The Annals of Statistics
2016, Vol. 44, No. 6, 2661–2694
DOI: 10.1214/15-AOS1431

HIGH-DIMENSIONAL GENERALIZATIONS OF ASYMMETRIC LEAST SQUARES REGRESSION AND THEIR APPLICATIONS

BY YUWEN GU AND HUI ZOU[1]

University of Minnesota

Asymmetric least squares regression is an important method that has wide applications in statistics, econometrics and finance. The existing work on asymmetric least squares only considers the traditional low dimension and large sample setting. In this paper, we systematically study the Sparse Asymmetric LEast Squares (SALES) regression under high dimensions where the penalty functions include the Lasso and nonconvex penalties. We develop a unified efficient algorithm for fitting SALES and establish its theoretical properties. As an important application, SALES is used to detect heteroscedasticity in high-dimensional data. Another method for detecting heteroscedasticity is the sparse quantile regression. However, both SALES and the sparse quantile regression may fail to tell which variables are important for the conditional mean and which variables are important for the conditional scale/variance, especially when there are variables that are important for both the mean and the scale. To that end, we further propose a COupled Sparse Asymmetric LEast Squares (COSALES) regression which can be efficiently solved by an algorithm similar to that for solving SALES. We establish theoretical properties of COSALES. In particular, COSALES using the SCAD penalty or MCP is shown to consistently identify the two important subsets for the mean and scale simultaneously, even when the two subsets overlap. We demonstrate the empirical performance of SALES and COSALES by simulated and real data.

1. Introduction. High-dimensional data have received tremendous attention in the last decade due to the advance of data collection technology. Sparse estimation, which uses penalization or regularization techniques to perform variable selection and estimation simultaneously, has become a mainstream approach for analyzing high-dimensional data. Popular penalized estimators include the L_1-type selectors such as the Lasso [Tibshirani (1996)] and Dantzig [Candes and Tao (2007)] selectors and the nonconvex penalized estimators such as the SCAD [Fan and Li (2001)] and MCP [Zhang (2010)] estimators. Some embrace the L_1-regularization for its computational efficiency, while others prefer to use the nonconvex penalization due to its oracle [Fan and Li (2001)] property.

The current literature on sparse estimation often assumes homoscedasticity. For example, the existing theory for the sparse linear regression model is based on the

Received June 2015; revised November 2015.
[1]Supported in part by NSF Grant DMS-15-05111.
MSC2010 subject classifications. 62J07.
Key words and phrases. Asymmetric least squares, COSALES, high dimensions, SALES.

classical linear model assumption in which the mean function is linear and the errors are i.i.d. with zero mean and constant variance. The heteroscedasticity issue is often overlooked for theoretical convenience. However, heteroscedasticity often exists due to heterogeneity in measurement units or accumulation of outlying observations from numerous sources of inputs. This is particularly relevant with high-dimensional data. For example, in genomics experiments, tens of thousands of genes are often analyzed simultaneously by microarrays and occasional outlying measurements appearing in numerous experimental and data-preprocessing steps can accumulate to form heteroscedasticity in the data obtained therein. These data sets are often of high dimension since only a small number of subjects are available for the study. Several studies on expression quantitative trait loci (eQTLs) [Daye, Chen and Li (2012), Wang, Wu and Li (2012)] confirmed the presence of heteroscedasticity in these high-dimensional data and it was shown that genetic variants have effects on both the mean and the scale (i.e., standard deviation) of gene expression levels. In such scenarios, it is important to incorporate heteroscedasticity to make inference from the limited amount of data. To our knowledge, most existing work on high-dimensional data analysis fails to address the heteroscedasticity issue.

The sparse quantile regression was proposed in Wang, Wu and Li (2012) to detect heteroscedasticity in high-dimensional data. Quantile regression [Koenker and Bassett (1978)] is appropriate under heteroscedasticity, because it uses an asymmetric absolute value loss. The key word is "asymmetric," not the absolute value loss. The absolute value loss is computationally more challenging than the squared error loss. Computational efficiency is always one of the primary considerations in high-dimensional data analysis. This motivates us to study the asymmetric least squares (ALS) regression under high dimensionality. The ALS regression has been studied in Efron (1991). It is also known as the expectile regression in econometrics and finance. See Kuan, Yeh and Hsu (2009), Newey and Powell (1987), Taylor (2008), Xie, Zhou and Wan (2014). The key idea in ALS is to assign different squared error loss to the positive and negative residuals, respectively. By doing so, one can infer a more complete description of the conditional distribution than ordinary least squares (OLS). Thus, ALS and quantile regression share a common virtue although they differ technically. The most notable advantage of ALS over quantile regression is that the former employs a smooth differentiable loss, which considerably alleviates the computational effort involved and also makes the theoretical analysis more amenable. These two are desirable properties under high dimensionality.

In this paper, we develop the methodology and theory for the Sparse Asymmetric LEast Squares (SALES) regression and show its applications in detecting heteroscedasticity in a general class of sparse models in which the set of relevant covariates may vary from segment to segment on the conditional distribution. For the nonconvex penalized SALES regression, we prove its strong oracle property. We then discuss an important issue overlooked by existing methods dealing with

heteroscedasticity in high dimensional data, that is, how to exactly differentiate the sets of relevant covariates for the mean and scale when they have overlaps. To resolve this issue, we propose a novel COupled Sparse Asymmetric LEast Squares (COSALES) regression method to select important variables for the mean and scale of the conditional distribution simultaneously. The strong oracle property is also shown for the nonconvex penalized COSALES estimator. We develop novel efficient algorithms for computing both SALES and COSALES.

The remainder of the article is organized as follows. We study SALES in Section 2 and demonstrate its application in detecting heteroscedasticity in Section 3. In Section 4, we introduce and study COSALES. The performance of COSALES is illustrated by two simulation examples. In Section 5, we apply SALES and COSALES to analyze a real microarray dataset. The proofs of all main theoretical results are relegated to Section 6.

2. High-dimensional SALES regression.

2.1. *Background and setup.* We start by defining the τ-mean of a random variable $Z \in \mathbb{R}$,

$$\mathscr{E}^{\tau}(Z) \equiv \arg\min_{a \in \mathbb{R}} \mathbb{E}\{\Psi_{\tau}(Z-a)\}, \qquad \tau \in (0,1), \tag{2.1}$$

where $\Psi_{\tau}(u) = |\tau - I(u<0)|u^2$ is the asymmetric squared error loss [see, e.g., Efron (1991), Newey and Powell (1987)] and $I(\cdot)$ represents the indicator function. Similar definition can be found in Efron (1991). As a matter of fact, our τ-mean corresponds to Efron's w-mean, where $w = \tau/(1-\tau)$. Hereafter, we call $\mathscr{E}^{\tau}$ the asymmetric expectation operator (with asymmetry coefficient τ). Note that $\mathscr{E}^{0.5}$ coincides with the usual expectation operator $\mathbb{E}$. The τ-mean is also called the τ-expectile in the econometrics literature [Newey and Powell (1987)]. By varying τ, the τ-mean quantifies different "locations" of a distribution, and thus it can be viewed as a generalization of the mean and an alternative measure of "location" of a distribution.

The asymmetric squared error loss $\Psi_{\tau}(\cdot)$ gives rise to the ALS regression, in which the squared error loss is given different weights depending on whether the residual is positive or negative. Let $\mathbf{X} = (X_1, \ldots, X_p)$ be the $n \times p$ design matrix with $X_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$, $j = 1, \ldots, p$, and $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be the n-dimensional response vector. The design matrix may also be written as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, where $\mathbf{x}_i = (x_{i1} \ldots, x_{ip})^{\mathrm{T}}, i = 1, \ldots, n$. The ALS regression is done via

$$\hat{\boldsymbol{\beta}}_{\tau}^{\mathrm{ALS}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \Psi_{\tau}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}).$$

When $\tau = 0.5$, the ALS regression reduces to the OLS regression. When $\tau \neq 0.5$, due to the asymmetric nature and relative smoothness of $\Psi_{\tau}(\cdot)$, the ALS regression provides a convenient and computationally efficient way of summarizing the

conditional distribution of a response variable given the covariates [Efron (1991), Newey and Powell (1987)]. Applications of the ALS regression include estimation of the value at risk and expected shortfall [Kuan, Yeh and Hsu (2009), Taylor (2008)], medical baseline correction [Eilers and Boelens (2005)], and small area estimation [Chambers and Tzavidis (2006), Salvati et al. (2012)] among others.

In the literature, the underlying model considered for studying the theoretical property of the ALS regression is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{\tau} + \boldsymbol{\varepsilon}^{\tau}, \tag{2.2}$$

where $\boldsymbol{\beta}^{\tau}$ is a p-dimensional vector of unknown parameters and $\boldsymbol{\varepsilon}^{\tau}$ is the vector of n independent errors, which satisfy $\mathscr{E}^{\tau}(\varepsilon_i^{\tau}|\mathbf{x}_i) = 0, i = 1, \ldots, n$ for some $\tau \in (0, 1)$. It follows that $\mathscr{E}^{\tau}(y_i|\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{\tau}$, which means that the conditional τ-mean of y_i is a linear combination of $\mathbf{x}_i, i = 1, \ldots, n$. A model similar to (2.2) was considered in Wang, Wu and Li (2012) for quantile regression where the conditional quantile of the response variable was modeled as a linear combination of the covariates. In model (2.2), it is important to realize that the coefficient vector $\boldsymbol{\beta}^{\tau}$ is allowed to change with τ, which makes modeling for different "locations" of the conditional distribution possible, and as a result heteroscedasticity in the data, when it exists, can be inspected by this model. For convenience, we will drop the superscript for $\boldsymbol{\beta}^{\tau}$ and $\boldsymbol{\varepsilon}^{\tau}$ when no confusion arises.

To accommodate high-dimensional data in model (2.2), we allow the number of covariates p to increase with the sample size n, and moreover, we are primarily interested in cases where p exceeds n ($p > n$). We adopt the sparsity assumption that only a small number of covariates contribute to the response. Suppose $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^{\mathrm{T}}$ is the parameter vector of the true underlying model that generates the data and assume $\boldsymbol{\beta}^*$ is s-sparse, where $s = |A|$ with $A \equiv \operatorname{supp}(\boldsymbol{\beta}^*) = \{j: \beta_j^* \neq 0\}$.

2.2. *Methodology.* To select important variables and estimate $\boldsymbol{\beta}$ in model (2.2) when the dimension is high, let us consider the following penalized SALES regression:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} n^{-1}\sum_{i=1}^{n}\Psi_{\tau}\big(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\big) + \sum_{j=1}^{p} p_{\lambda}(\beta_j), \tag{2.3}$$

where $\Psi_{\tau}(\cdot)$ is the asymmetric squared error loss and $p_{\lambda}(\cdot)$ is a nonnegative penalty function with regularization parameter $\lambda \in (0, \infty)$. In the remainder of this article, we mainly focus on the Lasso and nonconvex penalties.

2.2.1. *L_1-Penalized SALES regression.* For ease of notation, let $\mathcal{L}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\Psi_{\tau}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$. The L_1-penalized SALES estimator or SALES Lasso estimator $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ is defined as the solution to the minimization problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda_{\mathrm{lasso}}\sum_{j=1}^{p}|\beta_j|, \qquad \lambda_{\mathrm{lasso}} \in (0, \infty). \tag{2.4}$$

This is to take $p_\lambda(u) = \lambda|u|$ in (2.3). The Lasso is computationally attractive and can be solved by efficient algorithms such as the LARS [Efron et al. (2004)], the coordinate descent method [Friedman, Hastie and Tibshirani (2010)] and the generalized coordinate descent algorithm [Yang and Zou (2013)].

