—	bp∗hdl	0.47 (0.68)	—	—

We also calculated the median model size by each method: 18 by the lasso, eight by hard thresholding and eight by SICA. For each method, we computed the percentage of times that each predictor was selected and listed the most frequently chosen m predictors in Table 5, with m equal to the median model size by the method. Table 5 also reports the t-statistics of selected predictors as the ratio of the mean to standard deviation, with the means and standard deviations of their coefficients calculated over 100 random splittings. We see that the set of most frequently selected predictors for hard thresholding is identical to that for SICA, which is further a subset of that for the lasso. Some of these selected predictors have t-statistics with magnitude below 2, indicating less significance. We also observe that the coefficients for predictors sex and hdl estimated by all methods are negative. It is interesting to note that the interaction term sex∗age is found to be significant, although the predictor age is an insignificant variable on the basis of each method.

Acknowledgements

We sincerely thank the Joint Editor, an Associate Editor and a referee for their valuable comments that significantly improved the paper. This work was supported by National Science Foundation 'CAREER' awards DMS-0955316 and DMS-1150318 and grants DMS-0806030 and DMS-0906784, the 2010 Zumberge individual award from the University of Southern California's James H. Zumberge Faculty Research and Innovation Fund, and University of Southern California Marshall summer research funding.

Appendix A: Proofs of main results

A.1. Proof of theorem 1

The proof of theorem 1 contains two parts. The first part establishes the model selection consistency property of $\hat{\boldsymbol{\beta}}$ with a suitably chosen λ. The second part proves the oracle prediction properties by using the model selection consistency property from the first part.

A.1.1. Part 1: model selection consistency property

We prove $\mathrm{supp}(\hat{\boldsymbol{\beta}}) = \mathrm{supp}(\boldsymbol{\beta}_0)$ in two steps. In the first step, it will be shown that the number of non-zero elements in $\hat{\boldsymbol{\beta}}$ is no larger than s conditioning on event $\mathcal{E}$ defined in expression (A.1) in section A.2 of the on-line supplementary material, when $(c_2/c)\sqrt{\{(2s+1)\log(\tilde{p})/n\}} < \lambda < b_0$. We prove this by using the global optimality of $\hat{\boldsymbol{\beta}}$.

By lemma 1 and $\lambda < b_0$, any non-zero component of the true regression coefficient vector $\boldsymbol{\beta}_0$ or of the global minimizer $\hat{\boldsymbol{\beta}}$ is greater than λ, which ensures that $\|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 = \lambda^2\|\hat{\boldsymbol{\beta}}\|_0/2$ and $\|p_\lambda(\boldsymbol{\beta}_0)\|_1 = s\lambda^2/2$. Thus, $\|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 - \|p_\lambda(\boldsymbol{\beta}_0)\|_1 = (\|\hat{\boldsymbol{\beta}}\|_0 - s)\lambda^2/2$. Denote $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ by $\boldsymbol{\delta}$. Direct calculations yield

$$\begin{aligned} Q(\hat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}_0) &= 2^{-1}\|n^{-1/2}\mathbf{X}\boldsymbol{\delta}\|_2^2 - n^{-1}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\boldsymbol{\delta} + \|p_\lambda(\boldsymbol{\beta})\|_1 - \|p_\lambda(\boldsymbol{\beta}_0)\|_1 \\ &= 2^{-1}\|n^{-1/2}\mathbf{X}\boldsymbol{\delta}\|_2^2 - n^{-1}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\boldsymbol{\delta} + (\|\hat{\boldsymbol{\beta}}\|_0 - s)\lambda^2/2. \end{aligned} \tag{8}$$

However, conditional on event $\mathcal{E}$, we have

$$|n^{-1}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\boldsymbol{\delta}| \leqslant \|n^{-1}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\|_\infty\|\boldsymbol{\delta}\|_1 \leqslant c_2\sqrt{\{\log(\tilde{p})/n\}}\|\boldsymbol{\delta}\|_1 \leqslant c_2\sqrt{\{\log(\tilde{p})/n\}}\|\boldsymbol{\delta}\|_0^{1/2}\|\boldsymbol{\delta}\|_2. \tag{9}$$

In addition, by definition and condition 2, we obtain $\|\boldsymbol{\delta}\|_0 \leqslant \|\boldsymbol{\beta}_0\|_0 + \|\hat{\boldsymbol{\beta}}\|_0 < M$, with M being the robust spark of $\mathbf{X}$. Therefore, definition 1 entails

$$\|n^{-1/2}\mathbf{X}\boldsymbol{\delta}\|_2 \geqslant c\|\boldsymbol{\delta}\|_2. \tag{10}$$

Combining equation (8) with the inequalities (9) and (10) established above gives

$$Q(\hat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}_0) \geqslant 2^{-1}c^2\|\boldsymbol{\delta}\|_2^2 - c_2\sqrt{\{\log(\tilde{p})/n\}}\|\boldsymbol{\delta}\|_0^{1/2}\|\boldsymbol{\delta}\|_2 + (\|\hat{\boldsymbol{\beta}}\|_0 - s)\lambda^2/2. \tag{11}$$

Thus, the global optimality of $\hat{\boldsymbol{\beta}}$ ensures that

$$2^{-1}c^2\|\boldsymbol{\delta}\|_2^2 - c_2\sqrt{\{\log(\tilde{p})/n\}}\|\boldsymbol{\delta}\|_0^{1/2}\|\boldsymbol{\delta}\|_2 + (\|\hat{\boldsymbol{\beta}}\|_0 - s)\lambda^2/2 \leqslant 0.$$

