Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Achieving the oracle property of OEM with nonconvex penalties

Shifeng Xiong, Bin Dai & Peter Z. G. Qian

To cite this article: Shifeng Xiong, Bin Dai & Peter Z. G. Qian (2017) Achieving the oracle property of OEM with nonconvex penalties, Statistical Theory and Related Fields, 1:1, 28-36, DOI: 10.1080/24754269.2017.1326079

To link to this article: https://doi.org/10.1080/24754269.2017.1326079

Published online: 19 May 2017.

STATISTICAL THEORY AND RELATED FIELDS, 2017
VOL. 1, NO. 1, 28–36
https://doi.org/10.1080/24754269.2017.1326079

Taylor & Francis
Taylor & Francis Group

Achieving the oracle property of OEM with nonconvex penalties

Shifeng Xiong[a], Bin Dai[b] and Peter Z. G. Qian[c]

[a]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China; [b]Tower Research Capital, New York, NY, USA; [c]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT

Thepenalised least square estimator of non-convex penalties such as the smoothly clipped absolute deviation (SCAD) and the minimax concave penalty (MCP) is highly nonlinear and has many local optima. Finding a local solution to achieve the so-called oracle property is a challenging problem. We show that the orthogonalising EM (OEM) algorithm can indeed find such a local solution with the oracle property under some regularity conditions for a moderate but diverging number of variables.

ARTICLE HISTORY
Received 7 March 2017
Revised 26 April 2017
Accepted 30 April 2017

KEYWORDS
EM algorithm; non-convex optimisation; penalised regression; variable selection

1. Introduction

Consider a linear model

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

where $\boldsymbol{X} = (x_{ij})$ is the $n \times p$ regression matrix, $\mathbf{y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is the vector of random error. Throughout, let $\|\cdot\|$ denote the Euclidean norm. A *regularised least squares estimator* of $\boldsymbol{\beta}$ with the smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001) is given by solving

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{n}\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2\sum_{j=1}^{p} P(|\beta_j|;\ a, b) \right], \tag{2}$$

where for $\theta > 0$,

$$P'(\theta; a, b) = bI(\theta \leqslant b) + (ab - \theta)_+ I(\theta > b)/(a-1), \tag{3}$$

$a > 2$ and $b > 0$ are the tuning parameters, and I is the indicator function. In order to apply the penalty P equally on all the variables, $\boldsymbol{X}$ can be standardised so that

$$\sum_{i=1}^{n} x_{ij}^2 = n, \ \text{ for } j = 1, \ldots, p. \tag{4}$$

Fan and Li (2001) used the SCAD penalty in (3) to achieve simultaneous estimation and variable selection. Zhang (2010) studied a class of non-convex penalties and introduced the minimax concave penalty (MCP) method, which replaces P in (2) with P_{MCP} satisfying

$$P'_{\mathrm{MCP}}(\theta;\ a, b) = (b - \theta/a)I(\theta \leqslant ab), \tag{5}$$

where $a > 1, b > 0$, and $\theta > 0$. Non-convex penalties are now commonly used in statistics. Existing algorithms for solving the SCAD or MCP problem include local quadratic approximation (Fan & Li, 2001; Hunter & Li, 2005), local linear approximation (Zou & Li, 2008), the ConCave Convex procedure (CCCP) (Kim, Choi, & Oh, 2008), the minimisation by iterative soft thresholding algorithm (Schifano, Strawderman, & Wells, 2010), and the coordinate descent algorithm (Breheny & Huang, 2011; Mazumder, Friedman, & Hastie, 2011; Tseng, 2001; Tseng & Yun, 2009), among others.

The non-convex nature of the SCAD and MCP penalties has an interesting interface between computation and statistical properties. Fan and Li (2001) proposed the SCAD penalty and also an important property, called the oracle property. An estimator of $\boldsymbol{\beta}$ having this property can not only select the correct submodel asymptotically, but also estimate the nonzero coefficients as efficiently as if the correct submodel were known in advance. Fan and Li (2001) proved that there exists a local solution of the SCAD problem in (2) enjoying this property for fixed p. The corresponding results with a diverging p were presented in Fan and Peng (2004) and Fan and Lv (2011). However, it is challenging to single out the one with the oracle property since the non-convex penalised problems like SCAD or MCP can have many local optima (Huo & Chen, 2010; Huo & Ni, 2007). On the algorithmic aspect, different initial points in a specific algorithm can yield different solutions for $n \geqslant p$ or $n < p$ case. For example, let all rows of $\boldsymbol{X}$ in (1) be identically and independently generated from a zero-mean multi-normal distribution with correlation 0.5, $\boldsymbol{\beta} = \mathbf{0}$, and the components of $\boldsymbol{\varepsilon}$ be identically and independently generated from $N(0, 3^2)$. We use the coordinate descent algorithm to solve the SCAD problem with $a = 3.7$ and $b = 1/(n\ \log(n))$. For one simulation, we use 100 initial points, which are independently generated from a multi-normal distribution

CONTACT Peter Z. G. Qian peterq@stat.wisc.edu

Table 1. Averages and maxima (in parentheses) of the solution numbers of SCAD.

$n = 20, p = 10$	$n = 50, p = 20$	$n = 100, p = 30$	$n = 150, p = 40$	$n = 200, p = 50$
3.65 (35)	7.26 (62)	3.56 (54)	7.06 (59)	4.62 (86)

$N(\mathbf{0}, 10^2\boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ denotes the $p \times p$ identity matrix. The solution numbers for different pairs of (n, p) over 100 simulations are described in Table 1.

To get a good estimate from the multiple solutions, one needs to elaborately select the algorithm and the initial point. Fan and Li (2001) suggested using the ordinary least squares (OLS) estimator as the initial point in their local quadratic approximation algorithm for $n > p$. This method performs well in simulations but lacks theoretical justification. In recent years, related theoretical study has been carried out. Zhang (2010) devised a novel penalised linear unbiased selection algorithm and proved that it can achieve selection consistency of local solutions to MCP. Loh and Wainwright (2013) discussed the convergence rate of local solutions given by certain algorithms for a general class of non-convex penalised problems. When the oracle property is concerned, Zou and Li (2008) showed that the one-step solution of the local linear approximation algorithm, which stops the algorithm after one iteration, has the oracle property with a good initial estimator for a fixed p. Fan, Xue, and Zou (2014) extended such a result to a general penalised estimation problem. Wang, Kim, and Li (2013) proposed a two-step CCCP with different tuning parameters in the two steps, called calibrated CCCP, and showed that it produces the oracle estimator with probability approaching one. These theoretical results are useful for high-dimensional statistics. However, since a local solution is achieved when the iteration number goes to infinity, after only one or several iterations, the estimator is not guaranteed to be a local minimum of the non-convex problem in finite-sample cases. To find a local solution with the oracle property, we need to study the oracle property of a k-step estimator as k goes to infinity. As Meng (2008) pointed out, if we believe that the SCAD objective function is a valid criterion for sparse estimation, then the estimator from a monotonic algorithm is expected to become better as the iteration number increases. Therefore, a k-step estimator with the oracle property as k goes to infinity matches Fan and Li's original purpose of introducing the SCAD penalty. To the best of our knowledge, no such estimator has been presented in the literature, even for the case of $n > p$.

In this paper, we will show a rather surprising result. In our early results summarised in an unpublished paper (Xiong, Dai, & Qian, 2011), we introduced the orthogonalising EM (OEM) algorithm for general least squares problems. For the SCAD and MCP problems, each OEM iteration has a simple closed form. This feature makes it possible to study the theoretical properties of a k-step estimator as k goes to infinity, and thus motivates us to consider whether the local solution of SCAD or MCP given by OEM has the oracle property. Since OEM is more suitable for big tall data with $n > p$ (Xiong, Dai, Huling, & Qian, 2016), here our study mainly focuses on such cases. We prove that an OEM sequence for SCAD or MCP can indeed achieve the oracle property if the iteration number goes to infinity with an initial estimator having certain consistency property. We allow the dimensionality p to depend on n of the order $p = O(n^q)$ with $q \in [0, 3/2)$. For p exceeding this order, our result can be applied to the submodel after a screening stage (Fan & Lv, 2008), which reduces the initial p to the order $p = o(n)$. We therefore present a detailed discussion on the applications of our result to the case of $n > p$. For this case, with the OLS estimator being the initial estimator, our result holds for a broad class of deterministic or random $\boldsymbol{X}$ in (1), which can be allowed to be nearly degenerate.