For efficient computation of $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ in (2.4), we propose an algorithm called SALES which combines the cyclic coordinate descent [Tseng (2001)] and proximal gradient algorithms [Parikh and Boyd (2013)]. Our algorithm solves the following more general "weighted" L_1-minimization problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{L}_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} w_j|\beta_j| \tag{2.5}$$

with constants $w_j \geq 0$ for all j. Our consideration of formulation (2.5) is twofold. First, it not only can be directly applied to the SALES Lasso problem (2.4) by setting $w_j = \lambda_{\text{lasso}}$ for all j, but also can be used to solve the convex approximations to the nonconvex penalized SALES estimation [see step (a) of Algorithm 2]. Second, leaving some coefficients unpenalized is simply a matter of setting their corresponding weights to zero. Doing so gives us the flexibility to decide which covariates should always be kept in the model. The algorithm is described as follows.

For $\mathbf{v} = (v_1, \ldots, v_d)^{\mathrm{T}} \in \mathbb{R}^d$, denote $\mathbf{v}_{-k} = (v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_d)^{\mathrm{T}}$ the subvector of $\mathbf{v}$ with its kth component removed. Recover $\mathbf{v}$ from $\mathbf{v}_{-k}$ by writing $\mathbf{v} = [v_k, \mathbf{v}_{-k}]$. Let $\boldsymbol{\beta}^r = (\beta_1^r, \ldots, \beta_p^r)^{\mathrm{T}}$ be the update of $\boldsymbol{\beta}$ after the rth ($r \geq 0$) cycle of the coordinate descent algorithm. For ease of notation, denote

$$\mathbf{b}_{-k}^{r+1} = (\beta_1^{r+1}, \ldots, \beta_{k-1}^{r+1}, \beta_{k+1}^r, \ldots, \beta_p^r)^{\mathrm{T}}, \qquad 1 \leq k \leq p, r \geq 0.$$

Applying the coordinate descent method, to update β_k in the $(r+1)$th cycle, we solve the following minimization problem:

$$\min_{\beta_k\in\mathbb{R}} \ell_n(\beta_k; \mathbf{b}_{-k}^{r+1}) + w_k|\beta_k|, \tag{2.6}$$

where $\ell_n(\beta_k; \mathbf{b}_{-k}^{r+1}) = \mathcal{L}_n([\beta_k, \mathbf{b}_{-k}^{r+1}]) = n^{-1}\sum_{i=1}^n \Psi_\tau(y_i - \mathbf{x}_{i,-k}^{\mathrm{T}}\mathbf{b}_{-k}^{r+1} - x_{ik}\beta_k)$. One can show that $\ell_n'(\beta_k; \mathbf{b}_{-k}^{r+1})$ is Lipschitz continuous with constant $L_k = 2\bar{c}n^{-1}\|X_k\|_2^2$, where $\|\cdot\|_2$ is the Euclidean norm. Thus, the proximal gradient method can be employed to solve problem (2.6):

$$\beta_k^{r,0} := \beta_k^r, \qquad \beta_k^{r,s+1} := \mathbb{S}_{L_k^{-1}w_k}(\beta_k^{r,s} - L_k^{-1}\ell_n'(\beta_k^{r,s}; \mathbf{b}_{-k}^{r+1})), \qquad s \geq 0, \tag{2.7}$$

where $\mathbb{S}_v(u) = \operatorname{sgn}(u)(|u| - v)^+$ denotes the soft thresholding operator with $u^+ = uI(u > 0)$. We let (2.7) run for s_k^r iterations and set $\beta_k^{r+1} := \beta_k^{r,s_k^r}$. Our algorithm is summarized in Algorithm 1. We prove in Gu and Zou (2015) that Algorithm 1 converges at least linearly.

Algorithm 1: SALES—The cyclic coordinate descent plus proximal gradient algorithm for solving the weighted L_1-minimization problem (2.5)

1. Initialize the algorithm with $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)^{\mathrm{T}}$.
2. For $r = 0, 1, 2, \ldots, m-1$,
   - (2.1) For $k = 1, \ldots, p$,
     - (2.1.1) Initialize $\beta_k^{r,0} := \beta_k^r$.
     - (2.1.2) For $s = 0, 1, 2, \ldots, s_k^r - 1$,
       - (2.1.2.1) Calculate $\beta_k^{r,s+1} := \mathbb{S}_{L_k^{-1} w_k}(\beta_k - L_k^{-1} \ell_n'(\beta_k^{r,s}; \mathbf{b}_{-k}^{r+1}))$.
     - (2.1.3) Set $\beta_k^{r+1} := \beta_k^{r,s_k^r}$.
   - (2.2) Set $\boldsymbol{\beta}^{r+1} := (\beta_1^{r+1}, \ldots, \beta_p^{r+1})^{\mathrm{T}}$.
3. Output $\hat{\boldsymbol{\beta}} := \boldsymbol{\beta}^m$.

2.2.2. *Nonconvex penalized SALES regression*. Nonconvex penalties have been used in a broad type of sparse regression models [Fan and Lv (2011), Fan, Xue and Zou (2014), Wang, Kim and Li (2013)]. The most popular nonconvex penalties include the smoothly clipped absolute deviation (SCAD) penalty [Fan and Li (2001)] and the minimax concave penalty [MCP, Zhang (2010)]. For some constant $\gamma > 2$, the SCAD penalty is given by

$$
\begin{aligned}
p_\lambda(u) = {} & \lambda|u| I\big(|u| \le \lambda\big) + \left\{\lambda|u| - \frac{(\lambda - |u|)^2}{2(\gamma - 1)}\right\} I\big(\lambda < |u| \le \gamma\lambda\big) \\
& + \frac{(\gamma + 1)\lambda^2}{2} I\big(|u| > \gamma\lambda\big).
\end{aligned} \tag{2.8}
$$

The use of $\gamma = 3.7$ for the SCAD penalty is recommended in Fan and Li (2001) from a Bayesian perspective. The MCP is characterized by

$$
p_\lambda(u) = \lambda\left(|u| - \frac{u^2}{2\gamma\lambda}\right) I\big(|u| \le \gamma\lambda\big) + \frac{\gamma\lambda^2}{2} I\big(|u| > \gamma\lambda\big) \tag{2.9}
$$

for some $\gamma > 1$. The use of $\gamma = 2$ is suggested in Zhang (2010). In this article, we consider both SCAD and MCP penalized SALES regression.

The main motivation for using the nonconvex penalties is to achieve the oracle property. For the SALES regression, the oracle estimator is

$$
\hat{\boldsymbol{\beta}}^{\mathrm{oracle}} = \mathop{\arg\min}_{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}_{A^{\mathrm{c}}} = \mathbf{0}} \mathcal{L}_n(\boldsymbol{\beta}). \tag{2.10}
$$

In practice, the oracle estimator is infeasible, but it sets a benchmark for evaluation of other estimators. Many papers have shown that the nonconvex penalized least squares can find the oracle estimator with high probability [Fan, Xue and

Algorithm 2: The local linear approximation (LLA) algorithm for solving the nonconvex penalized SALES estimation problem (2.3)

1. Initialize $\hat{\boldsymbol{\beta}}^0 := \hat{\boldsymbol{\beta}}^{\text{initial}}$. Compute weights $\hat{w}_j^0 = p'_\lambda(|\hat{\beta}_j^0|)$, $j = 1, \ldots, p$.
2. For $m = 1, 2, \ldots,$ repeat the LLA iteration in (a) and (b) until convergence
   (a) Solve the following convex optimization problem for $\hat{\boldsymbol{\beta}}^m$

$$\hat{\boldsymbol{\beta}}^m := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{L}_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} \hat{w}_j^{m-1} |\beta_j|.$$

   (b) Update the weights $\hat{w}_j^m = p'_\lambda(|\hat{\beta}_j^m|)$, $j = 1, \ldots, p$.

Zou (2014), Wang, Kim and Li (2013)]. In particular, Fan, Xue and Zou (2014) showed that the local linear approximation (LLA) algorithm [Zou and Li (2008)] converges to the oracle estimator under regularity conditions. The LLA algorithm fits a sequence of weighted L_1-regularization problems. Since we already have Algorithm 1 for computing any weighted L_1-penalized SALES regression, we adopt the LLA algorithm for solving the nonconvex penalized SALES estimation problem (2.3). The details of the LLA algorithm are shown in Algorithm 2. Note that step (a) can be readily solved by Algorithm 1.

In our numerical examples, we tried using both the SALES Lasso estimator and zero as the initial values of the LLA algorithm for computing the nonconvex penalized SALES estimator. Our practice is based on theoretical results in Section 2.3.

2.3. *Theory.* In this section, we theoretically analyze the SALES regression. We consider the case where the covariates are from a fixed design.

The following notation will be used. For any vector $\mathbf{v} = (v_1, \ldots, v_p)^{\mathrm{T}} \in \mathbb{R}^p$ and an arbitrary index set $I \subset \{1, \ldots, p\}$, we write $\mathbf{v}_I = (v_j, j \in I)^{\mathrm{T}}$ and denote by $\mathbf{X}_I = (\mathbf{x}_j, j \in I)$ the submatrix consisting of the columns of $\mathbf{X}$ with indices in I. The complement of I is denoted by $I^{\mathrm{c}} = \{1, \ldots, p\} \setminus I$. For $q \in [1, \infty]$, the L_q-norm of $\mathbf{v}$ is denoted by $\|\mathbf{v}\|_q$. Sub-Gaussian norm [Rudelson and Vershynin (2013)] of a random variable Z is denoted by $\|Z\|_{\mathrm{SG}} = \sup_{k \geq 1} k^{-1/2} (\mathbb{E}|Z|^k)^{1/k}$. Let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$ for real numbers a and b. For a differentiable function $f : \mathbb{R}^p \to \mathbb{R}$, we write $\nabla f(\mathbf{v}) = \partial f(\mathbf{v})/\partial \mathbf{v}$ and $\nabla_I f(\mathbf{v}) = (\partial f(\mathbf{v})/\partial v_j, j \in I)^{\mathrm{T}}$. We use $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ to represent respectively the smallest and largest eigenvalues of a symmetric matrix. We also let $\underline{c} = \tau \wedge (1 - \tau)$ and $\bar{c} = \tau \vee (1 - \tau)$.

2.3.1. *L_1-Penalized SALES regression.* The estimation accuracy of the Lasso has been extensively studied in the literature; see, for example, Negahban et al. (2012) and Ye and Zhang (2010). Let $\mathscr{C} = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_{A^{\mathrm{c}}}\|_1 \leq 3\|\boldsymbol{\delta}_A\|_1 \neq 0\}$ be a

cone in $\mathbb{R}^p$. Denote $\rho_{\min} = \lambda_{\min}(n^{-1}\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)$ and $\rho_{\max} = \lambda_{\max}(n^{-1}\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)$. We assume $\rho_{\min} > 0$ so that the important variables are not linearly dependent. To study the estimation accuracy of the SALES Lasso, we impose the following conditions on the design matrix $\mathbf{X}$ and the random errors $\boldsymbol{\varepsilon}$.