Reorganizing this inequality and collecting terms, we obtain

$$\left[c\|\boldsymbol{\delta}\|_2 - \frac{c_2}{c}\sqrt{\left\{\frac{\log(\tilde{p})}{n}\right\}}\|\boldsymbol{\delta}\|_0^{1/2}\right]^2 - \left(\frac{c_2}{c}\right)^2\frac{\log(\tilde{p})}{n}\|\boldsymbol{\delta}\|_0 + (\|\hat{\boldsymbol{\beta}}\|_0 - s)\lambda^2 \leqslant 0,$$

which gives

$$(\|\hat{\boldsymbol{\beta}}\|_0 - s)\lambda^2 \leqslant \left(\frac{c_2}{c}\right)^2\frac{\log(\tilde{p})}{n}\|\boldsymbol{\delta}\|_0. \tag{12}$$

We next bound the value of $\|\hat{\boldsymbol{\beta}}\|_0$ by using inequality (12). Let $k = \|\hat{\boldsymbol{\beta}}\|_0$; then $\|\boldsymbol{\delta}\|_0 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_0 \leqslant k + s$. Thus, it follows from inequality (12) that

$$(k - s)\lambda^2 \leqslant \left(\frac{c_2}{c}\right)^2\frac{\log(\tilde{p})}{n}(k + s).$$

Organizing it in terms of k and s, we obtain

$$k\left\{\lambda^2 - \left(\frac{c_2}{c}\right)^2\frac{\log(\tilde{p})}{n}\right\} \leqslant s\left\{\lambda^2 + \left(\frac{c_2}{c}\right)^2\frac{\log(\tilde{p})}{n}\right\}. \tag{13}$$

Since $\lambda > (c_2/c)\sqrt{\{(2s+1)\log(\tilde{p})/n\}}$, we have $\lambda^2 - (c_2/c)^2(2s+1)\log(\tilde{p})/n > 0$ and $\lambda^2c^2n - c_2^2\log(\tilde{p}) > 2c_2^2 s\log(\tilde{p})$. Then it follows from inequality (13) that

$$k \leqslant s\frac{\lambda^2 + (c_2/c)^2\log(\tilde{p})/n}{\lambda^2 - (c_2/c)^2\log(\tilde{p})/n} = s\left\{1 + \frac{2c_2^2\log(\tilde{p})}{\lambda^2c^2n - c_2^2\log(\tilde{p})}\right\} < s + 1.$$

Therefore, the number of non-zero elements in $\hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\|_0 \leqslant s$.

The second step is based on the first step, where we shall use proof by contradiction to show that $\mathrm{supp}(\boldsymbol{\beta}_0) \subset \mathrm{supp}(\hat{\boldsymbol{\beta}})$ with the additional assumption $\lambda < b_0c/\sqrt{2}$ of theorem 1. Suppose that $\mathrm{supp}(\boldsymbol{\beta}_0) \not\subset \mathrm{supp}(\hat{\boldsymbol{\beta}})$; then the number of missed true coefficients $k = |\mathrm{supp}(\boldsymbol{\beta}_0)\backslash\mathrm{supp}(\hat{\boldsymbol{\beta}})| \geqslant 1$. Thus we have $\|\hat{\boldsymbol{\beta}}\|_0 \geqslant s - k$ and $\|\boldsymbol{\delta}\|_0 \leqslant \|\hat{\boldsymbol{\beta}}\|_0 + \|\boldsymbol{\beta}_0\|_0 \leqslant 2s$. Combining these two results with inequality (11) yields

$$Q(\hat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}_0) \geqslant \left[2^{-1}c^2\|\boldsymbol{\delta}\|_2 - c_2\sqrt{\left\{\frac{2s\log(\tilde{p})}{n}\right\}}\right]\|\boldsymbol{\delta}\|_2 - \frac{k\lambda^2}{2}. \tag{14}$$

For each $j \in \mathrm{supp}(\boldsymbol{\beta}_0) \setminus \mathrm{supp}(\hat{\boldsymbol{\beta}})$, we have $|\delta_j| = |\beta_{0,j}| \geqslant b_0$ with b_0 being the lowest signal strength in condition 2. Thus, $\|\boldsymbol{\delta}\|_2 \geqslant b_0\sqrt{k}$, which together with condition 2 entails

$$4^{-1}c^2\|\boldsymbol{\delta}\|_2 \geqslant 4^{-1}c^2b_0\sqrt{k} \geqslant 4^{-1}c^2b_0 > c_2\sqrt{\{2s\log(\tilde{p})/n\}}.$$

Thus, it follows from inequality (14) that

$$Q(\hat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}_0) \geqslant 4^{-1}c^2\|\boldsymbol{\delta}\|_2^2 - k\lambda^2/2 \geqslant 4^{-1}c^2kb_0^2 - k\lambda^2/2 > 0,$$

where the last step is because of the additional assumption $\lambda < b_0c/\sqrt{2}$. The above inequality contradicts the global optimality of $\hat{\boldsymbol{\beta}}$. Thus, we have $\mathrm{supp}(\boldsymbol{\beta}_0) \subset \mathrm{supp}(\hat{\boldsymbol{\beta}})$. Combining this with $\|\hat{\boldsymbol{\beta}}\|_0 \leqslant s$ from the first step, we know that $\mathrm{supp}(\hat{\boldsymbol{\beta}}) = \mathrm{supp}(\boldsymbol{\beta}_0)$.

It follows from lemma 1 and the characterization of the penalized least squares estimator in theorem 1 in Lv and Fan (2009) that the hard thresholded estimator $\hat{\boldsymbol{\beta}}$ on its support $\mathrm{supp}(\hat{\boldsymbol{\beta}})$ is exactly the ordinary least squares estimator constructed by using covariates in $\mathrm{supp}(\hat{\boldsymbol{\beta}})$. With the model selection consistency property proved above, we have the explicit form of $\hat{\boldsymbol{\beta}}$ on its support as $(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{y}$, where $\mathbf{X}_0$ is the submatrix of the design matrix $\mathbf{X}$ consisting of columns in $\mathrm{supp}(\boldsymbol{\beta}_0)$. Now we derive bounds for the prediction and estimation losses of $\hat{\boldsymbol{\beta}}$.