It should be pointed out that, the technical report (Xiong et al., 2011) summarised some earlier results we have got on OEM. We have divided the report into two papers for publication. One paper (Xiong et al., 2016) focuses on the details of the algorithm. The current paper deals with the oracle property. The two papers have no overlapping.

Section 2 reviews the OEM algorithm. Section 3 presents the main result that the local solution of SCAD given by OEM has the oracle property. Section 4 discusses our result in the case of $n > p$. Section 5 presents simulation results. Section 6 concludes with some discussion. All proofs are given in the Appendix.

2. The OEM algorithm

Consider the SCAD problem in (2) with the regression matrix $\boldsymbol{X}$ standardised as in (4). For a matrix, denote by $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ its largest and smallest eigenvalues, respectively. In the OEM algorithm, the first step of OEM is active orthogonalisation, which computes $\boldsymbol{\Delta}$ such that

$$\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}'\boldsymbol{\Delta} = nd_n\boldsymbol{I}_p \tag{6}$$

with $d_n \geqslant \gamma_1 = \lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}/n)$. Therefore, $\boldsymbol{X}_c = (\boldsymbol{X}' \ \boldsymbol{\Delta}')'$ is column orthogonal. Consider the linear model

$$\mathbf{y}_c = \boldsymbol{X}_c\boldsymbol{\beta} + \boldsymbol{\varepsilon}_c, \tag{7}$$

where $\boldsymbol{y}_c = (\boldsymbol{y}', \boldsymbol{y}_m')'$ is the complete response vector including a missing part $\boldsymbol{y}_m$. Based on the complete

model (7), we can solve (2) by iteratively imputing $\mathbf{y}_{\mathrm{m}}$. Let $\boldsymbol{\beta}^{(0)}$ be an initial point. For $k = 0, 1, \ldots,$ impute $\mathbf{y}_{\mathrm{m}}$ as $\mathbf{y}_{\mathrm{imp}} = \boldsymbol{\Delta}\boldsymbol{\beta}^{(k)}$, let $\mathbf{y}_{\mathrm{c,\,imp}} = (\mathbf{y}',\ \mathbf{y}'_{\mathrm{imp}})'$, and solve

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{n} \|\mathbf{y}_{\mathrm{c,\,imp}} - \boldsymbol{X}_{\mathrm{c}}\boldsymbol{\beta}\|^2 + 2\sum_{j=1}^{p} P(|\beta_j|;\ a,\ b) \right],$$

Since $\boldsymbol{X}_{\mathrm{c}}$ is orthogonal, the above problem becomes

$$\min_{\beta_j \in \Theta_j} \left[d_n \beta_j^2 - 2u_j^{(k)}\beta_j + 2P(|\beta_j|;\ a,\ b) \right], \tag{8}$$

where $u_j^{(k)}$ is the jth component of $\mathbf{u}^{(k)} = \mathbf{u}(\boldsymbol{\beta}^{(k)})$ and

$$\mathbf{u}(\mathbf{z}) = (u_1(\mathbf{z}), \ldots, u_p(\mathbf{z}))' = \frac{\boldsymbol{X}'\mathbf{y}}{n} + \left(d_n \boldsymbol{I}_p - \frac{\boldsymbol{X}'\boldsymbol{X}}{n} \right) \mathbf{z}. \tag{9}$$

For $u \in \mathbb{R}$, define

$$s(u;\ a,\ b) = \begin{cases} \mathrm{sign}(u)\big(|u| - b\big)_+ / d_n, & \text{when } |u| \leqslant b(d_n + 1), \\ \mathrm{sign}(u)\Big\{(a-1)|u| - ab\Big\} / \\ \quad \Big\{(a-1)d_n - 1\Big\}, & \text{when } (d_n+1)b < |u| \leqslant abd_n, \\ u/d_n, & \text{when } |u| > abd_n, \end{cases}$$

and

$$\mathbf{s}(\mathbf{u};\ a,\ b) = \big[s(u_1;\ a,\ b), \ldots, s(u_p;\ a,\ b)\big]'. \tag{10}$$

We fix a in (3) and thus omit it in $\mathbf{s}$ hereinafter. Let $b = \lambda_n/n$ that depends on n. The problem in (8) has a closed-form solution that leads to the OEM iteration formula

$$\boldsymbol{\beta}^{(k+1)} = \mathbf{s}(\mathbf{u}^{(k)};\ \lambda_n/n). \tag{11}$$

For the regularised least squares problem with the MCP penalty in (5), the OEM iteration formula is $\boldsymbol{\beta}^{(k+1)} = \mathbf{s}_{\mathrm{MCP}}(\mathbf{u}^{(k)};\ \lambda_n/n)$ with $\mathbf{s}_{\mathrm{MCP}}(\mathbf{u};\ a,\ b) = \big[s_{\mathrm{MCP}}(u_1;\ a,\ b), \ldots, s_{\mathrm{MCP}}(u_p;\ a,\ b)\big]'$ and

$$s_{\mathrm{MCP}}(u;\ a,\ b) = \begin{cases} \mathrm{sign}(u)a\big(|u_j| - b\big)_+ / (ad_n - 1), & \text{when } |u| \leqslant abd_n, \\ u/d_n, & \text{when } |u| > abd_n. \end{cases} \tag{12}$$

A simple choice of d_n in (6) is $d_n = \gamma_1$, which can be computed efficiently by the Lanczos algorithm (Lanczos, 1950). With this choice, it suffices to compute γ_1 for obtaining $\boldsymbol{\beta}^{(k+1)}$ in (11) instead of computing the whole $\boldsymbol{\Delta}$.

The above OEM procedure can be easily used in other regularised least squares problems including ridge regression (Hoerl & Kennard, 1970), the nonnegative garrote (Breiman, 1995), and the lasso (Tibshirani, 1996). Xiong et al. (2011) showed that this is an EM algorithm, and derived its monotonicity and convergence properties for general regularised least squares problems.

3. Main results

Suppose that the number of nonzero coefficients of $\boldsymbol{\beta}$ in (1) is p_1 and partition $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')', \tag{13}$$

where $\boldsymbol{\beta}_2 = \mathbf{0}$ and no component of $\boldsymbol{\beta}_1$ is zero. Divide columns of the regression matrix $\boldsymbol{X}$ in (1) to $(\boldsymbol{X}_1\ \boldsymbol{X}_2)$ with $\boldsymbol{X}_1$ corresponding to $\boldsymbol{\beta}_1$. We assume $\mathrm{rank}(\boldsymbol{X}_1) = p_1$. Let $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_1^{*\prime}, \hat{\boldsymbol{\beta}}_2^{*\prime})'$ denote the oracle estimator with $\hat{\boldsymbol{\beta}}_1^* = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\mathbf{y}$ and $\hat{\boldsymbol{\beta}}_2^* = \mathbf{0}$. A regularised least squares estimator of $\boldsymbol{\beta}$ in (1) has the oracle property if it can not only select the correct submodel asymptotically, but also estimate the nonzero coefficients $\boldsymbol{\beta}_1$ in (13) as efficiently as if the correct submodel were known in advance. Specifically, an estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')'$ has this property if $\mathrm{P}(\hat{\boldsymbol{\beta}}_2 = \mathbf{0}) \to 1$ and $\hat{\boldsymbol{\beta}}_1$ has the same asymptotic distribution as $\hat{\boldsymbol{\beta}}_1^*$.

In virtue of the simplicity of the OEM iteration in (11), we can study the oracle property of the local solution of SCAD given by OEM. First, we prove that, under certain conditions, a fixed point of the OEM iterations for SCAD is the oracle estimator with probability tending to one. Some notation, definitions, and assumptions are needed. Hereinafter, p, p_1, and $\boldsymbol{\beta}_1$ can depend on n. Let $\beta_{\min}$ denote the component of $\boldsymbol{\beta}_1$ that has the smallest absolute value. The matrix $\boldsymbol{X}$ is standardised as in (4). Recall that $d_n \geq \gamma_1$ in (8).

Definition 3.1: For a series of numbers $c_n \to \infty$ and a positive constant κ, an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is said to be c_n-concentratively consistent of order κ if there exists a constant $C > 0$ such that for sufficiently large t, $\mathrm{P}(c_n\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \geqslant t) \leqslant C/t^{\kappa}$.

Remark 3.1: By the Markov inequality, $\hat{\boldsymbol{\beta}}$ is c_n-concentratively consistent of order κ if $\mathrm{E}\big[c_n\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|\big]^{\kappa} = O(1)$.