(C1) The columns of $\mathbf{X}$ are normalizable, that is, $M_0 = \max_{1\le j\le p} \frac{\|X_j\|_2}{\sqrt{n}} \in (0,\infty)$.

(C2) The random errors ε_i are i.i.d. sub-Gaussian random variables satisfying $\mathscr{E}^\tau(\varepsilon_i) = 0, i = 1,\dots,n$.

(C3) $\kappa = \inf_{\boldsymbol{\delta}\in\mathscr{C}} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}\|_2^2} \in (0,\infty)$.

(C4) $\varrho = \inf_{\boldsymbol{\delta}\in\mathscr{C}} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_A\|_1\|\boldsymbol{\delta}\|_\infty} \in (0,\infty)$.

Condition (C3) is called the restricted eigenvalue condition and has been frequently assumed in the literature to study the Lasso and Dantzig selectors. See Bickel, Ritov and Tsybakov (2009), Meier, van de Geer and Bühlmann (2009), and Negahban et al. (2012). Condition (C4), the generalized invertability factor (GIF) condition, is closely related to condition (C3) and has also been often adopted to study the Lasso and Dantzig selectors. See discussion of these conditions in Ye and Zhang (2010) and Huang and Zhang (2012). Both conditions (C3) and (C4) are crucial assumptions to establish estimation consistency of the Lasso for high-dimensional data.

THEOREM 1. *Suppose in model* (2.2) *the true coefficients* $\boldsymbol{\beta}^*$ *are* s*-sparse and assume conditions* (C1)–(C2) *hold. Let* $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ *be any optimal solution to the SALES Lasso problem* (2.4). *Then with probability at least* $1 - p_1^{\text{ALS}}$, $\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^*\|_2 \le 3s^{1/2}\lambda_{\text{lasso}}(4\kappa\underline{c})^{-1}$ *if condition* (C3) *holds, and* $\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^*\|_\infty \le 3\lambda_{\text{lasso}}(4\varrho\underline{c})^{-1}$ *if condition* (C4) *holds, where*

$$p_1^{\text{ALS}} = 2p\exp\Big(-\frac{Cn\lambda_{\text{lasso}}^2}{4K_0^2M_0^2}\Big),$$

$K_0 = \|\Psi_\tau'(\varepsilon_i)\|_{\text{SG}}$ *with* $\Psi_\tau'(\cdot)$ *being the derivative of* $\Psi_\tau(\cdot)$ *and* $C > 0$ *is an absolute constant.*

REMARK 1. In some applications, it is natural to leave a given subset of the parameters unpenalized in the penalized framework (2.3). Let $\mathscr{R}$ denote the index set of such parameters. For example, when X_1 is a vector consisting of all ones, $\mathscr{R} = \{1\}$ reflects the common practice of leaving the intercept term not penalized. In this case, it is natural to modify the penalized SALES estimation problem (2.3) to be

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{L}_n(\boldsymbol{\beta}) + \sum_{j\in\mathscr{R}^{\text{c}}} p_\lambda(\beta_j).$$

With Lasso penalty, the SALES algorithm can be readily used to solve the above case. Moreover, similar theoretical analysis can be carried out with slight modifications. For instance, in the SALES Lasso problem (2.4) we can define $A' \equiv \operatorname{supp}(\boldsymbol{\beta}^*_{\mathscr{R}^c})$ and $\mathscr{C}' = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_{(A'\cup\mathscr{R})^c}\|_1 \le 3\|\boldsymbol{\delta}_{A'\cup\mathscr{R}}\|_1 \ne 0\}$. Conditions (C3) and (C4) can be then modified respectively as

$$\kappa' = \inf_{\boldsymbol{\delta}\in\mathscr{C}'} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}\|_2^2} \in (0,\infty) \quad \text{and} \quad \varrho' = \inf_{\boldsymbol{\delta}\in\mathscr{C}'} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_{A'\cup\mathscr{R}}\|_1\|\boldsymbol{\delta}\|_\infty} \in (0,\infty).$$

To establish the selection consistency of the Lasso, it is almost necessary to impose the irrepresentable condition; see Zou (2006) and Zhao and Yu (2006). When the focus is on identifying the underlying sparsity pattern, the nonconvex penalized regression is a competitive alternative as it requires weaker conditions to achieve selection consistency.

2.3.2. *Nonconvex penalized SALES regression.* To offer a unified treatment of the SCAD and MCP penalized SALES regression, our theoretical analysis handles the following class of nonconvex penalties:

(P1) $p_\lambda(u) = p_\lambda(-u)$;
(P2) $p_\lambda(u)$ is nondecreasing and concave in $u \in [0,\infty)$ and $p_\lambda(0) = 0$;
(P3) $p_\lambda(u)$ is differentiable in $u \in (0,\infty)$;
(P4) $p'_\lambda(u) \ge a_1\lambda$ for $u \in (0, a_2\lambda]$ and $p'_\lambda(0) := p'_\lambda(0+) \ge a_1\lambda$;
(P5) $p'_\lambda(u) = 0$ for $u \in [a\lambda,\infty)$ with some prespecified constant $a > a_2$,

where a_1 and a_2 are fixed constants characteristic of the penalty functions. It is easy to verify that both the SCAD penalty and MCP are in the above class.

We show that the sparse solutions obtained by the LLA algorithm in Section 2.2.2 possess the oracle property. Assume sufficient signal strength in the nonzero components of $\boldsymbol{\beta}^*$

(A1) $\min_{j\in A}|\beta^*_j| > (a+1)\lambda$.

THEOREM 2. *Suppose in model* (2.2) *the true coefficients* $\boldsymbol{\beta}^*$ *are* s*-sparse and satisfy assumption* (A1). *Assume conditions* (C1)–(C2) *hold and take* $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ *as the initial value. Let* $a_0 = 1 \wedge a_2$. *Take* $\lambda \ge 3s^{1/2}\lambda_{\text{lasso}}(4a_0\kappa\underline{c})^{-1}$ *when* (C3) *holds, or take* $\lambda \ge 3\lambda_{\text{lasso}}(4a_0\varrho\underline{c})^{-1}$ *when* (C4) *holds, or take* $\lambda \ge [3s^{1/2}\lambda_{\text{lasso}}(4a_0\kappa\underline{c})^{-1}] \wedge [3\lambda_{\text{lasso}}(4a_0\varrho\underline{c})^{-1}]$ *when both* (C3) *and* (C4) *hold. The LLA algorithm* (*Algorithm* 2) *converges to* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *after two iterations with probability at least* $1 - p_1^{\text{ALS}} - p_2^{\text{ALS}} - p_3^{\text{ALS}}$, *where* p_1^{ALS} *is given in Theorem* 1,

$$p_2^{\text{ALS}} = 2(p-s)\exp\Big(-\frac{Ca_1^2n\lambda^2}{4K_0^2M_0^2}\Big) + \Gamma(Q_1\lambda; n, s, K_0, M_0, \rho_{\max}, \nu_0)$$

and

$$p_3^{\mathrm{ALS}} = \Gamma(2\underline{c}\rho_{\min}R; n, s, K_0, M_0, \rho_{\max}, \nu_0),$$

where $Q_1 = a_1\underline{c}\rho_{\min}(2\bar{c}\rho_{\max}^{1/2}M_0)^{-1}$, $\nu_0 = \mathrm{var}(\Psi_\tau'(\varepsilon_i))$, $R = \min_{j\in A}|\beta_j^*| - a\lambda$, K_0 *is defined in Theorem* 1 *and* $\Gamma(\cdot)$ *is a function defined by*

$$\Gamma(x; n, s, K, M, \rho, \nu) = 2s\exp\left(-\frac{Cnx^2}{K^2M^2s}\right) \wedge 2\exp\left(-\frac{C\nu^2[(n^{1/2}x - \nu\rho^{1/2}s^{1/2})^+]^2}{K^4\rho}\right),$$

and $C > 0$ *is an absolute constant.*

It is interesting to note that with the SCAD penalty or MCP, a three-step LLA algorithm starting from the zero vector may also work. Indeed, for these two penalties we have $p_\lambda'(0) = \lambda$, so if we can take $\lambda = \lambda_{\mathrm{lasso}}$, this would give us the SALES Lasso estimator in the second step.

COROLLARY 1. *Assume the same framework of Theorem* 2 *and suppose the SCAD penalty* (2.8) *or MCP* (2.9) *is used. If condition* (C3) *holds and* $4a_0\kappa\underline{c} \geq 3s^{1/2}$, *or if condition* (C4) *holds and* $4a_0\varrho\underline{c} \geq 3$, *or if both* (C3) *and* (C4) *hold and* $[3s^{1/2}(\kappa)^{-1}] \wedge [3(\varrho)^{-1}] \leq 4a_0\underline{c}$, *the LLA algorithm* (*Algorithm* 2) *initialized by zero converges to the oracle estimator after three iterations with probability at least* $1 - 2p\exp\{-Cn\lambda^2(4K_0^2M_0^2)^{-1}\} - p_2^{\mathrm{ALS}} - p_3^{\mathrm{ALS}}$, *where* p_2^{ALS} *and* p_2^{ALS} *are given in Theorem* 2.

3. Application of SALES: Detecting heteroscedasticity. Due to asymmetry of the squared error loss, the SALES regression (2.3) can be employed to detect heteroscedasticity in high-dimensional data. In the following, we use a simulation example to illustrate this application. For the nonconvex penalty functions used in the simulation, we fix $\gamma = 3.7$ for the SCAD penalty (2.8) and $\gamma = 2$ for the MCP (2.9).

EXAMPLE 1. We adopt a model from Wang, Wu and Li (2012). In the model, the covariates are generated in two steps. First, we generate copies of $(z_1, \ldots, z_p)^{\mathrm{T}}$ from the multivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (0.5^{|i-j|})_{p\times p}$. In the second step, for each copy of $(z_1, \ldots, z_p)^{\mathrm{T}}$, we set $x_1 = \Phi(z_1)$ and $x_j = z_j$ for $j = 2, 3, \ldots, p$, where $\Phi(\cdot)$ is the standard normal CDF. The response is then simulated from the following normal linear heteroscedastic model:

$$y = x_6 + x_{12} + x_{15} + x_{20} + (0.7x_1)\varepsilon, \tag{3.1}$$

where $\varepsilon \sim N(0, 1)$ is independent of the covariates. This model was considered in Wang, Wu and Li (2012) for the sparse quantile regression, where a sample size $n = 300$ and covariate dimensions $p = 400$ and 600 were considered.

We apply the SALES regression (2.3) instead to select active variables and estimate the coefficients for this model. For the purpose of demonstration, we choose $n = 300$ and $p = 600$. A validation set of size $n = 300$ is generated independently to tune the regularization parameter by minimizing the validation error $\sum_{i\in\text{validation}} \Psi_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})$ for the computed estimate $\hat{\boldsymbol{\beta}}$, where $\tau = 0.5$ and 0.85 are considered.

For comparison purpose, we included in this simulation the SALES Lasso (2.4) and two variations of the LLA algorithm for each nonconvex penalized SALES regression: the two-step LLA algorithm initialized by the Lasso estimator (SCAD*, MCP*), and the three-step LLA algorithm initialized by zero (SCAD^0, MCP^0).