A.1.2. Part 2: prediction and estimation losses

The idea is to obtain the L_2-estimation loss bound by the global optimality of $\hat{\boldsymbol{\beta}}$, conditional on the event $\mathcal{E}_1 = \mathcal{E} \cap \mathcal{E}'$ with $\mathcal{E}$ and $\mathcal{E}'$ defined in lemma 2 in section A.2 of the on-line supplementary material.

Conditional on $\mathcal{E}_1$, we have $\|\boldsymbol{\delta}\|_0 \leqslant s$ by the model selection consistency property that was proved above. Thus, by the Cauchy–Schwarz inequality we have

$$|n^{-1}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}_0\boldsymbol{\delta}| \leqslant \|n^{-1}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}_0\|_\infty \|\boldsymbol{\delta}\|_1 \leqslant c_2'\sqrt{\left\{\frac{\log(n)}{n}\right\}}\|\boldsymbol{\delta}\|_1 \leqslant c_2'\sqrt{\left\{\frac{s\log(n)}{n}\right\}}\|\boldsymbol{\delta}\|_2. \tag{15}$$

Since expressions (8) and (10) are still true as they depend only on condition 2 and definition 1, it follows from inequalities (15) and the model selection consistency property $\|\hat{\boldsymbol{\beta}}\|_0 = s$ that

$$\begin{aligned} Q(\hat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}_0) &= 2^{-1}\|n^{-1}\mathbf{X}\boldsymbol{\delta}\|_2^2 - n^{-1}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\boldsymbol{\delta} + \frac{(\|\hat{\boldsymbol{\beta}}\|_0 - s)\lambda^2}{2} \\ &\geqslant 2^{-1}c^2\|\boldsymbol{\delta}\|_2^2 - n^{-1}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\boldsymbol{\delta} \geqslant \left[2^{-1}c^2\|\boldsymbol{\delta}\|_2 - c_2'\sqrt{\left\{\frac{s\log(n)}{n}\right\}}\right]\|\boldsymbol{\delta}\|_2. \end{aligned}$$

Then it follows from the global optimality of $\hat{\boldsymbol{\beta}}$ that $2^{-1}c^2\|\boldsymbol{\delta}\|_2 - c_2'\sqrt{\{s\log(n)/n\}} \leqslant 0$, which gives the L_2- and L_∞-estimation bounds as

$$\begin{aligned} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = \|\boldsymbol{\delta}\|_2 &\leqslant 2c^{-2}c_2'\sqrt{\{s\log(n)/n\}}, \\ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty \leqslant \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 &\leqslant 2c^{-2}c_2'\sqrt{\{s\log(n)/n\}}. \end{aligned}$$

For L_q-estimation loss with $1 \leqslant q < 2$, applying Hölder's inequality gives

$$\begin{aligned} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q = \Big(\sum_{j=1}^n |\delta_j|^q\Big)^{1/q} &\leqslant \Big(\sum_{j=1}^n |\delta_j|^2\Big)^{1/2}\Big(\sum_{\delta_j \neq 0} 1^{2/(2-q)}\Big)^{1/q-1/2} = \|\boldsymbol{\delta}\|_2\|\boldsymbol{\delta}\|_0^{1/q-1/2} \\ &\leqslant 2c^{-2}c_2's^{1/q}\sqrt{\{\log(n)/n\}}. \end{aligned} \tag{16}$$

Finally we prove the bound for oracle prediction loss. Since $\hat{\boldsymbol{\beta}}$ is the global minimizer, it follows from equation (8) and the model selection consistency property that conditioning on $\mathcal{E}_1$

$$\begin{aligned} 2^{-1/2}n^{-1/2}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2 &\leqslant \{n^{-1}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\boldsymbol{\delta} - (\|\hat{\boldsymbol{\beta}}\|_0 - s)\lambda^2/2\}^{1/2} \\ &\leqslant \{\|n^{-1}\mathbf{X}_0^{\mathrm{T}}\boldsymbol{\varepsilon}\|_\infty\|\boldsymbol{\delta}\|_1\}^{1/2} \leqslant c_2'c^{-1}\sqrt{\{2s\log(n)/n\}}, \end{aligned}$$

where the last step is because of the L_1-estimation bound proved above. This completes the proof.

A.2. Proof of theorem 2

In the proof of theorem 2, we apply mathematical techniques such as singular value decomposition and Taylor series expansion to study the explicit forms of risks of the refitted estimator $\hat{\boldsymbol{\beta}}_{\text{refitted}}$ under squared L_2-loss and squared prediction loss, and to find out the orders and leading terms of the optimal tuning parameter λ_1 and the corresponding minimized risks. The proof consists of two parts.

A.2.1. Part 1: risk properties for $\hat{\boldsymbol{\beta}}_{\text{refitted}}$ under L_q-estimation loss

We first consider the risk of $\hat{\boldsymbol{\beta}}_{\text{refitted}}$ under the squared L_2-loss and find the order and leading term of the corresponding optimal λ_1. The main idea is to divide the risk into two parts, and then to minimize the first part conditional on event $\mathcal{E}$ defined in expression (A.1) of the on-line supplementary material, and to show that the other part has a smaller order. By default, all arguments below are conditional on $\mathcal{E}$.