Assumption 3.1: The random errors $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with $\mathrm{E}\,\varepsilon_1 = 0$ and $\mathrm{E}|\varepsilon_1|^r < \infty$ for some $r \geqslant 2$.

Assumption 3.2: As $n \to \infty$, $p_1/(n^{1/2}d_n|\beta_{\min}|)^r \to 0$, $p_1/(c_n|\beta_{\min}|)^{\kappa} \to 0$, $\lambda_n/(n|\beta_{\min}|) \to 0$, $\lambda_n/(n^{1/2}p^{1/r}) \to \infty$, and $c_n\lambda_n/(nd_np^{1/\kappa}) \to \infty$.

Remark 3.2: Consider the case of $p \geqslant n$. We can assume $c_n = \sqrt{n/\log(p)}$ (Bühlmann & van de Geer 2011). Since $\boldsymbol{X}'\boldsymbol{X}$ has at most n non-zero eigenvalues, $\gamma_1 \geqslant \mathrm{trace}(\boldsymbol{X}'\boldsymbol{X}/n)/n = p/n$. The equality holds for some $\boldsymbol{X}$, and thus d_n can be assumed to have the same order of p/n. For $c_n = \sqrt{n/\log(p)}$, $d_n \sim p/n$, and sufficiently large r and κ, if we fix p_1 and $\boldsymbol{\beta}_1$ and set $p \sim n^q$ for some $q \geqslant 1$, then λ_n can be chosen as $\lambda_n \sim n^{\alpha}$, where $1 > \alpha > \max\{1/2 + q/r, q - 1/2 + q/\kappa\}$. It is clear that q should be smaller than 3/2. In other words, our results in this paper can handle dimensionality of order $p = O(n^q)$ for $q \in [1, 3/2)$.

Theorem 3.1: *Let $\hat{\boldsymbol{\beta}}^{\mathrm{f}}$ be a fixed point of the OEM iterations for SCAD. Suppose that $\hat{\boldsymbol{\beta}}^{\mathrm{f}}$ is a c_n-concentratively consistent estimator of order κ. Under Assumptions 3.1 and 3.2, as $n \to \infty$,*

$$\mathrm{P}(\hat{\boldsymbol{\beta}}^{\mathrm{f}} = \hat{\boldsymbol{\beta}}^{*}) \to 1.$$

Theorem 3.1 indicates that a fixed point of OEM consistent to the true parameter has the oracle property even when p grows faster than n.

Sometimes it is difficult to know whether a fixed point is consistent. We next show that, with an initial point concentratively consistent to $\boldsymbol{\beta}$, an OEM sequence can converge to that fixed point and possess the oracle property. In addition, its limit point is indeed the oracle estimator with probability tending to one.

Let $\{\boldsymbol{\beta}^{(k)},\ k = 0, 1, \ldots,\}$ be the OEM sequence from (11). Denote $\zeta_1 = \lambda_{\max}(\boldsymbol{X}_1'\boldsymbol{X}_1/n)$, $\zeta_{p_1} = \lambda_{\min}(\boldsymbol{X}_1'\boldsymbol{X}_1/n)$, $c_n^* = \min\{c_n, (n\zeta_{p_1}/p_1)^{1/2}\}$, and $\eta_n = \lambda_{\max}(\boldsymbol{I}_{p_1} - \boldsymbol{X}_1'\boldsymbol{X}_1/(nd_n)) = 1 - \zeta_{p_1}/d_n \in [0, 1)$.

Assumption 3.3: As $n \to \infty$, $p_1^{1+r/2}/((n\zeta_{p_1})^{r/2}|\beta_{\min}|^r) \to 0$, and $c_n^*\lambda_n/(nd_n) \to \infty$.

Theorem 3.2: *If $\boldsymbol{\beta}^{(0)}$ is c_n-concentratively consistent of order κ, then, under Assumptions 3.1–3.3, we have (i) as $n \to \infty$,*

$$\mathrm{P}\left(\lim_{k\to\infty} \boldsymbol{\beta}^{(k)} = \hat{\boldsymbol{\beta}}^{*}\right) \to 1; \tag{14}$$

(ii) for all $k = 1, 2, \ldots$, $\mathrm{P}(\boldsymbol{\beta}_2^{(k)} = \mathbf{0}) \to 1$ and $\|\boldsymbol{\beta}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^{*}\| = O_p(\eta_n^k/c_n^*)$.

By (ii) of Theorem 3.2, the OEM sequence can possess the oracle property only if k is sufficiently large. In particular, $\hat{\boldsymbol{\beta}}_1$ has the same asymptotic normality as $\hat{\boldsymbol{\beta}}_1^{*}$. We state this result as a corollary.

Assumption 3.4: As $n \to \infty$, $\zeta_{p_1}/d_n \to \delta \in [0, 1)$, and $k = k_n$ satisfies $(n\zeta_1)^{1/2}\exp(-k\log(1-\delta)^{-1})/c_n^* \to 0$ for $\delta > 0$ and $(n\zeta_1)^{1/2}\exp(-\zeta_{p_1}k/d_n)/c_n^* \to 0$ for $\delta = 0$.

Corollary 3.1: *Suppose that $\boldsymbol{\beta}^{(0)}$ is c_n-concentratively consistent of order κ. If*

$$\max_{1\leqslant i\leqslant n} \mathbf{x}_{1i}'(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\mathbf{x}_{1i} \to 0, \tag{15}$$

where $\mathbf{x}_{1i} = (x_{i1}, \ldots, x_{ip_1})'$, then under Assumptions 3.1–3.4, as $n \to \infty$,

(i) $\mathrm{P}(\boldsymbol{\beta}_2^{(k)} = \mathbf{0}) \to 1$;

(ii) for any non-zero $p_1 \times 1$ vector $\boldsymbol{\alpha}_n$, $\boldsymbol{\alpha}_n'(\boldsymbol{\beta}_1^{(k)} - \boldsymbol{\beta}_1)/[\boldsymbol{\alpha}_n'(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{\alpha}_n]^{1/2} \to N(0, \sigma^2)$ in distribution, where $\sigma^2 = \mathrm{E}\,\varepsilon_1^2$.

Remark 3.3: A sufficient condition for (15) is $p_1 = o(n\zeta_{p_1})$ and $\max_{1\leqslant i\leqslant n} p_1^{-1}\sum_{j=1}^{p_1} x_{ij}^2 = O(1)$.

Remark 3.4: The proofs of Theorems 3.1 and 3.2 only use the convergence rates of $\mathrm{P}(\beta_j^{(k+1)} = 0) = \mathrm{P}(|u_j^{(k)}| < \lambda_n/n)$ and $\mathrm{P}(\beta_j^{(k+1)} = u_j^{(k)}/d_n) = \mathrm{P}(|u_j^{(k)}| > ad_n\lambda_n/n)$. Since an OEM sequence for MCP has the same structure from (12), all results in this section hold for MCP with almost the same proofs.

From Corollary 3.1, the OEM sequence $\boldsymbol{\beta}^{(k)}$ can have the oracle property for sufficiently large k. For example, let p_1 and $\boldsymbol{\beta}_1$ be fixed and $\boldsymbol{X}_1'\boldsymbol{X}_1/n \to \boldsymbol{\Sigma}_1$, where $\boldsymbol{\Sigma}_1$ is a positive-definite matrix. For this case, if $d_n \to \infty$ and $\sqrt{n}\exp(-\lambda_{\min}(\boldsymbol{\Sigma}_1)k_n/d_n) \to 0$ as $n \to \infty$, then Assumption 3.4 holds and $\sqrt{n}(\boldsymbol{\beta}_1^{(k)} - \boldsymbol{\beta}_1) \to N(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}_1^{-1})$ in distribution.

Huo and Chen (2010) showed that, for the SCAD penalty, solving the global minimum of the SCAD problem leads to an NP-hard problem. Theorem 3.2 indicates that as far as the oracle property is concerned, the local solution given by OEM will suffice.

Theorems 3.1 and 3.2 can handle dimensionality of order $p = O(n^q)$ for $q < 3/2$. For p exceeding this order, penalised regression methods can perform poorly. A practical approach is a two-stage procedure in Fan and Lv (2008). The first stage uses an efficient screening method like the sure independence screening (SIS) (Fan & Lv, 2008) to reduce the dimensionality. OEM can be used in the second stage to obtain a SCAD estimator with the oracle property. In fact, OEM can also be used to screen variables even with a p increasing at an exponential rate, since the one-step OEM estimator with a proper λ_n using $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ is equivalent to the SIS method.