Let $\hat{\boldsymbol{\beta}}$ be the coefficient estimates from a given method. Based on 100 replicates, the following measurements are calculated to evaluate the sparsity recovery and estimation performance of that method:

$|\hat{A}|$: the average size of the active set $\hat{A} = \{j: \hat{\beta}_j \neq 0\}$ of $\hat{\boldsymbol{\beta}}$.

p_a: proportion of the event $A \subset \hat{A}$, where A is the active set of $\boldsymbol{\beta}^*$. When $\tau = 0.5$, $A = \{6, 12, 15, 20\}$ and when $\tau \neq 0.5$, $A = \{1, 6, 12, 15, 20\}$.

p_1: proportion of the event that $\{1\} \subset \hat{A}$.

R_1: the average L_1 risk $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$.

R_2: the average L_2 risk $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$.

The simulation results are shown in Table 1. The following conclusions can be made:

TABLE 1

Numerical summary of simulation results from the Lasso, SCAD and MCP penalized SALES regression for model (3.1): $y = x_6 + x_{12} + x_{15} + x_{20} + (0.7x_1)\varepsilon$. The sparsity recovery performance is measured by the selected active set size $|\hat{A}|$, the proportion p_a of covering the true active set and the proportion p_1 of selecting the signature variable X_1. The estimation accuracy is measured by the L_1 risk R_1 and the L_2 risk R_2. The results are shown as averages over 100 *replicates with standard errors listed in the parentheses when available*

	Method	$\|\hat{A}\|$	p_a	p_1	R_1	R_2
$\tau = 0.5$	SALES-Lasso	25.82 (1.15)	100%	0%	0.399 (0.015)	0.120 (0.003)
	SALES-SCAD*	7.75 (0.68)	100%	0%	0.103 (0.006)	0.049 (0.002)
	SALES-SCAD^0	6.65 (0.68)	100%	0%	0.100 (0.006)	0.050 (0.002)
	SALES-MCP*	6.39 (0.48)	100%	0%	0.099 (0.005)	0.049 (0.002)
	SALES-MCP^0	5.75 (0.29)	100%	0%	0.093 (0.004)	0.049 (0.002)
$\tau = 0.85$	SALES-Lasso	34.17 (1.26)	100%	100%	0.714 (0.016)	0.249 (0.005)
	SALES-SCAD*	7.52 (0.51)	100%	100%	0.160 (0.009)	0.083 (0.005)
	SALES-SCAD^0	8.19 (0.59)	100%	100%	0.166 (0.007)	0.084 (0.003)
	SALES-MCP*	6.30 (0.25)	100%	100%	0.148 (0.005)	0.079 (0.003)
	SALES-MCP^0	6.35 (0.23)	100%	100%	0.147 (0.005)	0.078 (0.003)

(1) The variable x_1 in the scale function is often not recovered by penalized least-squares ($\tau = 0.5$). However, when several τ-means (e.g., $\tau = 0.85$) are inspected together, it is possible to detect this variable with high probability. This shows that indeed the SALES regression can be used to detect heteroscedasticity.

(2) Compared to the SALES Lasso, the nonconvex penalized SALES regression selects much fewer irrelevant covariates and has better estimation accuracy.

(3) The three-step LLA algorithm starting from zero produces similar results to the two-step LLA algorithm starting from the Lasso solution.

4. High-dimensional COSALES regression. In Section 3, we showed that the SALES regression provides a means of detecting heteroscedasticity in high-dimensional data. Indeed, in the linear heteroscedastic model (3.1), the signature variable x_1, which appears in the scale function, was detected through comparison of different τ-means. However, in high-dimensional heteroscedastic models, often of more interest are the sparsity patterns in both the mean and the scale functions of the conditional distribution. The SALES regression and methods proposed by other authors, for example, Wang, Wu and Li (2012), are not sufficient to fulfill this task. To see it, consider a linear heteroscedastic model in which the active set for the mean is $\{1, 2\}$ and the active set for the scale is $\{1, 3\}$. Suppose the SALES regression can exactly recover the active variables. Then the method picks x_1 and x_2 when $\tau = 0.5$ and hopefully x_1, x_2, and x_3 when $\tau \neq 0.5$. A natural question is whether the scale function depends on x_1. With the SALES regression, we cannot answer this question. This motivates us to consider the COSALES regression for a general class of models and gain some insight into analyzing heteroscedasticity in high-dimensional data.

4.1. *Formulation and computation.* Consider the following model of systematic heteroscedasticity:

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma} + (\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\omega})\varepsilon_i, \qquad i = 1, \ldots, n, \tag{4.1}$$

where ε_i are i.i.d. random errors that are independent of the covariates and that have distribution F_0 with $\mathbb{E}(\varepsilon_i) = \int_{\mathbb{R}} x \, \mathrm{d}F_0(x) = 0$; $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ are unknown p-dimensional parameter vectors controlling the conditional mean and scale; and $\boldsymbol{\omega}$ is assumed to satisfy $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\omega} > 0$ for all i. The intercept can be included by letting $x_{i1} = 1$. The linear scale model of heteroscedasticity (4.1) is an important model considered by many authors [Efron (1991), Koenker and Bassett (1982), Koenker and Zhao (1994)] for analyzing heteroscedasticity.

Let $A_1 \equiv \operatorname{supp}(\boldsymbol{\gamma}^*) = \{j : \gamma_j^* \neq 0\}$ and $A_2 \equiv \operatorname{supp}(\boldsymbol{\omega}^*) = \{j : \omega_j^* \neq 0\}$ be the active sets of $\boldsymbol{\gamma}^*$ and of $\boldsymbol{\omega}^*$, respectively. Suppose $|A_1| = s_1$ and $|A_2| = s_2$. Let $e_\tau = \mathscr{E}^\tau(\varepsilon_1)$ be the τ-mean of the random error for $\tau \in (0, 1)$. It follows that the τ-mean of y_i given $\mathbf{x}_i$ is $\mathscr{E}^\tau(y_i|\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\gamma} + \boldsymbol{\omega}e_\tau)$. To select significant variables in both the mean and the scale functions, we now propose the COSALES regression. Write $\boldsymbol{\varphi} = \boldsymbol{\omega}e_\tau$. Note that we omit the dependency of $\boldsymbol{\varphi}$ on τ to ease exposition.

In the COSALES regression, we will deal with $\boldsymbol{\varphi}$ instead of $\boldsymbol{\omega}$. However, when $e_\tau \neq 0$, it should be noted that since $\operatorname{supp}(\boldsymbol{\varphi}) = \operatorname{supp}(\boldsymbol{\omega})$, the selection result on $\boldsymbol{\varphi}$ applies to $\boldsymbol{\omega}$. Moreover, $\boldsymbol{\omega}$ can be estimated up to a scale from the estimate of $\boldsymbol{\varphi}$. Ideally, if the distribution F_0 of ε_i is known, exact estimation of $\boldsymbol{\omega}$ is possible.

For some $\tau \in (0, 1)$ and $\tau \neq 0.5$, let

$$S_n(\boldsymbol{\gamma}, \boldsymbol{\varphi}) = n^{-1} \sum_{i=1}^{n} \{\Psi_{0.5}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}) + \Psi_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\varphi})\}.$$

The COSALES regression tries to minimize

$$Q_n(\boldsymbol{\gamma}, \boldsymbol{\varphi}) = S_n(\boldsymbol{\gamma}, \boldsymbol{\varphi}) + \sum_{j=1}^{p} p_{\lambda_1}(\gamma_j) + \sum_{j=1}^{p} p_{\lambda_2}(\varphi_j), \tag{4.2}$$

over $\boldsymbol{\gamma}, \boldsymbol{\varphi} \in \mathbb{R}^p$, where $p_{\lambda_1}(\cdot)$ and $p_{\lambda_2}(\cdot)$ are penalty functions with regularization parameters $\lambda_1, \lambda_2 \in (0, \infty)$, respectively. Let $\hat{\boldsymbol{\gamma}}^{\text{oracle}}$ and $\hat{\boldsymbol{\varphi}}^{\text{oracle}}$ be the oracle estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi} = \boldsymbol{\omega} e_\tau$, respectively, in model (4.1),

$$(\hat{\boldsymbol{\gamma}}^{\text{oracle}}, \hat{\boldsymbol{\varphi}}^{\text{oracle}}) = \underset{\boldsymbol{\gamma}, \boldsymbol{\varphi} \in \mathbb{R}^p : \boldsymbol{\gamma}_{A_1^{\mathrm{c}}} = \mathbf{0}, \boldsymbol{\varphi}_{A_2^{\mathrm{c}}} = \mathbf{0}}{\arg\min} S_n(\boldsymbol{\gamma}, \boldsymbol{\varphi}). \tag{4.3}$$

In what follows, let us focus on the Lasso and nonconvex penalties.

4.1.1. *L_1-penalized COSALES regression.* For $\lambda_1^{\text{lasso}}, \lambda_2^{\text{lasso}} \in (0, \infty)$, the L_1-penalized COSALES estimators or the COSALES Lasso estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ can be achieved simultaneously by

$$(\hat{\boldsymbol{\gamma}}^{\text{lasso}}, \hat{\boldsymbol{\varphi}}^{\text{lasso}}) = \underset{\boldsymbol{\gamma}, \boldsymbol{\varphi} \in \mathbb{R}^p}{\arg\min}\, S_n(\boldsymbol{\gamma}, \boldsymbol{\varphi}) + \lambda_1^{\text{lasso}} \|\boldsymbol{\gamma}\|_1 + \lambda_2^{\text{lasso}} \|\boldsymbol{\varphi}\|_1. \tag{4.4}$$

We note that problem (4.4) is a special case of the minimization problem in step (a) of Algorithm 4 (Section 4.1.2) and efficient computation of the solutions can be carried out by an algorithm similar to Algorithm 1. The algorithm applies the cyclic coordinate descent and proximal gradient descent methods to $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ alternately. We call this algorithm COSALES and display it in Algorithm 3. Note that COSALES solves the general coupled weighted L_1-minimization problem

$$\min_{\boldsymbol{\gamma}, \boldsymbol{\varphi} \in \mathbb{R}^p} S_n(\boldsymbol{\gamma}, \boldsymbol{\varphi}) + \sum_{j=1}^{p} w_j |\gamma_j| + \sum_{j=1}^{p} v_j |\varphi_j|. \tag{4.5}$$

To facilitate the presentation, in Algorithm 3, we let $\boldsymbol{\gamma}^r$ and $\boldsymbol{\varphi}^r$ be the updates of $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ respectively after the rth cycle of the coordinate descent algorithm and denote

$$\mathbf{g}_{-k}^{r+1} = (\gamma_1^{r+1}, \ldots, \gamma_{k-1}^{r+1}, \gamma_{k+1}^{r}, \ldots, \gamma_p^{r}), \qquad 1 \leq k \leq p, r \geq 0,$$

and

$$\mathbf{p}_{-k}^{r+1} = (\varphi_1^{r+1}, \ldots, \varphi_{k-1}^{r+1}, \varphi_{k+1}^{r}, \ldots, \varphi_p^{r}), \qquad 1 \le k \le p, r \ge 0.$$

Theoretical justification of the estimation accuracy of the COSALES Lasso will be deferred to the next section.