The proof of theorem 1 ensures that $\text{supp}(\hat{\boldsymbol{\beta}}) = \text{supp}(\boldsymbol{\beta}_0)$ conditional on event $\mathcal{E}$ under conditions 1 and 2. Thus, if we denote $\mathbf{X}_0$ as the oracle design matrix, then $\mathbf{X}_1 = \mathbf{X}_0$ and $s_1 = s_0$. Let I_s be the $s \times s$ identity matrix for a positive integer s. It follows that

$$\hat{\boldsymbol{\beta}}_{\text{refitted}} = (\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1 + \lambda_1 I_{s_1})^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} = (\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\boldsymbol{\beta}_0 + (\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-1}\mathbf{X}_0^{\mathrm{T}}\boldsymbol{\varepsilon},$$

where in the last step we used $\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$. So the difference between $\hat{\boldsymbol{\beta}}_{\text{refitted}}$ and $\boldsymbol{\beta}_0$ is

$$\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0 = -\lambda_1(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-1}\boldsymbol{\beta}_0 + (\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-1}\mathbf{X}_0^{\mathrm{T}}\boldsymbol{\varepsilon}.$$

Set $\boldsymbol{\mu} = -\lambda_1(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-1}\boldsymbol{\beta}_0$ and $A = \mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-1}$; then $\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0 = \boldsymbol{\mu} + A^{\mathrm{T}}\boldsymbol{\varepsilon}$. Thus, conditioning on $\mathcal{E}$ we have

$$\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2 = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} + 2\boldsymbol{\mu}^{\mathrm{T}}A^{\mathrm{T}}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^{\mathrm{T}}AA^{\mathrm{T}}\boldsymbol{\varepsilon}. \tag{17}$$

In view of equation (17), we consider the expectation of $\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2$ by using the decomposition

$$\begin{aligned} E\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2 &= E\{\mathbf{1}_{\mathcal{E}}\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2\} + E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2\} \\ &\leqslant E\{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} + 2\boldsymbol{\mu}^{\mathrm{T}}A^{\mathrm{T}}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^{\mathrm{T}}AA^{\mathrm{T}}\boldsymbol{\varepsilon}\} + E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2\}. \end{aligned}$$

Since $P(\mathcal{E}^{\mathrm{c}}) = o(1)$ by lemma 2 in section A.2 of the on-line supplementary material, the above inequality becomes an equation asymptotically by the dominated convergence theorem, which provides the basis for determining the orders of the risks. To ease the presentation, we do not distinguish between these two representations hereafter. The above decomposition, along with equation (17), condition 1 and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$, gives

$$\begin{aligned} E\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2 &\leqslant \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} + \sigma^2 \operatorname{tr}(AA^{\mathrm{T}}) + E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2\} \\ &= I_1(\lambda_1) + I_2(\lambda_1) + I_3(\lambda_1), \end{aligned} \tag{18}$$

where

$$I_1(\lambda_1) = \lambda_1^2 \boldsymbol{\beta}_0^{\mathrm{T}}(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-2}\boldsymbol{\beta}_0, \tag{19}$$

$$I_2(\lambda_1) = \sigma^2 \operatorname{tr}\{\mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-2}\mathbf{X}_0^{\mathrm{T}}\}, \tag{20}$$

$$I_3(\lambda_1) = E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2\}. \tag{21}$$

We analyse the first two terms, $I_1(\lambda_1)$ and $I_2(\lambda_1)$ in equation (18), by singular value decomposition. Since $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$ is symmetric and positive semidefinite, there exists an $s \times s$ orthonormal matrix P such that $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 = P^{\mathrm{T}}DP$, where D is a diagonal matrix with non-negative elements $d_i, i = 1, \ldots, s$, the eigenvalues of $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$. Replacing $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$ with $P^{\mathrm{T}}DP$, we obtain

$$\begin{aligned} \mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s &= P^{\mathrm{T}}(D + \lambda_1 I_s)P, \\ (\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-2} &= P^{\mathrm{T}}(D + \lambda_1 I_s)^{-2}P. \end{aligned}$$

Set $\mathbf{b} = (b_1, \ldots, b_s)^{\mathrm{T}} = P\boldsymbol{\beta}_0$. Then $\|\mathbf{b}\|_2 = \|\boldsymbol{\beta}_0\|_2$ and the first term becomes

$$I_1(\lambda_1) = \lambda_1^2 \boldsymbol{\beta}_0^{\mathrm{T}} P^{\mathrm{T}}(D + \lambda_1 I_s)^{-2}P\boldsymbol{\beta}_0 = \sum_{i=1}^{s} \frac{\lambda_1^2 b_i^2}{(d_i + \lambda_1)^2}$$

and the second term can be simplified as

$$\begin{aligned} I_2(\lambda_1) &= \sigma^2 \operatorname{tr}\{\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-2}\} = \sigma^2 \operatorname{tr}\{P^{\mathrm{T}}DPP^{\mathrm{T}}(D + \lambda_1 I_s)^{-2}P\} \\ &= \sigma^2 \operatorname{tr}\{D(D + \lambda_1 I_s)^{-2}\} = \sum_{i=1}^{s} \frac{\sigma^2 d_i}{(d_i + \lambda_1)^2}. \end{aligned}$$

Substituting the above two terms into equation (18), we obtain $E\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2 \leqslant f(\lambda_1) + I_3(\lambda_1)$, where

$$f(\lambda_1) = \sum_{i=1}^{s} \frac{\lambda_1^2 b_i^2}{(d_i + \lambda_1)^2} + \sum_{i=1}^{s} \frac{\sigma^2 d_i}{(d_i + \lambda_1)^2}. \tag{22}$$

Note that $f(\lambda_1)$ is a sum of two terms, with the first term increasing with λ_1 and the second term decreasing with λ_1. Besides $f(\lambda_1)$, we have another term $E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\hat{\boldsymbol{\beta}}_{\text{refitted}} - \boldsymbol{\beta}_0\|_2^2\}$, which will be shown to be of a strictly smaller order than $f(\lambda_1)$.

A.2.2. Part 1.1: identifying orders of optimal λ_1 and corresponding $f(\lambda_1)$ for L_2-risk

It is difficult to find the exact λ_1 that minimizes $f(\lambda_1)$ since the denominators in the sum are different, but we can surely identify its order, in the following three steps.