4. Discussion on the $n > p$ case

The OEM algorithm is originally motivated by applications with Big Data, i.e., large n (Xiong et al., 2016). In this section, we discuss the case of $p = o(n)$, and show that Theorems 3.1 and 3.2 hold under fairly weak conditions on $\boldsymbol{X}$ with the OLS estimator being the initial point of OEM. Like the previous sections, the matrix $\boldsymbol{X}$ in (1) is standardised as in (4). As in Section 3, the results in this section allow $\zeta_{p_1} \to 0$ and/or $\zeta_1 \to \infty$ as $n \to \infty$, that is, the regression matrix $\boldsymbol{X}$ and $\boldsymbol{X}_1$ can be nearly degenerate. Let $\gamma_p = \lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}/n)$. By Lemma A.5 in the Appendix, we immediately obtain the following theorem.

Theorem 4.1: *Suppose that $\gamma_p > 0$ and $p = o(n\gamma_p)$. Then, the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{y}$ is $\sqrt{n\gamma_p/p}$-concentratively consistent of order r under Assumption 3.1.*

For $p = o(n\gamma_p)$, we use the OLS estimator as the initial estimator in Theorem 3.2 and take $d_n = \gamma_1$. Since $p_1\gamma_1^2 \geqslant \zeta_{p_1}$ and $p\gamma_1^2 \geqslant \gamma_p$, under Assumption 3.1, Assumptions 3.2 and 3.3 become

Assumption 4.1: As $n \to \infty$, $p_1^{1+r/2}/\left((n\zeta_{p_1})^{r/2}|\beta_{\min}|^r\right) \to 0$, $\lambda_n/(n|\beta_{\min}|) \to 0$, and $\gamma_p^{1/2}\lambda_n/(n^{1/2}p^{1/2+1/r}\gamma_1) \to \infty$.

Assumption 3.4 becomes

Assumption 4.2: As $n \to \infty$, $\zeta_{p_1}/\gamma_1 \to \delta \in [0, 1)$, and $k = k_n$ satisfies $(p\zeta_1/\gamma_p)^{1/2}\exp(-k\log(1-\delta)^{-1}) \to 0$ for $\delta > 0$ and $(p\zeta_1/\gamma_p)^{1/2}\exp(-\zeta_{p_1}k/\gamma_1) \to 0$ for $\delta = 0$.

By Theorem 3.2 and Corollary 3.1, under Assumption 3.1, Assumptions 4.1 and 4.2 imply the oracle property of the OEM sequence for SCAD with the OLS estimator being the initial estimator.

Under above conditions and suppose further that there exist two positive constants $\underline{C}$ and $\overline{C}$ such that $\underline{C} \leqslant \gamma_p \leqslant \gamma_1 \leqslant \overline{C}$. Then, Assumption 4.1 reduces further to that, as $n \to \infty$, $p_1^{1+r/2}/\left(n^{r/2}|\beta_{\min}|^r\right) \to 0$, $\lambda_n/(n|\beta_{\min}|) \to 0$, and $\lambda_n/(n^{1/2}p^{1/2+1/r}) \to \infty$. Assumption 4.2 reduces further to that, as $n \to \infty$, $p^{1/2}\exp(-k\log(1-\underline{C}/\overline{C})^{-1}) \to 0$.

We next show an example where Assumption 4.1 holds with $\zeta_{p_1} \to 0$ and $\zeta_1 \to \infty$. Let $p \sim n^q$ with $q \in [0, 1)$, $\zeta_{p_1} \sim 1/\log(n)$, $\zeta_1 \sim \log(n)$, $\gamma_p \sim 1/\log(n)^2$, $\gamma_1 \sim \log(n)^2$, $\lambda_n \sim n^\alpha$, and $|\beta_{\min}|$ be a constant. Then, Assumption 4.1 holds if $p_1 = o((n/\log(n))^{r/(r+2)})$ and $1 > \alpha > 1/2 + q/2 + q/r$. Such α exists for any $q \in [0, 1)$ if r is sufficiently large.

Remark 4.1: In many papers on high-dimensional asymptotics, the random errors are assumed to follow sub-Gaussian distributions. Under the sub-Gaussian assumption, we can show that the OLS estimator has an exponential tail probability bound using the results in Hsu, Kakade, and Zhang (2012). Therefore, when using it as the initial point, Assumption 4.1 can be relaxed with more choices of λ_n.

Assumption 4.1 requires that the convergence rates of γ_1, $1/\gamma_p$, ζ_1, and $1/\zeta_{p_1}$ to infinity should be relatively slow, which holds for commonly encountered regression matrix $\boldsymbol{X}$. The following theorem derives the bounds of the eigenvalues under random designs from a broad class of correlated multivariate distributions.

Theorem 4.2: *Let z_{ij}, $i = 1, \ldots, n, j = 1, \ldots, p$, be i.i.d. random variables with $Ez_{11} = 0$, $Ez_{11}^2 = 1$, and $E|z_{11}|^4 < \infty$. Suppose that $\boldsymbol{U}$ is a $p \times p$ positive-definite matrix and $\boldsymbol{U}_1$ is the $p_1 \times p_1$ sub-matrix constructed by its first p_1 rows and columns. Denote $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})'$, $\mathbf{z}_{1i} = (z_{i1}, \ldots, z_{ip_1})'$, $\mathbf{x}_i = \boldsymbol{U}^{1/2}\mathbf{z}_i$, and $\mathbf{x}_{1i} = \boldsymbol{U}_1^{1/2}\mathbf{z}_{1i}$ for $i = 1, \ldots, n$. Let $\boldsymbol{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$. Then, for $p = o(n)$,*

$$\gamma_1 = O(\lambda_{\max}(\boldsymbol{U})),\ 1/\gamma_p = O(1/\lambda_{\min}(\boldsymbol{U})),\ \zeta_1 = O(\lambda_{\max}(\boldsymbol{U}_1)),\ \text{and } 1/\zeta_{p_1} = O(1/\lambda_{\min}(\boldsymbol{U}_1)) \tag{16}$$

with probability one.

Theorem 4.2 indicates that the convergence rates of γ_1, $1/\gamma_p$, ζ_1, and $1/\zeta_{p_1}$ to infinity will be relatively slow if the eigenvalues of $\boldsymbol{U}$ and $\boldsymbol{U}_1$ are restrictive. It is clear that a nearly degenerate $\boldsymbol{U}_1$ can also yield $\gamma_1, \gamma_p, \zeta_1$, and ζ_{p_1} satisfying Assumption 4.1 asymptotically.

5. Simulations

This section presents some simulation results of the SCAD solution given by OEM to support our theoretical discoveries. Our main purpose is to show that our method is at least comparable with other estimators having the oracle property.

We focus on the $p < n$ case, and compare the SCAD solution given by OEM (SCAD_{OEM}) with other four methods, including the lasso (Tibshirani, 1996), the adaptive lasso (AdaLasso) (Zou, 2006), Zou and Li (2008)'s one-step LLA estimator, and the SCAD solution given by the coordinate descent algorithm (SCAD_{CD}). The regression matrix $\boldsymbol{X}$ in (1) is constructed as $\boldsymbol{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently generated from $N(\mathbf{0}, \boldsymbol{\Sigma})$, and the (i, j) entry of $\boldsymbol{\Sigma}$ is $\rho^{|i-j|}$. The random errors $\varepsilon_1, \ldots, \varepsilon_n \sim N(0, 1)$, $p = 8$, and

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_8)' = (3, 1.5, 0, 0, 2, 0, 0, 0)'.$$

The sample size n is chosen as 40, 60, and 80. We first use the OEM algorithm to compute the SCAD solution with the initial point being the OLS estimator. The tuning parameter a in (3) is set as 3.7 as recommended in Fan and Li (2001). The other parameter $b = \lambda_n/n$ is selected by BIC (Wang, Li, & Tsai, 2007). With the same b, we compute the one-step estimator, and compare the variable selection errors (VSEs) and the model errors (MEs) of the two estimators. Here, the VSE and ME of an estimator $\hat{\boldsymbol{\beta}}$ are, respectively, defined as

$$\begin{aligned}\text{VSE}(\hat{\boldsymbol{\beta}}) = &|\{j\ :\ j \in \mathcal{A}(\boldsymbol{\beta}) \text{ but } j \notin \mathcal{A}(\hat{\boldsymbol{\beta}})\}| \\ &+|\{j\ :\ j \in \mathcal{A}(\hat{\boldsymbol{\beta}}) \text{ but } j \notin \mathcal{A}(\boldsymbol{\beta})\}|\end{aligned}$$

and

$$\text{ME}(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/n,$$

where $|\cdot|$ denotes cardinality and $\mathcal{A}(\boldsymbol{\beta}) = \{j :\ \beta_j \neq 0,\ j = 1, \ldots, p\}$.