4.1.2. *Nonconvex penalized COSALES regression.* In (4.2), let $p_{\lambda_1}(\cdot)$ and $p_{\lambda_2}(\cdot)$ be nonconvex penalties having properties (P1)–(P5). This nonconvex penalized COSALES estimation problem can be solved by the LLA algorithm shown in Algorithm 4. Note that the minimization problem in step (a) was solved in Algorithm 3. Oracle properties of the sparse solutions will be established in the following section.

4.2. *Theory.* In this section, we show the selection and estimation accuracy of the COSALES regression for both Lasso and nonconvex penalties.

Algorithm 3: COSALES—The coordinate descent plus proximal gradient algorithm for solving the coupled weighted L_1-minimization problem (4.5)

1. Initialize the algorithm with $\boldsymbol{\gamma}^0 = (\gamma_1^0, \ldots, \gamma_p^0)^{\mathrm{T}}$ and $\boldsymbol{\varphi}^0 = (\varphi_1^0, \ldots, \varphi_p^0)^{\mathrm{T}}$.
2. For $r = 1, \ldots, m-1$,
   - (2.1) For $k = 1, \ldots, p$,
     - (2.1.1) Initialize $\gamma_k^{r,0} := \gamma_k^r$.
     - (2.1.2) For $s = 0, 1, \ldots, s_{1k}^r - 1$,
       - (2.1.2.1) Compute $\gamma_k^{r,s+1} := \mathbb{S}_{L_{1k}^{-1} w_k}(\gamma_k^{r,s} - L_{1k}^{-1} h_n'(\gamma_k^{r,s}; \mathbf{g}_{-k}^{r+1}, \boldsymbol{\varphi}^r))$, where $L_{1k} = (2\bar{c}+1)n^{-1}\|X_k\|_2^2$; $h_n(\gamma_k; \mathbf{g}_{-k}^{r+1}, \boldsymbol{\varphi}^r) = S_n([\gamma_k, \mathbf{g}_{-k}^{r+1}], \boldsymbol{\varphi}^r)$.
     - (2.1.3) Set $\gamma_k^{r+1} := \gamma_k^{r, s_{1k}^r}$.
   - (2.2) Set $\boldsymbol{\gamma}^{r+1} := (\gamma_1^{r+1}, \ldots, \gamma_p^{r+1})^{\mathrm{T}}$.
   - (2.3) For $k = 1, \ldots, p$,
     - (2.3.1) Initialize $\varphi_k^{r,0} := \varphi_k^r$.
     - (2.3.2) For $s = 0, 1, \ldots, s_{2k}^r - 1$,
       - (2.3.2.1) Compute $\varphi_k^{r,s+1} := \mathbb{S}_{L_{2k}^{-1} v_k}(\varphi_k^{r,s} - L_{2k}^{-1} \hbar_n'(\varphi_k^{r,s}; \boldsymbol{\gamma}^{r+1}, \mathbf{p}_{-k}^{r+1}))$, where $L_{2k} = 2\bar{c}n^{-1}\|X_k\|_2^2$; $\hbar_n(\varphi_k; \boldsymbol{\gamma}^{r+1}, \mathbf{p}_{-k}^{r+1}) = S_n(\boldsymbol{\gamma}^{r+1}, [\varphi_k, \mathbf{p}_{-k}^{r+1}])$.
     - (2.3.3) Set $\varphi_k^{r+1} := \varphi_k^{r, s_{2k}^r}$.
   - (2.4) Set $\boldsymbol{\varphi}^{r+1} := (\varphi_1^{r+1}, \ldots, \varphi_p^{r+1})^{\mathrm{T}}$.
3. Output $\hat{\boldsymbol{\gamma}} := \boldsymbol{\gamma}^m$ and $\hat{\boldsymbol{\varphi}} := \boldsymbol{\varphi}^m$.

Algorithm 4: The local linear approximation (LLA) algorithm for solving the nonconvex penalized COSALES estimation problem (4.2)

1. Initialize $\hat{\boldsymbol{\gamma}}^0 = \hat{\boldsymbol{\gamma}}^{\text{initial}}$ and $\hat{\boldsymbol{\varphi}}^0 = \hat{\boldsymbol{\varphi}}^{\text{initial}}$. Compute weights
$$\hat{w}_j^0 = p'_{\lambda_1}(|\hat{\gamma}_j^0|), \qquad \bar{w}_j^0 = p'_{\lambda_2}(|\hat{\varphi}_j^0|), \qquad j = 1, \ldots, p.$$
2. For $m = 1, 2, \ldots,$ repeat the LLA iteration in (a) and (b) until convergence.
   (a) Solve the following convex optimization problem for $\hat{\boldsymbol{\gamma}}^m$ and $\hat{\boldsymbol{\varphi}}^m$
$$\min_{\boldsymbol{\gamma}, \boldsymbol{\varphi} \in \mathbb{R}^p} S_n(\boldsymbol{\gamma}, \boldsymbol{\varphi}) + \sum_{j=1}^p \hat{w}_j^{m-1} |\gamma_j| + \sum_{j=1}^p \bar{w}_j^{m-1} |\varphi_j|.$$
   (b) Update the weights
$$\hat{w}_j^m = p'_{\lambda_1}(|\hat{\gamma}_j^m|), \qquad \bar{w}_j^m = p'_{\lambda_2}(|\hat{\varphi}_j^m|), \qquad j = 1, \ldots, p.$$

4.2.1. *L_1-penalized COSALES regression.* For the Lasso problem (4.4), let $\breve{M} = (\lambda_1^{\text{lasso}}/\lambda_2^{\text{lasso}}) \vee (\lambda_2^{\text{lasso}}/\lambda_1^{\text{lasso}})$ and define set $A_0 = (A_1, A_2')$, where $A_2' = \{j + p : \omega_j^* \neq 0\}$. For $M \geq 1$, define $\mathscr{C}_M = \{\boldsymbol{\delta} \in \mathbb{R}^{2p} : \|\boldsymbol{\delta}_{A_0^c}\|_1 \leq M \|\boldsymbol{\delta}_{A_0}\|_1 \neq 0\}$. For $k = 1, 2$, let $\rho_{k\bullet\min} = \lambda_{\min}(n^{-1}\mathbf{X}_{A_k}^{\mathrm{T}}\mathbf{X}_{A_k})$ and $\rho_{k\bullet\max} = \lambda_{\max}(n^{-1}\mathbf{X}_{A_k}^{\mathrm{T}}\mathbf{X}_{A_k})$. Denote $\phi_{\min} = \rho_{1\bullet\min} \wedge \rho_{2\bullet\min}$ and $\phi_{\max} = \rho_{1\bullet\max} \vee \rho_{2\bullet\max}$. Assume $\phi_{\min} > 0$. Let $\mathbf{I}_2$ be a 2×2 identity matrix and let $\otimes$ denote the Kronecker product. To establish an error bound on the COSALES Lasso estimators, the following conditions on the design matrix $\mathbf{X}$ and the random errors $\boldsymbol{\varepsilon}$ are imposed:

(C1′) The columns of $\mathbf{X}$ is normalizable, that is, $M_0 = \max_{1 \leq j \leq p} \frac{\|X_j\|_2}{\sqrt{n}} \in (0, \infty)$.
(C2′) $M_1 = \|\mathbf{X}^{\mathrm{T}}\boldsymbol{\omega}^*\|_\infty \in (0, \infty)$.
(C3′) The random errors ε_i are i.i.d. mean zero sub-Gaussian random variables.
(C4′) $\bar{\kappa} = \kappa(3\breve{M}) \in (0, \infty)$, where $\kappa(M) = \inf_{\boldsymbol{\delta} \in \mathscr{C}_M} \boldsymbol{\delta}^{\mathrm{T}}[\mathbf{I}_2 \otimes (n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})]\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2$.
(C5′) $\bar{\varrho} = \varrho(3\breve{M}) \in (0, \infty)$, where $\varrho(M) = \inf_{\boldsymbol{\delta} \in \mathscr{C}_M} \frac{\boldsymbol{\delta}^{\mathrm{T}}[\mathbf{I}_2 \otimes (n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})]\boldsymbol{\delta}}{\|\boldsymbol{\delta}_{A_0}\|_1 \|\boldsymbol{\delta}\|_\infty}$.

THEOREM 3. *In model* (4.1), *suppose the true parameter vectors $\boldsymbol{\gamma}^*$ and $\boldsymbol{\omega}^*$ are respectively s_1-sparse and s_2-sparse and assume conditions* (C1′)–(C3′) *hold. Let $\hat{\boldsymbol{\gamma}}^{\text{lasso}}$ and $\hat{\boldsymbol{\varphi}}^{\text{lasso}}$ be any optimal solutions to the L_1-penalized COSALES estimation problem* (4.4). *Then with probability at least $1 - \pi_1^{\text{ALS}}$,*
$$\left\| \begin{pmatrix} \hat{\boldsymbol{\gamma}}^{\text{lasso}} \\ \hat{\boldsymbol{\varphi}}^{\text{lasso}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\gamma}^* \\ \boldsymbol{\varphi}^* \end{pmatrix} \right\|_2 \leq 3(s_1 + s_2)^{1/2} (\lambda_1^{\text{lasso}} \vee \lambda_2^{\text{lasso}}) (2\bar{\kappa} c_0)^{-1}$$

if condition (C4′) *holds and*

$$\left\|\begin{pmatrix}\hat{\boldsymbol{\gamma}}^{\text{lasso}}\\ \hat{\boldsymbol{\varphi}}^{\text{lasso}}\end{pmatrix}-\begin{pmatrix}\boldsymbol{\gamma}^*\\ \boldsymbol{\varphi}^*\end{pmatrix}\right\|_\infty \le 3\big(\lambda_1^{\text{lasso}}\vee\lambda_2^{\text{lasso}}\big)(2\bar{\varrho}c_0)^{-1}$$

if condition (C5′) *holds*, *where*

$$\pi_1^{\text{ALS}}=2p\exp\Big(-\frac{Cn(\lambda_1^{\text{lasso}})^2}{4M_0^2M_1^2(K_1+K_2)^2}\Big)+2p\exp\Big(-\frac{Cn(\lambda_2^{\text{lasso}})^2}{4M_0^2M_1^2K_2^2}\Big),$$

$c_0=2^{-1}[(1+4\underline{c})-(1+16\underline{c}^2)^{1/2}]$, $K_1=\|\varepsilon_i\|_{\text{SG}}$, $K_2=\|\Psi_\tau'(\varepsilon_i-e_\tau)\|_{\text{SG}}$ *and* $C>0$ *is an absolute constant*.

4.2.2. *Nonconvex penalized COSALES regression*. We show that the oracle estimators $\hat{\boldsymbol{\gamma}}^{\text{oracle}}$ and $\hat{\boldsymbol{\varphi}}^{\text{oracle}}$ can be achieved with overwhelming probability by Algorithm 4 under rather general conditions. Indeed, suppose the minimal signal strength of $\boldsymbol{\gamma}^*$ and $\boldsymbol{\omega}^*$ satisfies

(A0′) $\min_{j\in A_1}|\gamma_j^*|>(a+1)\lambda_1$ and $\min_{j\in A_2}|\omega_j^*|>(a+1)|e_\tau|^{-1}\lambda_2$.