First of all, we claim that $c^2 n \leqslant d_i \leqslant c_3^2 n$ for all i. It suffices to show that the maximum and minimum eigenvalues of $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$, which are denoted as $\lambda_{\max}$ and $\lambda_{\min}$, can be bounded as $c^2 \leqslant \lambda_{\min}/n \leqslant \lambda_{\max}/n \leqslant c_3^2$. For

this, note that, as $\mathbf{X}_0$ is the submatrix of $\mathbf{X}$ formed by columns with indices in $\mathrm{supp}(\boldsymbol{\beta}_0)$ and $|\mathrm{supp}(\boldsymbol{\beta}_0)| = s < M/2$ by condition 2, we have $\lambda_{\min}/n \geqslant c^2$ by the property of the robust spark. However, since we assumed that $|\mathrm{supp}(\boldsymbol{\beta}_0)| < M/2$, condition 3 ensures that $\lambda_{\max}/n \leqslant c_3^2$. So we have proved $c^2 \leqslant \lambda_{\min}/n \leqslant \lambda_{\max}/n \leqslant c_3^2$. Since d_is are the eigenvalues of $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$, it follows that

$$c^2 n \leqslant \lambda_{\min} \leqslant d_i \leqslant \lambda_{\max} \leqslant c_3^2 n. \tag{23}$$

In fact, the same argument applies to any submatrix of $\mathbf{X}$ with number of columns less than $M/2$. Since $|\mathrm{supp}(\hat{\boldsymbol{\beta}}_1)| \leqslant \|\hat{\boldsymbol{\beta}}\|_0 < M/2$ by expression (6), we also have

$$c^2 n \leqslant \lambda_{\min}(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1) \leqslant \lambda_{\max}(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1) \leqslant c_3^2 n, \tag{24}$$

which will be used later for analysing $I_3(\lambda_1)$.

Second, we show that the optimal λ_1 that minimizes $f(\lambda_1)$, which is denoted as $\lambda_{1,\mathrm{opt}}$, is of the order $o(n)$. If it is not true, then there is some constant $k > 0$ such that $\lambda_{1,\mathrm{opt}} \geqslant kn$. By inequalities (23) and condition 3, and since $\|\boldsymbol{\beta}_0\|_2 \geqslant O(1)$, we have

$$f(\lambda_{1,\mathrm{opt}}) \geqslant \sum_{i=1}^{s} \frac{\lambda_{1,\mathrm{opt}}^2 b_i^2}{(d_i + \lambda_{1,\mathrm{opt}})^2} \geqslant \sum_{i=1}^{s} \frac{k^2 n^2 b_i^2}{(c_3^2 n + kn)^2} = \sum_{i=1}^{s} \frac{k^2 b_i^2}{(c_3^2 + k)^2} = \frac{k^2 \|\boldsymbol{\beta}_0\|_2^2}{(c_3^2 + k)^2} \geqslant O(1). \tag{25}$$

However, by the optimality of $\lambda_{1,\mathrm{opt}}$, $f(\lambda_{1,\mathrm{opt}}) \leqslant f(0) = \sum_{i=1}^{s} \sigma^2/d_i = O(s/n) = o(1)$. It is a contradiction and thus we must have $\lambda_{1,\mathrm{opt}} = o(n)$.

In the third step, we go one step further to show that the order of $\lambda_{1,\mathrm{opt}}$ is indeed $O(s\|\boldsymbol{\beta}_0\|_2^{-2}) + O(s^2 n^{-1}\|\boldsymbol{\beta}_0\|_2^{-4})$ by applying Taylor series expansion on $f'(\lambda_1)$ with $\lambda_{1,\mathrm{opt}} = o(n)$. Direct calculations yield

$$f'(\lambda_1) = \sum_{i=1}^{s} \frac{2\lambda_1 b_i^2 d_i}{(d_i + \lambda_1)^3} - \sum_{i=1}^{s} \frac{2\sigma^2 d_i}{(d_i + \lambda_1)^3} = \sum_{i=1}^{s} \frac{2\lambda_1 b_i^2 d_i - 2\sigma^2 d_i}{(d_i + \lambda_1)^3}. \tag{26}$$

Since the optimal λ_1 satisfies $f'(\lambda_{1,\mathrm{opt}}) = 0$, we have

$$\sum_{i=1}^{s} \frac{\lambda_{1,\mathrm{opt}} b_i^2 d_i}{(d_i + \lambda_{1,\mathrm{opt}})^3} = \sum_{i=1}^{s} \frac{\sigma^2 d_i}{(d_i + \lambda_{1,\mathrm{opt}})^3}. \tag{27}$$

We shall rearrange this equation as a quadratic equation for $\lambda_{1,\mathrm{opt}}$ by using Taylor series expansion. Since it has been proved that $\lambda_{1,\mathrm{opt}} = o(n)$ or, equivalently, $\lambda_{1,\mathrm{opt}} = o(d_i)$ for each $1 \leqslant i \leqslant s$, we can apply Taylor series expansion with Lagrange remainder to deal with the two fractions in equation (27). For the left-hand side of equation (27), we have

$$\sum_{i=1}^{s} \frac{\lambda_{1,\mathrm{opt}} b_i^2 d_i}{(d_i + \lambda_{1,\mathrm{opt}})^3} = \sum_{i=1}^{s} \lambda_{1,\mathrm{opt}} b_i^2 d_i \left\{ \frac{1}{d_i^3} - \frac{3\lambda_{1,\mathrm{opt}}}{(d_i + \omega_i)^4} \right\} = \lambda_{1,\mathrm{opt}} \left\{ \sum_{i=1}^{s} \frac{b_i^2}{d_i^2} - \sum_{i=1}^{s} \frac{3 b_i^2 d_i \lambda_{1,\mathrm{opt}}}{(d_i + \omega_i)^4} \right\},$$

where ω_is are numbers between 0 and $\lambda_{1,\mathrm{opt}}$. For the right-hand side of equation (27), we obtain