The average VSE and ME values of the two estimators over 1000 times are shown in Table 2. We can see that SCAD_{OEM} outperforms the one-step estimator, especially when ρ is large. Besides, SCAD_{OEM} is comparable to AdaLasso that also possesses the oracle property (Zou, 2006) in terms of ME. SCAD_{CD} gives almost the same results as SCAD_{OEM}, but its oracle property has not been proved in the literature.

Table 2. Comparisons of VSEs and MEs.

	VSE			ME		
	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
$n = 40$						
Lasso	2.560 (1.57)	2.292 (1.49)	2.144 (1.31)	0.166 (0.10)	0.158 (0.10)	0.139 (0.09)
AdaLasso	1.028 (1.31)	0.941 (1.21)	1.192 (1.29)	0.123 (0.09)	0.124 (0.09)	0.142 (0.10)
SCAD (one-step)	1.892 (1.92)	1.440 (1.74)	3.529 (1.19)	0.183 (0.11)	0.181 (0.14)	0.227 (0.28)
SCAD_{CD}	1.591 (1.63)	1.101 (1.35)	1.519 (0.80)	0.134 (0.08)	0.129 (0.08)	0.137 (0.13)
SCAD_{OEM}	1.591 (1.63)	1.101 (1.35)	1.513 (0.79)	0.134 (0.08)	0.129 (0.08)	0.138 (0.13)
$n = 60$						
Lasso	2.622 (1.52)	2.322 (1.51)	2.159 (1.34)	0.110 (0.07)	0.105 (0.07)	0.092 (0.06)
AdaLasso	1.000 (1.26)	0.960 (1.28)	1.033 (1.26)	0.080 (0.06)	0.080 (0.06)	0.087 (0.07)
SCAD (one-step)	1.842 (1.93)	1.448 (1.76)	3.585 (1.15)	0.135 (0.09)	0.119 (0.09)	0.152 (0.25)
SCAD_{CD}	1.589 (1.68)	1.114 (1.37)	1.415 (0.67)	0.089 (0.06)	0.083 (0.06)	0.078 (0.08)
SCAD_{OEM}	1.589 (1.68)	1.114 (1.37)	1.416 (0.67)	0.089 (0.06)	0.083 (0.06)	0.079 (0.08)
$n = 80$						
Lasso	2.510 (1.51)	2.303 (1.47)	2.220 (1.26)	0.082 (0.04)	0.079 (0.05)	0.070 (0.05)
AdaLasso	0.931 (1.23)	0.902 (1.19)	0.967 (1.20)	0.058 (0.04)	0.059 (0.04)	0.062 (0.05)
SCAD (one-step)	1.683 (1.93)	1.360 (1.73)	3.522 (1.15)	0.106 (0.08)	0.099 (0.07)	0.097 (0.05)
SCAD_{CD}	1.420 (1.67)	1.099 (1.42)	1.362 (0.59)	0.066 (0.04)	0.064 (0.04)	0.053 (0.05)
SCAD_{OEM}	1.420 (1.67)	1.099 (1.42)	1.364 (0.59)	0.066 (0.04)	0.064 (0.04)	0.053 (0.05)

6. Concluding remarks

Since Fan and Li (2001) pointed out that there exists a local solution of SCAD having the oracle property, it has become an interesting open problem to find such a local solution. We have proved that the OEM algorithm can indeed provide this local solution with the oracle property even with a diverging p. Compared with other estimators after one or several iterations with this property, our results provide a new way to compute the required local solution in Fan and Li (2001) and Zhang (2010) and indicates a new interface between optimisation and statistics for non-convex penalties.

Although our main results require $p = O(n^q)$ with $q \in [0, 3/2)$, the condition on the random error $\boldsymbol{\varepsilon}$ is accordingly weak. Especially, for $p = o(n)$ and under such a condition on $\boldsymbol{\varepsilon}$, our results easily hold for commonly encountered regression matrix $\boldsymbol{X}$, even nearly degenerate, with the OLS estimator being the initial point. This point also matches the spirit why Fan and Lv (2008) proposed the two-stage method for an ultra-high p. That is, regularised least squares methods like SCAD are more suitable for a moderate p.

This paper does not discuss the selection of the tuning parameter λ_n in the SCAD penalty. This issue has been intensively studied in the literature, and we refer the reader to Wang et al. (2007), Wang, Li, and Leng (2009), Wang et al. (2013), and Fan and Tang (2013), among others.

Acknowledgments

We thank the editor, the associate editor, and two referees for their helpful and constructive comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Xiong's research was supported by the National Natural Science Foundation of China [grant number 11471172], [grant number 11671386].

References

Bai, Z. D., & Yin, Y. Q. (1993). Limit of smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability, 21*, 1275–1294.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics, 37*, 373–384.

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics, 5*, 232–253.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Berlin: Springer.

Eicker, F. R. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics, 34*, 447–456.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*, 1348–1360.

Fan, J., & Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society: Series B, 70*, 849–911.

Fan, J., & Lv, J. (2011). Properties of non-concave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory, 57*, 5467–5484.

Fan, J., & Peng, H. (2004). Non-concave penalized likelihood with diverging number of parameters. *Annals of Statistics, 32*, 928–961.

Fan, J., Xue, L., & Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics, 42*, 819–849.

Fan, Y., & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society, Series B, 75*, 531–552.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67.

Hsu, D., Kakade, S. M., & Zhang, T. (2012). A tail inequality for quadratic forms of sub-Gaussian random vectors. *Electronic Journal of Probability, 17*, 1–6.
Hunter, D. R., & Li, R. (2005). Variable selection using MM algorithms. *Annals of Statistics, 33*, 1617–1642.
Huo, X., & Chen, J. (2010). Complexity of penalized likelihood estimation. *Journal of Statistical Computation and Simulation, 80*, 747–759.
Huo, X., & Ni, X. L. (2007). When do stepwise algorithms meet subset selection criteria? *Annals of Statistics, 35*, 870–887.
Kim, Y., Choi, H., & Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of American Statistical Association, 103*, 1665–1673.
Lai, T. L., & Wei, C. Z. (1984). Moment inequalities with applications to regression and time series models. *Inequalities in Statistics and Probability, IMS Lecture Notes-Monograph Series, 5*, 165–172.
Lanczos, C. (1950). An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards, 45*, 255–282.
Loh, P. L., & Wainwright, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Advances in Neural Information Processing Systems, 26*, 476–484.
Mazumder, R., Friedman, J., & Hastie, T. (2011). SparseNet: Coordinate descent with non-convex penalties. *Journal of American Statistical Association, 106*, 1125–1138.
Meng, X. L. (2008). Discussion on one-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics, 36*, 1542–1552.
Schifano, E. D., Strawderman, R., & Wells, M. T. (2010). Majorization–minimization algorithms for nonsmoothly penalized objective functions. *Electronic Journal of Statistics, 4*, 1258–1299.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, 58*, 267–288.
Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications, 109*, 475–494.
Tseng, P., & Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Programs in Mathematics, 117*, 387–423.
Wang, H., Li, B., & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B, 71*, 671–683.
Wang, H., Li, R., & Tsai, C-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika, 94*, 553–568.
Wang, L., Kim, Y., & Li, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Annals of Statistics, 41*, 2505–2536.
Wang, S., & Jia, Z. (1994). *Inequalities in matrix theory* (In Chinese). Hefei: Anhui Education Press.
Xiong, S., Dai, B., Huling, J., & Qian, P. (2016). Orthogonalizing EM: A design-based least squares algorithm. *Technometrics, 58*, 285–293.
Xiong, S., Dai, B., & Qian, P. (2011). *OEM algorithm for least squares problems* (Unpublished report). Retrieved from http://arxiv.org/abs/1108.0185
Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics, 38*, 894–942.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*, 1418–1429.
Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics, 36*, 1509–1533.

Appendix: Proofs

Lemma A.1: *Under Assumption 3.1, for* $\mathbf{a} \in \mathbb{R}^n$ *with* $\|\mathbf{a}\| = 1$ *and* $t \in \mathbb{R}$, $\mathrm{P}(|\mathbf{a}'\boldsymbol{\varepsilon}| \geqslant t) \leqslant C/t^r$, *where* $C > 0$ *is a constant that does not rely on* n *or* $\mathbf{a}$.