THEOREM 4. *Suppose in model* (4.1) $\boldsymbol{\gamma}^*$ *and* $\boldsymbol{\omega}^*$ *are respectively* s_1*-sparse and* s_2*-sparse and satisfy assumption* (A0′). *Take* $\hat{\boldsymbol{\gamma}}^{\text{lasso}}$ *and* $\hat{\boldsymbol{\varphi}}^{\text{lasso}}$ *as the initial values and assume conditions* (C1′)–(C3′) *hold. Take* $\lambda\ge 3s^{1/2}(\lambda_1^{\text{lasso}}\vee\lambda_2^{\text{lasso}})(2a_0c_0\bar{\kappa})^{-1}$ *when* (C4′) *holds, or take* $\lambda\ge 3(\lambda_1^{\text{lasso}}\vee\lambda_2^{\text{lasso}})(2a_0c_0\bar{\varrho})^{-1}$ *when* (C5′) *holds, or take* $\lambda\ge 3(\lambda_1^{\text{lasso}}\vee\lambda_2^{\text{lasso}})(2a_0c_0)^{-1}[(s^{1/2}\bar{\kappa}^{-1})\wedge\bar{\varrho}^{-1}]$ *when both* (C4′) *and* (C5′) *hold. The LLA algorithm* (*Algorithm* 4) *converges to the oracle estimators* $\hat{\boldsymbol{\gamma}}^{\text{oracle}}$ *and* $\hat{\boldsymbol{\varphi}}^{\text{oracle}}$ *in two iterations with probability at least* $1-\pi_1^{\text{ALS}}-\pi_2^{\text{ALS}}-\pi_3^{\text{ALS}}$, *where* π_1^{ALS} *is given in Theorem* 3,

$$\begin{aligned}\pi_2^{\text{ALS}}&=\Gamma\big(2^{-1}Q_2\lambda;n,s_1,K_1+K_2,M_0M_1,M_1^2\rho_{1\bullet\max},\nu_1\big)\\&\quad+\Gamma\big(2^{-1}Q_2\lambda;n,s_2,K_2,M_0M_1,M_1^2\rho_{2\bullet\max},\nu_2\big)\\&\quad+2(p-s_1)\exp\Big(-\frac{Ca_1^2n\lambda^2}{4M_0^2M_1^2(K_1+K_2)^2}\Big)\\&\quad+2(p-s_2)\exp\Big(-\frac{Ca_1^2n\lambda^2}{4M_0^2M_1^2K_2^2}\Big),\end{aligned}$$

and

$$\begin{aligned}pi_3^{\text{ALS}}&=\Gamma\big(2^{-1}c_0\phi_{\min}\bar{R};n,s_1,K_1+K_2,M_0M_1,M_1^2\rho_{1\bullet\max},\nu_1\big)\\&\quad+\Gamma\big(2^{-1}c_0\phi_{\min}\bar{R};n,s_2,K_2,M_0M_1,M_1^2\rho_{2\bullet\max},\nu_2\big),\end{aligned}$$

where $s=s_1+s_2$, $\lambda=\lambda_1\wedge\lambda_2$, $Q_2=a_1c_0\phi_{\min}[2(1+2\bar{c})M_0\phi_{\max}^{1/2}]^{-1}$, $\nu_1=\text{var}(\varepsilon_i+\Psi_\tau'(\varepsilon_i-e_\tau))$, $\nu_2=\text{var}(\Psi_\tau'(\varepsilon_i-e_\tau))$, $\bar{R}=(\min_{j\in A_1}|\gamma_j^*|-a\lambda_1)\wedge(\min_{j\in A_2}|\varphi_j^*|-a\lambda_2)$, $C>0$ *is an absolute constant*, c_0, K_1, K_2 *are given in Theorem* 3, *and* $\Gamma(\cdot)$ *is given in Theorem* 2.

For SCAD and MCP penalized COSALES regressions, the LLA algorithm (Algorithm 4) starting from the zero vector can also be used as long as we can take $\lambda_k = \lambda_k^{\text{lasso}}$, $k = 1, 2$.

COROLLARY 2. *Assume the same framework of Theorem 4 and suppose the SCAD penalty (2.8) or MCP (2.9) is used. If condition* (C4′) *holds and* $2a_0c_0\bar{\kappa} \geq 3\check{M}s^{1/2}$, *or if condition* (C5′) *holds and* $2a_0c_0\bar{\varrho} \geq 3\check{M}$, *or if both* (C4′) *and* (C5′) *hold and* $3\check{M}[(s^{1/2}\bar{\kappa}^{-1}) \wedge \bar{\varrho}^{-1}] \leq 2a_0c_0$, *then the LLA algorithm (Algorithm 4) initialized by zero converges to the oracle estimators* $\hat{\boldsymbol{\gamma}}^{\text{oracle}}$ *and* $\hat{\boldsymbol{\varphi}}^{\text{oracle}}$ *after three iterations with probability at least* $1 - \check{\pi}_1^{\text{ALS}} - \pi_2^{\text{ALS}} - \pi_3^{\text{ALS}}$, *where*

$$\check{\pi}_1^{\text{ALS}} = 2p\exp\Big(-\frac{Cn\lambda_1^2}{4M_0^2M_1^2(K_1+K_2)^2}\Big) + 2p\exp\Big(-\frac{Cn\lambda_2^2}{4M_0^2M_1^2K_2^2}\Big),$$

π_2^{ALS} *and* π_3^{ALS} *are given in Theorem 4, and* $s = s_1 + s_2$.

REMARK 2. We can easily modify (4.2) to allow certain subsets of coefficients not to be penalized. Let $\mathscr{R}_1$ and $\mathscr{R}_2$ be the index sets of unpenalized components of $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$, respectively. Then (4.2) can be modified as

$$\min_{\boldsymbol{\gamma},\boldsymbol{\varphi}\in\mathbb{R}^p} S_n(\boldsymbol{\gamma},\boldsymbol{\varphi}) + \sum_{j\in\mathscr{R}_1^c} p_{\lambda_1}(\gamma_j) + \sum_{j\in\mathscr{R}_2^c} p_{\lambda_2}(\varphi_j).$$

The COSALES algorithm can be readily used to solve the above problem. Moreover, similar theoretical results can be established with slight modifications.

4.3. *Simulation examples.* We demonstrate the selection and estimation accuracy of the COSALES regression through two numerical simulations. For the nonconvex penalties used in both simulations, we fix $\gamma = 3.7$ for the SCAD penalty and $\gamma = 2$ for the MCP.

EXAMPLE 2. We consider the same model (3.1) that was used in Example 1, but different from the approach used there, we estimate the coefficients through the nonconvex penalized COSALES regression (4.2). Again we choose $p = 600$ and independently simulate a training set of size $n = 300$ for fitting and a validation set of size $n = 300$ for tuning. The tuning parameter is selected by minimizing the validation error $\sum_{i\in\text{validation}}\{\Psi_{0.5}(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\gamma}}) + \Psi_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\gamma}} - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\varphi}})\}$ for the computed estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\varphi}}$. We pick a fairly extreme τ-value ($\tau = 0.95$) for easy separation of the conditional mean and scale functions. Both the COSALES Lasso and two variations of the LLA algorithm for each of the SCAD and MCP penalized COSALES regressions are implemented.

Based on 100 independent runs, the following measurements are calculated to evaluate the sparsity recovery and estimation performance of the COSALES estimators:

TABLE 2

Numerical summary of simulation results from the Lasso, SCAD and MCP penalized COSALES regression for model (3.1) $y = x_6 + x_{12} + x_{15} + x_{20} + (0.7x_1)\varepsilon$. The selection accuracy is measured by the number of selected variables $|\hat{A}_1|$ and $|\hat{A}_2|$, and the proportions p_{a_1} and p_{a_2} of covering the true active sets. The estimation accuracy is measured by the L_1 risks $R_1^{\boldsymbol{\gamma}}$ and $R_1^{\boldsymbol{\varphi}}$, and the L_2 risks $R_2^{\boldsymbol{\gamma}}$ and $R_2^{\boldsymbol{\varphi}}$. The results are shown as averages over 100 replicates with standard errors listed in the parentheses. A fairly extreme τ-value ($\tau = 0.95$) is used in the simulation for easy separation of the mean and scale

Method	$\lvert\hat{A}_1\rvert$	$\lvert\hat{A}_2\rvert$	p_{a_1}	p_{a_2}	$R_1^{\boldsymbol{\gamma}}$	$R_1^{\boldsymbol{\varphi}}$	$R_2^{\boldsymbol{\gamma}}$	$R_2^{\boldsymbol{\varphi}}$
COSALES-Lasso	26.88	13.36	100%	100%	0.407	0.378	0.124	0.294
	(1.04)	(0.45)	(0)	(0)	(0.012)	(0.008)	(0.002)	(0.006)
COSALES-SCAD*	7.24	1.01	100%	100%	0.095	0.072	0.048	0.072
	(0.10)	(0.01)	(0)	(0)	(0.004)	(0.005)	(0.002)	(0.005)
COSALES-SCAD^0	8.85	1.01	100%	100%	0.107	0.065	0.049	0.065
	(0.57)	(0.01)	(0)	(0)	(0.005)	(0.005)	(0.002)	(0.005)
COSALES-MCP*	6.46	1.01	100%	100%	0.089	0.070	0.045	0.070
	(0.38)	(0.01)	(0)	(0)	(0.004)	(0.005)	(0.002)	(0.005)
COSALES-MCP^0	7.08	1.01	100%	100%	0.102	0.067	0.052	0.067
	(0.44)	(0.01)	(0)	(0)	(0.006)	(0.005)	(0.003)	(0.005)

$|\hat{A}_1|, |\hat{A}_2|$: the average size of the active sets for $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\varphi}}$, respectively, $\hat{A}_1 = \{j : \hat{\gamma}_j \neq 0\}$ and $\hat{A}_2 = \{j : \hat{\varphi}_j \neq 0\}$.

p_{a_1}, p_{a_2}: proportions of the events $A_1 \subset \hat{A}_1$ and $A_2 \subset \hat{A}_2$, respectively, where $A_1 = \{6, 12, 15, 20\}$ denotes the active set of $\boldsymbol{\gamma}^*$ and $A_2 = \{1\}$ denotes the active set of $\boldsymbol{\varphi}^*$.

$R_1^{\boldsymbol{\gamma}}, R_1^{\boldsymbol{\varphi}}$: the average L_1 risks, $R_1^{\boldsymbol{\gamma}} = \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1$ and $R_1^{\boldsymbol{\varphi}} = \|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}^*\|_1$.

$R_2^{\boldsymbol{\gamma}}, R_2^{\boldsymbol{\varphi}}$: the average L_2 risks, $R_2^{\boldsymbol{\gamma}} = \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2$ and $R_2^{\boldsymbol{\varphi}} = \|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}^*\|_2$.

The results are summarized in Table 2, from which we can draw the following conclusions:

(1) The COSALES regression (with Lasso or nonconvex penalties) can recover the sparse patterns in both the mean and scale functions with overwhelming probabilities.

(2) The COSALES Lasso tends to select a lot more irrelevant covariates and has much larger estimation errors than the nonconvex penalized COSALES regression (with the SCAD penalty or MCP).

(3) The three-step LLA algorithm starting from zero produces similar results to the two-step LLA algorithm starting from the Lasso solution.