$$\sum_{i=1}^{s} \frac{\sigma^2 d_i}{(d_i + \lambda_{1,\mathrm{opt}})^3} = \sum_{i=1}^{s} \sigma^2 d_i \left\{ \frac{1}{d_i^3} - \frac{3\lambda_{1,\mathrm{opt}}}{d_i^4} + \frac{6\lambda_{1,\mathrm{opt}}^2}{(d_i + \gamma_i)^5} \right\} = \sum_{i=1}^{s} \frac{\sigma^2}{d_i^2} - \lambda_{1,\mathrm{opt}} \sum_{i=1}^{s} \frac{3\sigma^2}{d_i^3} + \sum_{i=1}^{s} \frac{6\sigma^2 d_i \lambda_{1,\mathrm{opt}}^2}{(d_i + \gamma_i)^5},$$

where γ_is are numbers between 0 and $\lambda_{1,\mathrm{opt}}$. Equalling the two sides yields

$$\lambda_{1,\mathrm{opt}} \left\{ \sum_{i=1}^{s} \frac{b_i^2}{d_i^2} - \sum_{i=1}^{s} \frac{3 b_i^2 d_i \lambda_{1,\mathrm{opt}}}{(d_i + \omega_i)^4} \right\} = \sum_{i=1}^{s} \frac{\sigma^2}{d_i^2} - \lambda_{1,\mathrm{opt}} \sum_{i=1}^{s} \frac{3\sigma^2}{d_i^3} + \sum_{i=1}^{s} \frac{6\sigma^2 d_i \lambda_{1,\mathrm{opt}}^2}{(d_i + \gamma_i)^5}.$$

Reorganizing it in terms of the power of $\lambda_{1,\mathrm{opt}}$, we obtain

$$\left\{ \sum_{i=1}^{s} \frac{6\sigma^2 d_i}{(d_i + \gamma_i)^5} + \sum_{i=1}^{s} \frac{3 b_i^2 d_i}{(d_i + \omega_i)^4} \right\} \lambda_{1,\mathrm{opt}}^2 - \left(\sum_{i=1}^{s} \frac{b_i^2}{d_i^2} + \sum_{i=1}^{s} \frac{3\sigma^2}{d_i^3} \right) \lambda_{1,\mathrm{opt}} + \sum_{i=1}^{s} \frac{\sigma^2}{d_i^2} = 0. \tag{28}$$

Its solution for $\lambda_{1,\mathrm{opt}}$ is $\{-b - \sqrt{(b^2 - 4ac)}\}/2a$, where

$$a=\sum_{i=1}^{s}\frac{6\sigma^2 d_i}{(d_i+\gamma_i)^5}+\sum_{i=1}^{s}\frac{3b_i^2 d_i}{(d_i+\omega_i)^4},$$

$$b=-\left(\sum_{i=1}^{s}\frac{b_i^2}{d_i^2}+\sum_{i=1}^{s}\frac{3\sigma^2}{d_i^3}\right)$$

and $c=\Sigma_{i=1}^{s}\sigma^2/d_i^2$. We drop the solution $\lambda_{1,\text{opt}}=\{-b+\sqrt{(b^2-4ac)}\}/2a$ since its order is $O(n)$, which can be proved by analysing the orders of a, b and c as follows.

With $c^2 n\leqslant d_i\leqslant c_3^2 n$, we can immediately calculate the orders of terms in a, b and c as

$$\sum_{i=1}^{s}\frac{6\sigma^2 d_i}{(d_i+\gamma_i)^5}=O(sn^{-4}),$$

$$\sum_{i=1}^{s}\frac{3b_i^2 d_i}{(d_i+\omega_i)^4}=O(n^{-3}\|\boldsymbol{\beta}_0\|_2^2),$$

$$\sum_{i=1}^{s}\frac{b_i^2}{d_i^2}=O(n^{-2}\|\boldsymbol{\beta}_0\|_2^2),$$

$$\sum_{i=1}^{s}\frac{3\sigma^2}{d_i^3}=O(sn^{-3}),$$

$$\sum_{i=1}^{s}\frac{\sigma^2}{d_i^2}=O(sn^{-2}).$$

Then we have $a=O(n^{-3}\|\boldsymbol{\beta}_0\|_2^2)$, $b=O(n^{-2}\|\boldsymbol{\beta}_0\|_2^2)$ and $c=O(sn^{-2})$. We know that $b^2=O(n^{-4}\|\boldsymbol{\beta}_0\|_2^4)$ is the leading term in b^2-4ac since $4ac=O(sn^{-5}\|\boldsymbol{\beta}_0\|_2^2)$. Since $b<0$, both $-b$ and $\sqrt{(b^2-4ac)}$ are positive and they are of the same order $O(n^{-2}\|\boldsymbol{\beta}_0\|_2^2)$. So the order for $\{-b+\sqrt{(b^2-4ac)}\}/2a$ is

$$O(n^{-2}\|\boldsymbol{\beta}_0\|_2^2)/O(n^{-3}\|\boldsymbol{\beta}_0\|_2^2)=O(n).$$

Since we have proved $\lambda_{1,\text{opt}}=o(n)$ before, this rules out the possibility of $\lambda_{1,\text{opt}}=\{-b+\sqrt{(b^2-4ac)}\}/2a$, which entails that $\lambda_{1,\text{opt}}=\{-b-\sqrt{(b^2-4ac)}\}/2a$. We further show that $\lambda_{1,\text{opt}}$ has a leading order $O(s\|\boldsymbol{\beta}_0\|_2^{-2})$ followed by a secondary order $O(s^2n^{-1}\|\boldsymbol{\beta}_0\|_2^{-4})$, in section B.1 of the on-line supplementary material.