Proof: By the Markov inequality, $\mathrm{P}(|\mathbf{a}'\boldsymbol{\varepsilon}| \geqslant t) \leqslant \mathrm{E}|\mathbf{a}'\boldsymbol{\varepsilon}|^r/t^r$. By the Lai-Wei inequality in Example 3 of Lai and Wei (1984) and Assumption 3.1, $\mathrm{E}|\mathbf{a}'\boldsymbol{\varepsilon}|^r$ is not greater than a positive constant that does not rely on n or $\mathbf{a}$. This completes the proof. ■

Proof of Theorem 3.1: Under Assumption 3.1, $\sigma^2 = \mathrm{E}\,\varepsilon_1^2 < \infty$. Without loss of generality, assume $\sigma^2 = 1$ in all the proofs. Let $\mathbf{x}_1, \ldots, \mathbf{x}_p$ denote the columns of $\boldsymbol{X}$.

Since $\hat{\boldsymbol{\beta}}^{\mathrm{f}}$ is a fixed point, $\hat{\boldsymbol{\beta}}^{\mathrm{f}} = \mathbf{s}(\hat{\mathbf{u}}; \lambda_n/n)$ with $\hat{\mathbf{u}} = \mathbf{u}(\hat{\boldsymbol{\beta}}^{\mathrm{f}})$, where $\mathbf{u}$ and $\mathbf{s}$ are defined in (9) and (10), respectively. Therefore,

$$\frac{\hat{\mathbf{u}}}{d_n} = \boldsymbol{\beta} + \frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{nd_n} + \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}'\boldsymbol{X}}{nd_n}\right)(\hat{\boldsymbol{\beta}}^{\mathrm{f}} - \boldsymbol{\beta}). \quad \text{(A1)}$$

We have

$$\begin{aligned}
&\mathrm{P}\big(\hat{\boldsymbol{\beta}}_1^{\mathrm{f}} = \hat{\mathbf{u}}_1/d_n,\ \hat{\boldsymbol{\beta}}_2^{\mathrm{f}} = \mathbf{0}\big) \\
&\quad = \mathrm{P}(|\hat{u}_j| > ad_n\lambda_n/n \text{ for } j = 1, \ldots, p_1,\ |\hat{u}_j| \\
&\qquad < \lambda_n/n \text{ for } j = p_1 + 1, \ldots, p) \\
&\quad \geqslant 1 - \sum_{j=1}^{p_1} \mathrm{P}(|\hat{u}_j| \leqslant ad_n\lambda_n/n) \\
&\qquad - \sum_{j=p_1+1}^{p} \mathrm{P}(|\hat{u}_j| \geqslant \lambda_n/n). \quad \text{(A2)}
\end{aligned}$$

Note that $d_n \geqslant \gamma_1 \geqslant p/\mathrm{rank}(\boldsymbol{X}) \geqslant 1$ and that $\hat{\boldsymbol{\beta}}^{\mathrm{f}}$ is concentratively consistent. For sufficiently large n, by (A1) and Assumption 3.2,

$$\begin{aligned}
&\sum_{j=1}^{p_1} \mathrm{P}(|\hat{u}_j| \leqslant ad_n\lambda_n/n) \\
&\quad \leqslant \sum_{j=1}^{p_1} \mathrm{P}(|\beta_{\min}| - |\mathbf{x}_j'\boldsymbol{\varepsilon}/(nd_n)| \\
&\qquad - \left\| \left(\boldsymbol{I}_p - (\boldsymbol{X}'\boldsymbol{X})/(nd_n)\right)(\hat{\boldsymbol{\beta}}^{\mathrm{f}} - \boldsymbol{\beta}) \right\| \leqslant a\lambda_n/n) \\
&\quad \leqslant \sum_{j=1}^{p_1} \left\{ \mathrm{P}(|\mathbf{x}_j'\boldsymbol{\varepsilon}/(nd_n)| + \|\hat{\boldsymbol{\beta}}^{\mathrm{f}} - \boldsymbol{\beta}\| \geqslant |\beta_{\min}|/2) \right\} \\
&\quad \leqslant \sum_{j=1}^{p_1} \mathrm{P}\left[|\mathbf{x}_j'\boldsymbol{\varepsilon}/\sqrt{n}| \geqslant \sqrt{n}d_n|\beta_{\min}|/4 \right]
\end{aligned}$$

$$+p_1\,\mathrm{P}\big(c_n\|\hat{\boldsymbol{\beta}}^{\mathrm{f}}-\boldsymbol{\beta}\|\geqslant c_n|\beta_{\min}|/4\big)$$
$$=O\left(p_1/(n^{1/2}d_n|\beta_{\min}|)^r\right)+O\left(p_1/(c_n|\beta_{\min}|)^{\kappa}\right). \tag{A3}$$

For the other part in (A2),

$$\begin{aligned}
&\sum_{j=p_1+1}^{p}\mathrm{P}(|\hat{u}_j|\geqslant\lambda_n/n)\\
&\leqslant\sum_{j=p_1+1}^{p}\mathrm{P}\big(|\mathbf{x}_j'\boldsymbol{\varepsilon}/\sqrt{n}|\geqslant\lambda_n/\sqrt{n}-\sqrt{n}d_n\|\\
&\quad\times\left(\boldsymbol{I}_p-(\boldsymbol{X}'\boldsymbol{X})/(nd_n)\right)(\hat{\boldsymbol{\beta}}^{\mathrm{f}}-\boldsymbol{\beta})\|\big)\\
&\leqslant\sum_{j=p_1+1}^{p}\mathrm{P}\big[|\mathbf{x}_j'\boldsymbol{\varepsilon}/\sqrt{n}|\geqslant\lambda_n/(2\sqrt{n})\big]\\
&\quad+p\,\mathrm{P}\big(c_n\|\hat{\boldsymbol{\beta}}^{\mathrm{f}}-\boldsymbol{\beta}\|\geqslant\lambda_nc_n/(2nd_n)\big)\\
&=O\left(p(\lambda_n/\sqrt{n})^{-r}\right)+O\left[p\left((c_n\lambda_n)/(nd_n)\right)^{-\kappa}\right].
\end{aligned} \tag{A4}$$

Plugging the above inequalities in (A2), we have

$$\mathrm{P}(\hat{\boldsymbol{\beta}}_1^{\mathrm{f}}=\hat{\mathbf{u}}_1/d_n,\ \hat{\boldsymbol{\beta}}_2^{\mathrm{f}}=\mathbf{0})\to 1.$$

Note that when $\hat{\boldsymbol{\beta}}_1^{\mathrm{f}}=\hat{\mathbf{u}}_1/d_n$ and $\hat{\boldsymbol{\beta}}_2^{\mathrm{f}}=\mathbf{0}$,

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{f}}=\frac{\hat{\mathbf{u}}_1}{d_n}=\boldsymbol{\beta}_1+\frac{\boldsymbol{X}_1'\boldsymbol{\varepsilon}}{nd_n}+\left(\boldsymbol{I}_{p_1}-\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{nd_n}\right)(\hat{\boldsymbol{\beta}}_1^{\mathrm{f}}-\boldsymbol{\beta}_1),$$

which implies that

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{f}}=\boldsymbol{\beta}_1+(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\varepsilon}=(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\mathbf{y}.$$

This completes the proof. ■

To prove Theorem 3.2, we need several lemmas. For $\phi>0$, define $E(\phi)=\{\mathbf{z}=(\mathbf{z}_1',\ \mathbf{z}_2')'\in\mathbb{R}^p:\ \|\mathbf{z}_1-\hat{\boldsymbol{\beta}}_1^*\|<\phi,\ \mathbf{z}_2=\mathbf{0}\}$ and $F=\{\mathbf{z}\in\mathbb{R}^p:\ |u_j(\mathbf{z})|>ad_n\lambda_n/n$ for $j=1,\ldots,p_1,\ |u_j(\mathbf{z})|<\lambda_n/n$ for $j=p_1+1,\ldots,p\}$, where the vector-valued function $\mathbf{u}=(u_1,\ldots,u_p)'$ is defined in (9).