EXAMPLE 3. In this example, we simulate data from the following normal linear heteroscedastic model:

$$y = x_6 + x_{12} + x_{15} + x_{20} + (0.7x_1 + 0.7x_{12})\varepsilon, \tag{4.6}$$

where the covariates are simulated by setting $x_1 = \Phi(z_1), x_{12} = \Phi(z_{12})$, and $x_j = z_j$, $j \neq 1, 12$, where $(z_1, \ldots, z_p)^{\mathrm{T}} \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (0.5^{|i-j|})_{p\times p}$, and $\Phi(\cdot)$ is the CDF of the standard normal distribution. The random error $\varepsilon \sim N(0, 1)$. Note that in model (3.1), the active sets of the true parameter vectors do not overlap, so the SALES regression can detect active variables in the scale. However, in model (4.6) the active set for the mean, $A_1 = \{6, 12, 15, 20\}$, overlaps with the active set for the scale, $A_2 = \{1, 12\}$. Thus, the SALES regression cannot recover the variable x_{12} in the scale function. We show by this Monte Carlo simulation that the COSALES regression can recover the sparse patterns in both the mean and scale functions. We fix $p = 600$ and independently simulate a training set of size $n = 500$ for fitting and a validation set of the same size for tuning. We select the regularization parameter by minimizing the validation error $\sum_{i\in\text{validation}}\{\Psi_{0.5}(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\gamma}}) + \Psi_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\gamma}} - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\varphi}})\}$ for the computed estimate $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\varphi}}$. In order to separate the mean and scale easily, we again pick $\tau = 0.95$. We implement the COSALES Lasso and two variations of the LLA algorithm as were done in Examples 2 for each of the SCAD and MCP penalized COSALES regressions.

Based on 100 independent runs, the same measurements of performance as in Example 2 are calculated to evaluate the sparsity recovery and estimation accuracy of the COSALES estimation. The results are summarized in Table 3. The same conclusions in Example 2 can be drawn here.

5. Real data example. We apply the SALES and COSALES regressions to a real data set reported in Scheetz et al. (2006). The data set consists of gene expression levels of more than 31,000 probes obtained from 120 rats. The expressions are analyzed on a logarithmic scale (base 2). As was done in Scheetz et al. (2006), we exclude the probes that were not expressed in the eye or that lacked sufficient variation. Among those 18,976 probes left, we study how the expressions of other genes are associated with the gene *TRIM32* (probe 1389163_at). This gene was found to be associated with Bardet–Biedl syndrome, which is a disorder that affects many parts of the body including the retina. For all the other genes, we first standardize them and select the 3000 probes with the largest variances. These 3000 probes are then ranked according to the magnitude of the correlations between their expressions and that of probe 1389163_at. We choose the top 300 probes with the largest correlations in magnitude for the analysis.

The third column of Table 4 lists the number of active variables selected by the SALES regressions with Lasso, SCAD and MCP penalties, fitted on the whole

TABLE 3

Numerical summary of simulation results from the Lasso, SCAD and MCP penalized COSALES regression for model (4.6): $y = x_6 + x_{12} + x_{15} + x_{20} + (0.7x_1 + 0.7x_{12})\varepsilon$. The selection accuracy is measured by the number of selected variables $|\hat{A}_1|$ and $|\hat{A}_2|$, and the proportions p_{a_1} and p_{a_2} of covering the true active sets. The estimation accuracy is measured by the L_1 risks R_1^{γ} and R_1^{φ}, and the L_2 risks R_2^{γ} and R_2^{φ}. The results are shown as averages over 100 replicates with standard errors listed in the parentheses. A fairly extreme τ-value ($\tau = 0.95$) is used in the simulation for easy separation of the mean and scale

Method	$\lvert\hat{A}_1\rvert$	$\lvert\hat{A}_2\rvert$	p_{a_1}	p_{a_2}	R_1^{γ}	R_1^{φ}	R_2^{γ}	R_2^{φ}
COSALES-Lasso	27.92	12.67	100%	100%	0.719	0.450	0.249	0.282
	(0.98)	(0.49)	(0)	(0)	(0.018)	(0.011)	(0.006)	(0.008)
COSALES-SCAD*	6.80	2.06	100%	100%	0.167	0.210	0.089	0.161
	(0.52)	(0.04)	(0)	(0)	(0.008)	(0.014)	(0.004)	(0.010)
COSALES-SCAD^0	5.70	2.02	100%	100%	0.157	0.199	0.090	0.148
	(0.25)	(0.01)	(0)	(0)	(0.006)	(0.013)	(0.003)	(0.009)
COSALES-MCP*	5.95	2.06	100%	100%	0.153	0.221	0.086	0.165
	(0.35)	(0.03)	(0)	(0)	(0.006)	(0.015)	(0.003)	(0.010)
COSALES-MCP^0	6.00	2.04	100%	100%	0.180	0.205	0.098	0.154
	(0.36)	(0.02)	(0)	(0)	(0.009)	(0.014)	(0.004)	(0.010)

data set of 120 subjects. For both SCAD and MCP penalized SALES regressions, the two variations of the LLA algorithm were used. The tuning parameter for each method is selected by five-fold cross-validation. The last two columns of Table 4 summarize the results from 50 random partitions. Each partition randomly splits the data into a training set with 80 observations and a validation set with 40 observations. We fit the model with the training set using five-fold cross-validation for tuning and calculate the predicted loss $(1/40)\sum_{i\in\text{validation}} \Psi_\tau(y_i - \hat{\beta}_0 - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})$ based on the validation set. The average number of active variables selected and the average predicted loss are calculated from the 50 partitions with their respective standard errors listed in the parentheses. Table 4 reveals two interesting findings. First, the nonconvex penalized SALES regression selects less variables than the SALES Lasso, but there is no obvious improvement of the nonconvex penalized SALES regression over the SALES Lasso in terms of predicted loss. Second, for all SALES regressions, the number of variables selected is different at different values of τ (0.3, 0.5 and 0.7). This is an indication of heteroscedasticity in the data.

To further explore the heterogeneous scale, we also apply the COSALES regression to the data. The results are summarized in Table 5. Columns 2 and 3 display the number of variables selected for the mean ($|\hat{A}_1|$) and scale ($|\hat{A}_2|$), and the number of variables that overlap ($|\hat{A}_1 \cap \hat{A}_2|$) for each method. For all penalties, τ is set to be 0.7 in the COSALES regression. Random partitions are done in the same way as the SALES regression and the predicted loss is calculated via

TABLE 4

Analysis of microarray data using SALES regressions with Lasso, SCAD and MCP penalties. Three different values of τ (0.3, 0.5 *and* 0.7) *are used for each method. The number of active variables selected using the whole data set is given in column* 3. *The average number of active variables selected and average predicted loss* $(1/40)\sum_{i\in\text{validation}}\Psi_\tau(y_i-\hat\beta_0-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol\beta})$ *listed in columns* 4 *and* 5 *are calculated from* 50 *random partitions of the original data with standard errors listed in parentheses*

Method	τ	All data	Random partition	
		$\lvert\hat{A}\rvert$	$\lvert\hat{A}\rvert$	Predicted loss
SALES-Lasso	0.3	22	22.00 (1.51)	0.007 (0.00055)
	0.5	25	25.38 (1.94)	0.005 (0.00036)
	0.7	20	21.90 (1.66)	0.005 (0.00022)
SALES-SCAD*	0.3	19	16.02 (2.09)	0.006 (0.00048)
	0.5	13	15.52 (1.80)	0.006 (0.00043)
	0.7	11	13.54 (1.98)	0.005 (0.00037)
SALES-SCAD^0	0.3	16	16.60 (2.03)	0.006 (0.00054)
	0.5	17	17.22 (2.36)	0.007 (0.00048)
	0.7	14	14.82 (2.18)	0.005 (0.00030)
SALES-MCP*	0.3	14	15.82 (2.56)	0.006 (0.00053)
	0.5	12	12.66 (2.58)	0.008 (0.00054)
	0.7	10	9.66 (1.78)	0.006 (0.00035)
SALES-MCP^0	0.3	11	11.74 (1.47)	0.006 (0.00057)
	0.5	13	13.24 (2.75)	0.007 (0.00058)
	0.7	13	14.18 (3.36)	0.006 (0.00034)

$(1/40)\sum_{i\in\text{validation}}\Psi_{0.5}(y_i-\hat\gamma_0-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol\gamma})+\Psi_\tau(y_i-\hat\gamma_0-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol\gamma}-\hat\varphi_0-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol\varphi})$. The results for the random partitions are shown in columns 4 to 6. It can be seen that the COSALES regression reveals more information about the heterogeneous scale which cannot be otherwise detected in the SALES regression or the sparse quantile regression [Wang, Wu and Li (2012)] due to overlaps.

6. Proofs. In this section, we give the proofs of the main theoretical results stated in previous sections. First of all, let us state two lemmas on the properties of the asymmetric squared error loss $\Psi_\tau(\cdot)$ given in (2.1). These properties play an important role in the proofs of many results to be presented below. Let $w_\tau(u)=|\tau-I(u<0)|$ and recall that $\underline{c}=\tau\wedge(1-\tau)$ and $\bar c=\tau\vee(1-\tau)$.

LEMMA 1. *The asymmetric squared error loss* $\Psi_\tau(\cdot)$ *is continuously differentiable, but is not twice differentiable at zero when* $\tau\neq0.5$. *Moreover, for any* $u,u_0\in\mathbb{R}$ *and* $\tau\in(0,1)$, *we have*

$$\underline{c}(u-u_0)^2\le\Psi_\tau(u)-\Psi_\tau(u_0)-\Psi'_\tau(u_0)(u-u_0)\le\bar c(u-u_0)^2.$$

It follows that $\Psi_\tau(\cdot)$ *is strongly convex.*

TABLE 5

Analysis of microarray data using COSALES regressions with Lasso, SCAD and MCP penalties. In this analysis, $\tau = 0.7$ is used. The number of active variables selected for the mean and scale using the whole data set is given in columns 2 and 3. The average number of active variables selected for the mean and scale and average predicted loss $(1/40)\sum_{i\in\text{validation}} \Psi_{0.5}(y_i - \hat{\gamma}_0 - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\gamma}}) + \Psi_\tau(y_i - \hat{\gamma}_0 - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\gamma}} - \hat{\varphi}_0 - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\varphi}})$ listed in columns 4 to 6 are calculated from 50 random partitions of the original data with standard errors listed in parentheses

	All data			**Random partition**			
Method	$\lvert\hat{A}_1\rvert$	$\lvert\hat{A}_2\rvert$	$\lvert\hat{A}_1 \cap \hat{A}_2\rvert$	$\lvert\hat{A}_1\rvert$	$\lvert\hat{A}_2\rvert$	$\lvert\hat{A}_1 \cap \hat{A}_2\rvert$	**Predicted loss**
COSALES-Lasso	22	10	9	22.62 (1.21)	9.80 (1.10)	7.86 (0.93)	0.010 (0.00056)
COSALES-SCAD*	19	7	6	18.92 (0.79)	5.58 (0.50)	3.90 (0.31)	0.011 (0.00067)
COSALES-SCAD^0	20	5	4	20.22 (0.98)	5.82 (0.64)	3.92 (0.44)	0.011 (0.00072)
COSALES-MCP*	10	3	1	10.96 (2.32)	3.08 (1.40)	1.38 (0.74)	0.014 (0.00096)
COSALES-MCP^0	10	4	3	12.94 (1.83)	4.56 (1.04)	1.46 (0.42)	0.012 (0.00083)

LEMMA 2. *For any $u, u_0 \in \mathbb{R}$ and $\tau \in (0, 1)$, we have*

$$2\underline{c}|u - u_0| \le \left|\Psi'_\tau(u) - \Psi'_\tau(u_0)\right| \le 2\bar{c}|u - u_0|.$$

It follows immediately that $\Psi'_\tau(\cdot)$ is Lipschitz continuous.