Plugging $\lambda_{1,\text{opt}}=O(s\|\boldsymbol{\beta}_0\|_2^{-2})+O(s^2n^{-1}\|\boldsymbol{\beta}_0\|_2^{-4})$ into $f(\lambda_1)$ defined in equation (22), we obtain

$$\sum_{i=1}^{s}\frac{\lambda_{1,\text{opt}}^2 b_i^2}{(d_i+\lambda_{1,\text{opt}})^2}=O\left(\frac{s^2}{n^2\|\boldsymbol{\beta}_0\|_2^2}\right)+O\left(\frac{s^3}{n^3\|\boldsymbol{\beta}_0\|_2^4}\right),$$

$$\sum_{i=1}^{s}\frac{\sigma^2 d_i}{(d_i+\lambda_{1,\text{opt}})^2}=O\left(\frac{s}{n}\right)+O\left(\frac{s^2}{n^2\|\boldsymbol{\beta}_0\|_2^2}\right).$$

Thus, the order for $f(\lambda_{1,\text{opt}})$ is $O(s/n)+O(s^2n^{-2}\|\boldsymbol{\beta}_0\|_2^{-2})$.

A.2.3. Part 1.2: bounding the leading term of order $O(s\|\boldsymbol{\beta}_0\|_2^{-2})$ in $\lambda_{1,\text{opt}}$

In fact, the leading order $O(s\|\boldsymbol{\beta}_0\|_2^{-2})$ in $\lambda_{1,\text{opt}}$ comes from $-t/(2a)$, which equals $-c/b$ since $4ac=2bt$. Plugging in the definitions of b and c gives

$$-\frac{c}{b}=\frac{\sum_{i=1}^{s}\sigma^2/d_i^2}{\sum_{i=1}^{s}b_i^2/d_i^2+\sum_{i=1}^{s}3\sigma^2/d_i^3}.$$

By inequalities (23), we see that $\Sigma_{i=1}^{s}3\sigma^2/d_i^3$ is a smaller order term compared with $\Sigma_{i=1}^{s}b_i^2/d_i^2$. Thus, the leading term for $-c/b$ is $(\Sigma_{i=1}^{s}\sigma^2/d_i^2)(\Sigma_{i=1}^{s}b_i^2/d_i^2)^{-1}$.

Recall that $\lambda_{\min}$ and $\lambda_{\max}$ stand for the smallest and largest eigenvalues of $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$. With $\lambda_{\min} \leqslant d_i \leqslant \lambda_{\max}$ and $\Sigma_{i=1}^{s} b_i^2 = \|\boldsymbol{\beta}_0\|_2^2$, we obtain that the leading term for $-c/b$ can be bounded as

$$\frac{s\sigma^2}{\|\boldsymbol{\beta}_0\|_2^2}\frac{\lambda_{\min}^2}{\lambda_{\max}^2} = \frac{\sum_{i=1}^{s}\sigma^2/\lambda_{\max}^2}{\sum_{i=1}^{s} b_i^2/\lambda_{\min}^2} \leqslant \frac{\sum_{i=1}^{s}\sigma^2/d_i^2}{\sum_{i=1}^{s} b_i^2/d_i^2} \leqslant \frac{\sum_{i=1}^{s}\sigma^2/\lambda_{\min}^2}{\sum_{i=1}^{s} b_i^2/\lambda_{\max}^2} = \frac{s\sigma^2}{\|\boldsymbol{\beta}_0\|_2^2}\frac{\lambda_{\max}^2}{\lambda_{\min}^2}.$$

So the leading term for $\lambda_{1,\mathrm{opt}}$, which is $O(s\|\boldsymbol{\beta}_0\|_2^{-2})$, is between $(s\sigma^2/\|\boldsymbol{\beta}_0\|_2^2)\lambda_{\min}^2/\lambda_{\max}^2$ and $(s\sigma^2/\|\boldsymbol{\beta}_0\|_2^2) \times \lambda_{\max}^2/\lambda_{\min}^2$. In particular, when $\lambda_{\max} = \lambda_{\min}$, which implies that all d_is are the same, we can solve equation (26) readily to obtain $\lambda_{1,\mathrm{opt}} = s\sigma^2/\|\boldsymbol{\beta}_0\|_2^2$, which coincides with the above bounds for the leading term.

A.2.4. Part 1.3: bounding term $E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0\|_2^2\}$ in equation (18)

Now let us turn to the last term in equation (18): $I_3(\lambda_1) = E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0\|_2^2\}$. We prove in section B.2 of the on-line supplementary material that, compared with $f(\lambda_{1,\mathrm{opt}})$, the order of $E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0\|_2^2\}$ is much smaller. Thus $f(\lambda_{1,\mathrm{opt}})$ is the leading term of $E\|\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0\|_2^2$ and

$$E\|\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0\|_2^2 = O(s/n) + O(s^2 n^{-2}\|\boldsymbol{\beta}_0\|_2^{-2}) \tag{29}$$

for the optimal choice of λ_1.

A.2.5. Part 1.4: bounds for the L_q-risks

On the basis of the risk for squared L_2-loss above, we can derive the bounds for the risks of L_q-losses by using Hölder's inequality, as shown in section B.3 of the on-line supplementary material. The bound under L_∞-loss follows directly from the inequality $E\|\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0\|_\infty \leqslant E\|\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0\|_2$.

A.2.6. Part 2: risk properties for $\hat{\boldsymbol{\beta}}_{\mathrm{refitted}}$ under prediction loss

In this part, we shall find the risk property of the prediction loss for the refitted estimator $\hat{\boldsymbol{\beta}}_{\mathrm{refitted}}$ in a very similar way to that before.