Lemma A.2: *If $\phi<\lambda_n/(nd_n)$ and $\phi<|\hat{\beta}_j^*|-a\lambda_n/n$ for $j=1,\ldots,p_1$, then $E(\phi)\subset F$.*

Proof: For $\mathbf{z}\in\mathbb{R}^p$, we have

$$\begin{aligned}
\mathbf{u}(\mathbf{z})&=\frac{\boldsymbol{X}'\mathbf{y}}{n}+\left(d_n\boldsymbol{I}_p-\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)(\hat{\boldsymbol{\beta}}^*+\mathbf{z}-\hat{\boldsymbol{\beta}}^*)\\
&=d_n\hat{\boldsymbol{\beta}}^*+\left(d_n\boldsymbol{I}_p-\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)(\mathbf{z}-\hat{\boldsymbol{\beta}}^*).
\end{aligned}$$

Then, for $\mathbf{z}=(\mathbf{z}_1',\ \mathbf{0}')'\in E(\phi)$ and $j=1,\ldots,p_1$, $|u_j(\mathbf{z})|\geqslant d_n|\hat{\beta}_j^*|-\|(d_n\boldsymbol{I}_p-\boldsymbol{X}'\boldsymbol{X}/n)(\mathbf{z}-\hat{\boldsymbol{\beta}}^*)\|>d_n(\phi+a\lambda_n/n)-d_n\phi=ad_n\lambda_n/n$; for $j=p_1+1,\ldots,p$, since $(u_{p_1+1}(\mathbf{z}),\ldots,u_p(\mathbf{z}))'=-\boldsymbol{X}_2'\boldsymbol{X}_1(\mathbf{z}_1-\hat{\boldsymbol{\beta}}_1^*)/n$, we have $|u_j(\mathbf{z})|\leqslant\left\{(\mathbf{z}_1-\hat{\boldsymbol{\beta}}_1^*)'\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{X}_2'\boldsymbol{X}_1(\mathbf{z}_1-\hat{\boldsymbol{\beta}}_1^*)/n^2\right\}^{1/2}\leqslant d_n\phi<\lambda_n/n$. ■

Lemma A.3: *Let $\{\boldsymbol{\beta}^{(k)},\ k=0,1,\ldots,\}$ be the OEM sequence from (11). If $\boldsymbol{\beta}^{(0)}\in E(\phi)$, then under the conditions of Lemma A.2, $\boldsymbol{\beta}_1^{(k)}\to\hat{\boldsymbol{\beta}}_1^*$ as $k\to\infty$, and $\boldsymbol{\beta}_2^{(k)}=\mathbf{0}$ for all $k=0,1,\ldots$.*

Proof: Since $\boldsymbol{\beta}^{(0)}\in E(\phi)$, by Lemma A.2, $\boldsymbol{\beta}^{(0)}\subset F$, which implies $\boldsymbol{\beta}_1^{(1)}=\mathbf{u}_1(\boldsymbol{\beta}^{(0)})/d_n=\hat{\boldsymbol{\beta}}_1^*+(\boldsymbol{I}_{p_1}-\boldsymbol{X}_1'\boldsymbol{X}_1/(nd_n))(\boldsymbol{\beta}_1^{(0)}-\hat{\boldsymbol{\beta}}_1^*)$ and $\boldsymbol{\beta}_2^{(1)}=\mathbf{0}$. Recall that $\eta_n=\lambda_{\max}(\boldsymbol{I}_{p_1}-\boldsymbol{X}_1'\boldsymbol{X}_1/(nd_n))\in(0,1)$. We have $\|\boldsymbol{\beta}_1^{(1)}-\hat{\boldsymbol{\beta}}_1^*\|\leqslant\eta_n\|\boldsymbol{\beta}_1^{(0)}-\hat{\boldsymbol{\beta}}_1^*\|<\|\boldsymbol{\beta}_1^{(0)}-\hat{\boldsymbol{\beta}}_1^*\|\leqslant\phi$. Consequently, $\boldsymbol{\beta}^{(1)}\in E(\phi)$. By recursive, we know that, for all $k=1,2,\ldots,$ $\boldsymbol{\beta}^{(k)}\in E(\phi)$, and $\|\boldsymbol{\beta}_1^{(k)}-\hat{\boldsymbol{\beta}}_1^*\|\leqslant\eta_n^k\|\boldsymbol{\beta}_1^{(0)}-\hat{\boldsymbol{\beta}}_1^*\|$. Letting $k\to\infty$, we complete the proof. ■

Lemma A.4: *Under Assumption 3.1, $\|\boldsymbol{X}_1'\boldsymbol{\varepsilon}\|/(np_1)^{1/2}=O_p(\zeta_1^{1/2})$.*

Proof: Let $\zeta_1\geqslant\cdots\geqslant\zeta_{p_1}$ be all eigenvalues of $\boldsymbol{X}_1'\boldsymbol{X}_1/n$. Therefore, $\boldsymbol{X}_1\boldsymbol{X}_1'/n$ can be written as $\boldsymbol{X}_1\boldsymbol{X}_1'/n=\sum_{j=1}^{p_1}\zeta_j\mathbf{a}_j\mathbf{a}_j'$ with $\|\mathbf{a}_j\|=1$ and $\mathbf{a}_i'\mathbf{a}_j=0$ for $i\neq j$. We have $\mathrm{E}\big(\|\boldsymbol{X}_1'\boldsymbol{\varepsilon}\|^2/(np_1)\big)=p_1^{-1}\sum_{j=1}^{p_1}\zeta_j\big(\mathrm{E}\|\mathbf{a}_j'\boldsymbol{\varepsilon}\|^2\big)\leqslant\zeta_1p_1^{-1}\sum_{j=1}^{p_1}\big(\mathrm{E}\|\mathbf{a}_j'\boldsymbol{\varepsilon}\|^2\big)$. The lemma follows from the Markov inequality and the Lai-Wei inequality. ■

Lemma A.5: *If $p_1=o(n\zeta_{p_1})$, then under Assumption 3.1, $\hat{\boldsymbol{\beta}}^*$ is $\sqrt{n\zeta_{p_1}/p_1}$-concentratively consistent of order r.*

Proof: Let $\zeta_1,\ldots,\zeta_{p_1}$ be the same as in the proof of Lemma A.4. Since all non-zero eigenvalues of matrix $\boldsymbol{A}=(np_1)^{-1}\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1/n)^{-2}\boldsymbol{X}_1'$ are $1/(\zeta_1p_1),\ldots,1/(\gamma_{p_1}p_1)$, we have $\boldsymbol{A}=\sum_{j=1}^{p_1}\mathbf{b}_j'\mathbf{b}_j/(\zeta_jp_1)$, where $\mathbf{b}_j\in\mathbb{R}^n$ with $\|\mathbf{b}_j\|=1$ and $\mathbf{b}_i'\mathbf{b}_j=0$ for $i\neq j$. By the C_r inequality, we have $\mathrm{E}(\sqrt{n\zeta_{p_1}/p_1}\|\hat{\boldsymbol{\beta}}^*-\boldsymbol{\beta}\|)^r=\mathrm{E}\,(\zeta_{p_1}\boldsymbol{\varepsilon}'\boldsymbol{A}\boldsymbol{\varepsilon})^{r/2}=\mathrm{E}(\zeta_{p_1}p_1^{-1}\sum_{j=1}^{p_1}\|\mathbf{b}_j'\boldsymbol{\varepsilon}\|^2/\zeta_j)^{r/2}\leqslant\mathrm{E}(p_1^{-1}\sum_{j=1}^{p_1}\|\mathbf{b}_j'\boldsymbol{\varepsilon}\|^2)^{r/2}\leqslant p_1^{-1}\sum_{j=1}^{p_1}\mathrm{E}|\mathbf{b}_j'\boldsymbol{\varepsilon}|^r$. By the Lai-Wei inequality, the above expression is not greater than a constant. By Remark 3.1, this completes the proof. ■

Lemma A.6: *Let $\{\boldsymbol{\beta}^{(k)},\ k=0,1,\ldots,\}$ be the OEM sequence from (11). If $\boldsymbol{\beta}^{(0)}$ is c_n-concentratively consistent of order κ, then under Assumptions 3.1 and 3.2, $\mathrm{P}(\boldsymbol{\beta}_1^{(1)}=\mathbf{u}_1^{(0)}/d_n,\ \boldsymbol{\beta}_2^{(1)}=\mathbf{0})\to 1$ as $n\to\infty$.*

Proof: By (9),

$$\begin{aligned}
\mathbf{u}^{(0)}/d_n&=\boldsymbol{\beta}+\boldsymbol{X}'\boldsymbol{\varepsilon}/(nd_n)\\
&\quad+\left(\boldsymbol{I}_p-\boldsymbol{X}'\boldsymbol{X}/(nd_n)\right)(\boldsymbol{\beta}^{(0)}-\boldsymbol{\beta}).
\end{aligned} \tag{A5}$$