Similarly to equation (17), we have $\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2 = (\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\varepsilon}^{\mathrm{T}}A)\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0(\boldsymbol{\mu} + A^{\mathrm{T}}\boldsymbol{\varepsilon}) = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\boldsymbol{\mu} + 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0A^{\mathrm{T}}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^{\mathrm{T}}A\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0A^{\mathrm{T}}\boldsymbol{\varepsilon}$. Taking expectation to the prediction loss, we have

$$\begin{aligned} n^{-1}E\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2 &= n^{-1}E\{\mathbf{1}_{\mathcal{E}}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2\} + n^{-1}E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2\} \\ &\leqslant n^{-1}E\{(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\boldsymbol{\mu} + 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0A^{\mathrm{T}}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^{\mathrm{T}}A\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0A^{\mathrm{T}}\boldsymbol{\varepsilon})\} + n^{-1}E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2\} \\ &= n^{-1}\{\boldsymbol{\mu}^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\boldsymbol{\mu} + \sigma^2\operatorname{tr}(A\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0A^{\mathrm{T}})\} + n^{-1}E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2\}. \end{aligned} \tag{30}$$

Using definitions $\boldsymbol{\mu} = -\lambda_1(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-1}\boldsymbol{\beta}_0$, $A = \mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda_1 I_s)^{-1}$ and $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 = P^{\mathrm{T}}DP$, we obtain

$$\boldsymbol{\mu}^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\boldsymbol{\mu} = \lambda_1^2(P\boldsymbol{\beta}_0)^{\mathrm{T}}(D + \lambda_1 I_s)^{-1}D(D + \lambda_1 I_s)^{-1}P\boldsymbol{\beta}_0 = \lambda_1^2\sum_{i=1}^{s}\frac{b_i^2 d_i}{(d_i + \lambda_1)^2},$$

and

$$\begin{aligned} \sigma^2\operatorname{tr}(A\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0A^{\mathrm{T}}) &= \sigma^2\operatorname{tr}\{P^{\mathrm{T}}DPP^{\mathrm{T}}(D + \lambda_1 I_s)^{-1}PP^{\mathrm{T}}DPP^{\mathrm{T}}(D + \lambda_1 I_s)^{-1}P\} \\ &= \sigma^2\operatorname{tr}\{D(D + \lambda_1 I_s)^{-1}D(D + \lambda_1 I_s)^{-1}\} = \sum_{i=1}^{s}\frac{\sigma^2 d_i^2}{(d_i + \lambda_1)^2}. \end{aligned}$$

Plugging the above two terms into equation (30) yields

$$\frac{1}{n}E\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2 \leqslant \frac{1}{n}\left\{\lambda_1^2\sum_{i=1}^{s}\frac{b_i^2 d_i}{(d_i + \lambda_1)^2} + \sum_{i=1}^{s}\frac{\sigma^2 d_i^2}{(d_i + \lambda_1)^2}\right\} + \frac{1}{n}E\{\mathbf{1}_{\mathcal{E}^{\mathrm{c}}}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2\}.$$

Set

$$g(\lambda_1) = \lambda_1^2\sum_{i=1}^{s}\frac{b_i^2 d_i}{(d_i + \lambda_1)^2} + \sum_{i=1}^{s}\frac{\sigma^2 d_i^2}{(d_i + \lambda_1)^2},$$

and note that it can be transformed from $f(\lambda_1)$ by multiplying d_i in the ith term of each sum. Denote the optimal λ_1 for minimizing $g(\lambda_1)$ as $\lambda'_{1,\mathrm{opt}}$. In view of expressions (25) and (28), the same argument applies; we also obtain $\lambda'_{1,\mathrm{opt}} = o(n)$ and, consequently, we can deduce that $\lambda'_{1,\mathrm{opt}} = O(s\|\boldsymbol{\beta}_0\|_2^{-2}) + O(s^2 n^{-1}\|\boldsymbol{\beta}_0\|_2^{-4})$ as the ratio of orders does not change. Then we can prove that the leading term for $\lambda'_{1,\mathrm{opt}}$ is between $(s\sigma^2/\|\boldsymbol{\beta}_0\|_2^2)\lambda_{\min}/\lambda_{\max}$ and $(s\sigma^2/\|\boldsymbol{\beta}_0\|_2^2)\lambda_{\max}/\lambda_{\min}$ and $g(\lambda'_{1,\mathrm{opt}}) = O(s) + O(s^2 n^{-1}\|\boldsymbol{\beta}_0\|_2^{-2})$. The term $n^{-1}E\{\mathbf{1}_{\mathcal{E}^c}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2\}$ can be shown to have a smaller order than $O(s^2 n^{-2}\|\boldsymbol{\beta}_0\|_2^{-2})$ similarly to before. Therefore, $n^{-1}E\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{refitted}} - \boldsymbol{\beta}_0)\|_2^2 = O(s/n) + O(s^2 n^{-2}\|\boldsymbol{\beta}_0\|_2^{-2})$, which concludes the proof.

References

Antoniadis, A. (1996) Smoothing noisy data with tapered coiflets series. *Scand. J. Statist.*, **23**, 313–330.

Antoniadis, A. and Fan, J. (2001) Regularization of wavelet approximations (with discussion). *J. Am. Statist. Ass.*, **96**, 939–967.

Barron, A., Birge, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probab. Theor. Reltd Flds*, **113**, 301–413.

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.*, **1**, 169–194.

Candès, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.*, **35**, 2313–2404.

Donoho, D. and Elad, M. (2003) Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proc. Natn. Acad. Sci. USA*, **100**, 2197–2202.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion). *Ann. Statist.*, **32**, 407–499.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, J. and Lv, J. (2011) Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theor.*, **57**, 5467–5484.

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

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**, 302–332.

van de Geer, S., Bühlmann, P. and Zhou, S. (2011) The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Statist.*, **5**, 688–749.

James, W. and Stein, C. (1961) Estimation with quadratic loss. *In Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 361–379. Berkeley: University of California Press.

Lin, W. and Lv, J. (2013) High-dimensional sparse additive hazards regression. *J. Am. Statist. Ass.*, **108**, 247–264.

Lv, J. and Fan, Y. (2009) A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, **37**, 3498–3528.

Stein, C. (1956) Inadmissibility of the usual estimator for the mean of a multivariate distribution. *In Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 197–206. Berkeley: University of California Press.

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

Wu, T. T. and Lange, K. (2008) Coordinate descent algorithms for Lasso penalized regression. *Ann. Appl. Statist.*, **2**, 224–244.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.*, **36**, 1509–1566.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material to "High-dimensional thresholded regression and shrinkage effect"'.