Similar to (A3) and (A4),

$$\begin{aligned}
&\mathrm{P}\left(\boldsymbol{\beta}_1^{(1)}=\mathbf{u}_1^{(0)}/d_n\right)\\
&\quad=\mathrm{P}\big(|u_j^{(0)}|>ad_n\lambda_n/n\text{ for }j=1,\ldots,p_1\big)\\
&\quad\geqslant 1-\sum_{j=1}^{p_1}\mathrm{P}\big(|u_j^{(0)}|\leqslant ad_n\lambda_n/n\big)
\end{aligned}$$

$$= 1 - O\left(p_1/(n^{1/2}d_n|\beta_{\min}|)^r\right) + O\left(p_1/(c_n|\beta_{\min}|)^\kappa\right) \quad \text{(A6)}$$

and

$$\begin{aligned} \mathrm{P}\left(\boldsymbol{\beta}_2^{(1)} = \mathbf{0}\right) &= \mathrm{P}\left(|u_j^{(0)}| \leqslant \lambda_n/n \text{ for } j = p_1+1, \ldots, p\right) \\ &\geqslant 1 - \sum_{j=p_1+1}^{p} \mathrm{P}\left(|u_j^{(0)}| > \lambda_n/n\right) \\ &= 1 - O\left(p(\lambda_n/\sqrt{n})^{-r}\right) + O\left[p\left((c_n\lambda_n)/(nd_n)\right)^{-\kappa}\right]. \end{aligned} \quad \text{(A7)}$$

By Assumption 3.2, we complete the proof. ■

Lemma A.7: *Under Assumptions 3.1–3.3,* $\mathrm{P}(\boldsymbol{\beta}^{(1)} \in E(\phi_n)) \to 1$ *as* $n \to \infty$, *where* $\phi_n = (\lambda_n/(nd_nc_n^*))^{1/2}$.

Proof: By Lemma A.6, $\mathrm{P}(\boldsymbol{\beta}_2^{(1)} = \mathbf{0}) \to 1$. It suffices to consider $\boldsymbol{\beta}_1^{(1)}$. If $\boldsymbol{\beta}_1^{(1)} = \mathbf{u}_1^{(0)}/d_n$, then by (A5) and Lemma A.4, $\|\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_1\| \leqslant \|\boldsymbol{\beta}_1^{(0)} - \boldsymbol{\beta}_1\| + \|\boldsymbol{X}_1\boldsymbol{\varepsilon}/(nd_n)\| = O_p(1/c_n^{**})$, where $c_n^{**} = \min\{c_n, n^{1/2}d_n/(p_1\zeta_1)^{1/2}\}$. Since $\|\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1\| = O_p((n\zeta_{p_1}/p_1)^{-1/2})$ by Lemma A.5, $\|\boldsymbol{\beta}_1^{(1)} - \hat{\boldsymbol{\beta}}_1^*\| \leqslant \|\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_1\| + \|\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1\| = O_p(1/c_n^{**}) + O_p((n\zeta_{p_1}/p_1)^{-1/2}) = O_p(1/c_n^*)$. By Assumption 3.3, $\phi_n c_n^* \to 1$ as $n \to \infty$. This implies $\mathrm{P}(\|\boldsymbol{\beta}_1^{(1)} - \hat{\boldsymbol{\beta}}_1^*\| < \phi_n) \to 1$ and completes the proof. ■

Proof of Theorem 3.2: (i) By Assumption 3.3, $\phi_n < \lambda_n/(nd_n)$ for sufficiently large n, where ϕ_n is defined in Lemma A.7. By Lemma A.3, $\mathrm{P}(\lim_{k\to\infty} \boldsymbol{\beta}^{(k)} = \hat{\boldsymbol{\beta}}^*) \geqslant \mathrm{P}(\phi_{n_0} < \min_{j=1,\ldots,p_1}\{|\hat{\beta}_j^*| - a\lambda_n/n_0\}, \boldsymbol{\beta}^{(1)} \in E(\phi_{n_0})$ for some $n_0 \in \mathbb{N}$). By Lemma A.7, it suffices to show

$$\mathrm{P}\left(\lambda_n/(nd_n) < \min_{j=1,\ldots,p_1}\{|\hat{\beta}_j^*| - a\lambda_n/n\}\right) \to 1. \quad \text{(A8)}$$

Since $\lambda_n/(n|\beta_{\min}|) \to 0$ in Assumption 3.2, for sufficiently large n, we have

$$\begin{aligned} &\mathrm{P}\left(\lambda_n/(nd_n) < \min_{j=1,\ldots,p_1}\{|\hat{\beta}_j^*| - a\lambda_n/n\}\right) \\ &\geqslant \mathrm{P}\left(\min_{j=1,\ldots,p_1}|\hat{\beta}_j^*| > a\lambda_n/n + \lambda_n/n\right) \\ &\geqslant \mathrm{P}\left(\min_{j=1,\ldots,p_1}|\hat{\beta}_j^*| > |\beta_{\min}|/2\right) \\ &\geqslant 1 - \sum_{j=1}^{p_1}\mathrm{P}\left(|\beta_j| - |\hat{\beta}_j^* - \beta_j| \leqslant |\beta_{\min}|/2\right) \\ &\geqslant 1 - p_1\mathrm{P}\left(\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\| \geqslant |\beta_{\min}|/2\right) \\ &\geqslant 1 - Cp_1\left(\sqrt{n\zeta_{p_1}/p_1}|\beta_{\min}|\right)^{-r}, \end{aligned}$$

where $C > 0$ is a constant. The last inequality is from Lemma A.5. By Assumption 3.3, (A8) holds and this completes the proof of (i).

(ii) By Lemma A.3 and its proof, when $\phi_{n_0} < \lambda_{n_0}/(n_0d_{n_0})$, $\phi_{n_0} < \min_{j=1,\ldots,p_1}\{|\hat{\beta}_j^*| - a\lambda_n/n_0\}$, and $\boldsymbol{\beta}^{(1)} \in E(\phi_{n_0})$ for some $n_0 \in \mathbb{N}$, $\boldsymbol{\beta}_2^{(k)} = \mathbf{0}$ and $\|\boldsymbol{\beta}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^*\| \leqslant \eta_n^k\|\boldsymbol{\beta}_1^{(0)} - \hat{\boldsymbol{\beta}}_1^*\|$ for all $k = 0, 1, \ldots$. The proof of (ii) is completed by noting $\|\boldsymbol{\beta}_1^{(0)} - \hat{\boldsymbol{\beta}}_1^*\| \leqslant \|\boldsymbol{\beta}_1^{(0)} - \boldsymbol{\beta}_1\| + \|\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1\| = O_p(1/c_n^*)$. ■

Proof of Corollary 3.1: We only need to prove (ii). With minor modifications from Eicker (1963), under Assumption 3.1 and (15), we have $\boldsymbol{\alpha}_n'(\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1)/[\boldsymbol{\alpha}_n'(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{\alpha}_n]^{1/2} \to N(0, \sigma^2)$ in distribution. Then, it suffices to show $|\boldsymbol{\alpha}_n'(\boldsymbol{\beta}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^*)|/[\boldsymbol{\alpha}_n'(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{\alpha}_n]^{1/2} = o_p(1)$. By (ii) of Theorem 3.2 and Assumption 3.4, the left side $\leqslant \|\boldsymbol{\beta}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^*\|/(n\zeta_1)^{-1/2} = O_p(\sqrt{n\zeta_1}\eta_n^k/c_n^*) = o_p(1)$. ■

Proof of Theorem 4.2: Note that $\boldsymbol{X}'\boldsymbol{X}/n = \boldsymbol{Z}'\boldsymbol{U}\boldsymbol{Z}/n$, where $\boldsymbol{Z} = (z_{ij})$. By the results in Bai and Yin (1993), $\liminf_{n\to\infty}\lambda_{\min}(\boldsymbol{Z}'\boldsymbol{Z}/n) \geqslant 1$ and $\limsup_{n\to\infty}\lambda_{\max}(\boldsymbol{Z}'\boldsymbol{Z}/n) \leqslant 1$ with probability one. Then, the right part of (16) can be proved by noting that $\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}/n) = \lambda_{\min}(\boldsymbol{Z}'\boldsymbol{U}\boldsymbol{Z}/n) = \lambda_{\min}(\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{U}/n) \geqslant \lambda_{\min}(\boldsymbol{Z}'\boldsymbol{Z}/n)\lambda_{\min}(\boldsymbol{U})$, where the last inequality is from Corollary 4.6.3 in Wang and Jia (1994). The proof of the left part is similar. ■