PROOF OF LEMMA 1. It is easy to see that $\underline{c} \le w_\tau(u) \le \bar{c}$ for any $u \in \mathbb{R}$. Note that $\Psi'_\tau(u) = 2w_\tau(u)u$, which is continuous and which is not differentiable at $u = 0$ when $\tau \ne 0.5$. To show the inequalities, consider the following situations. If $w_\tau(u) \ge w_\tau(u_0)$, it follows that

$$\begin{aligned}
&\Psi_\tau(u) - \Psi_\tau(u_0) - \Psi'_\tau(u_0)(u - u_0)\\
&\quad = w_\tau(u)u^2 - w_\tau(u_0)u_0^2 - 2w_\tau(u_0)u_0(u - u_0)\\
&\quad = w_\tau(u_0)(u - u_0)^2 + \{w_\tau(u) - w_\tau(u_0)\}u^2\\
&\quad \ge w_\tau(u_0)(u - u_0)^2 \ge \underline{c}(u - u_0)^2.
\end{aligned}$$

Otherwise, if $w_\tau(u) < w_\tau(u_0)$, then we know that $\underline{c} = w_\tau(u)$, $\bar{c} = w_\tau(u_0)$ and $u_0 u \le 0$. It follows that

$$\begin{aligned}
&\Psi_\tau(u) - \Psi_\tau(u_0) - \Psi'_\tau(u_0)(u - u_0)\\
&\quad = \underline{c}u^2 - \bar{c}u_0^2 - 2\bar{c}u_0(u - u_0)\\
&\quad \ge \underline{c}u^2 - 2\underline{c}u_0 u + \underline{c}u_0^2 = \underline{c}(u - u_0)^2.
\end{aligned}$$

Therefore, the first inequality holds. Similarly, we can show the second inequality. □

PROOF OF LEMMA 2. If $u = 0$ or $u_0 = 0$, then the inequalities hold trivially. If $uu_0 > 0$, we know that $w_\tau(u) = w_\tau(u_0)$. It follows that

$$2\underline{c}|u - u_0| \le \left|\Psi'_\tau(u) - \Psi'_\tau(u_0)\right| = 2w_\tau(u)|u - u_0| \le 2\bar{c}|u - u_0|.$$

If instead, $uu_0 < 0$, there are two cases: $u > 0, u_0 < 0$ or $u < 0, u_0 > 0$. For the first case, we have

$$2\underline{c}|u - u_0| \le \left|\Psi'_\tau(u) - \Psi'_\tau(u_0)\right| = 2\tau u - 2(1-\tau)u_0 \le 2\bar{c}|u - u_0|.$$

For the second case, we have

$$2\underline{c}|u - u_0| \le \left|\Psi'_\tau(u) - \Psi'_\tau(u_0)\right| = -2(1-\tau)u + 2\tau u_0 \le 2\bar{c}|u - u_0|.$$

This completes the proof. □

The following lemma deals with sub-Gaussian random variables.

LEMMA 3. *Suppose that $Z, Z_1, \ldots, Z_n \in \mathbb{R}$ are i.i.d. sub-Gaussian random variables. Let $\mathbf{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$, $K = \|Z\|_{\mathrm{SG}}$, $Z^+ = \max(Z, 0)$ and $Z^- = \max(-Z, 0)$.*

(1) *If $\mathbb{E}(Z) = 0$, then there exists an absolute constant $C > 0$ such that for any $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}} \in \mathbb{R}^n$ and any $t \ge 0$,*

$$\mathrm{P}\left(|\mathbf{a}^{\mathrm{T}}\mathbf{Z}| \ge t\right) \le 2\exp\left(-\frac{Ct^2}{K^2\|\mathbf{a}\|_2^2}\right).$$

(2) *Let $\mathbf{A}$ be a fixed $m \times n$ matrix. If $\mathbb{E}(Z) = 0$ and $\mathrm{var}(Z) = 1$, then there exists an absolute constant $C > 0$ such that for any $t \ge 0$,*

$$\mathrm{P}\left(\left|\|\mathbf{AZ}\|_2 - \|\mathbf{A}\|_F\right| \ge t\right) \le 2\exp\left(-\frac{Ct^2}{K^4\|\mathbf{A}\|_2^2}\right),$$

where $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_2$ represent the Frobenius and L_2 norms of matrix $\mathbf{A}$ respectively.

(3) *Let $\mathbf{A}$ be a fixed $m \times n$ matrix. Let $\mathbf{e}_j \in \mathbb{R}^m$ be the unit vector with its jth component one, $j = 1, \ldots, m$. Suppose $M \equiv \max_{1 \le j \le m} n^{-1/2}\|\mathbf{A}^{\mathrm{T}}\mathbf{e}_j\|_2 \in (0, \infty)$ and $\rho \equiv \lambda_{\max}(n^{-1}\mathbf{AA}^{\mathrm{T}}) \in (0, \infty)$. If $\mathbb{E}(Z) = 0$ and $\nu = \mathrm{var}(Z) \in (0, \infty)$, then there exists an absolute constant $C > 0$ such that for any $t \ge 0$,*

$$\begin{aligned}
&\mathrm{P}\left(\left\|n^{-1}\mathbf{AZ}\right\|_2 \ge t\right) \\
&\quad \le \Gamma(t; n, m, K, M, \rho, \nu) \\
&\quad = 2m\exp\left(-\frac{Cnt^2}{K^2M^2m}\right) \wedge 2\exp\left(-\frac{C\nu^2[(n^{1/2}t - \nu m^{1/2}\rho^{1/2})^+]^2}{K^4\rho}\right).
\end{aligned}$$

(4) *The random variables Z^+ and Z^- are also sub-Gaussian. Moreover, for any $c_1, c_2 \in \mathbb{R}$, $c_1 Z^+ + c_2 Z^-$ is sub-Gaussian.*

PROOF. (1) This part follows directly from Proposition 5.10 of Vershynin (2010).

(2) This part follows from Theorem 2.1 of Rudelson and Vershynin (2013).

(3) On one hand, we have by part (1) that

$$\mathrm{P}\Big(\Big\|\frac{\mathbf{A}}{n}\mathbf{Z}\Big\|_2 \ge t\Big) \le \mathrm{P}\Big(\Big\|\frac{\mathbf{A}}{\sqrt{n}}\mathbf{Z}\Big\|_\infty \ge \frac{t\sqrt{n}}{\sqrt{m}}\Big) \le 2m\exp\Big(-\frac{Cnt^2}{K^2M^2m}\Big).$$

One the other hand, note that $\|n^{-1/2}\mathbf{A}\|_F = \sqrt{\mathrm{Tr}(\mathbf{A}\mathbf{A}^{\mathrm{T}}/n)} \le \sqrt{m\rho}$ and $\|n^{-1/2} \times \mathbf{A}\|_2^2 = \lambda_{\max}(\mathbf{A}^{\mathrm{T}}\mathbf{A}/n) = \lambda_{\max}(\mathbf{A}\mathbf{A}^{\mathrm{T}}/n) = \rho$. We have by part (2) that

$$\begin{aligned}
\mathrm{P}\Big(\Big\|\frac{\mathbf{A}}{n}\mathbf{Z}\Big\|_2 \ge t\Big) &\le \mathrm{P}\Big(\Big\|\frac{\mathbf{A}}{\sqrt{n}}\frac{\mathbf{Z}}{\nu}\Big\|_2 - \Big\|\frac{\mathbf{A}}{\sqrt{n}}\Big\|_F \ge \frac{t\sqrt{n}}{\nu} - \sqrt{m\rho}\Big) \\
&\le \mathrm{P}\Big(\Big|\Big\|\frac{\mathbf{A}}{\sqrt{n}}\frac{\mathbf{Z}}{\nu}\Big\|_2 - \Big\|\frac{\mathbf{A}}{\sqrt{n}}\Big\|_F\Big| \ge \Big(\frac{t\sqrt{n}}{\nu} - \sqrt{m\rho}\Big)^+\Big) \\
&\le 2\exp\Big(-\frac{C\nu^2[(n^{1/2}t - \nu m^{1/2}\rho^{1/2})^+]^2}{K^4\rho}\Big).
\end{aligned}$$

(4) Note that by definition, we have $K \in (0, \infty)$ and $(\mathbb{E}|Z|^p)^{1/p} \le K\sqrt{p}$, $\forall p \ge 1$. It follows immediately that $(\mathbb{E}|Z^+|^p)^{1/p} \le (\mathbb{E}|Z|^p)^{1/p} \le K\sqrt{p}$ and $(\mathbb{E}|Z^-|^p)^{1/p} \le (\mathbb{E}|Z|^p)^{1/p} \le K\sqrt{p}$, $\forall p \ge 1$. Now by Lemma 5.5 of Vershynin (2010), we conclude that Z^+ and Z^- are both sub-Gaussian. For any $c_1, c_2 \in \mathbb{R}$, by Minkowski inequality,

$$\begin{aligned}
\big(\mathbb{E}|c_1Z^+ + c_2Z^-|^p\big)^{1/p} &\le |c_1|\big(\mathbb{E}|Z^+|^p\big)^{1/p} + |c_2|\big(\mathbb{E}|Z^-|^p\big)^{1/p} \\
&\le \big(|c_1| + |c_2|\big)K\sqrt{p}, \qquad \forall p \ge 1.
\end{aligned}$$

By Lemma 5.5 of Vershynin (2010) again, we can see that $c_1Z^+ + c_2Z^-$ is also sub-Gaussian. This completes the proof. □

Now we are ready to prove Theorems 1 and 2. Lemmas 4 and 5 are presented to facilitate the proofs.

LEMMA 4. *Let $\boldsymbol{\zeta} = (\zeta_i, 1 \le i \le n)^{\mathrm{T}}$ with $\zeta_i = \Psi'_\tau(\varepsilon_i) = 2|\tau - I(\varepsilon_i < 0)|\varepsilon_i$.*

(1) *For any $\boldsymbol{\beta}, \boldsymbol{\delta} \in \mathbb{R}^p$, $\langle \nabla\mathcal{L}_n(\boldsymbol{\beta} + \boldsymbol{\delta}) - \nabla\mathcal{L}_n(\boldsymbol{\beta}), \boldsymbol{\delta}\rangle \ge 2\underline{c}\|\mathbf{X}\boldsymbol{\delta}\|_2^2/n$.*

(2) *For any $d > 0$, $\mathrm{P}(\|\hat{\boldsymbol{\beta}}^{\mathrm{oracle}} - \boldsymbol{\beta}^*\|_2 \ge d) \le \mathrm{P}(\|n^{-1}\mathbf{X}_A^{\mathrm{T}}\boldsymbol{\zeta}\|_2 \ge 2\underline{c}\rho_{\min}d)$.*

PROOF. The first part follows from the strong convexity of $\Psi_\tau(\cdot)$. Specifically, by Lemma 1, we have

$$\mathcal{L}_n(\boldsymbol{\beta}+\boldsymbol{\delta})-\mathcal{L}_n(\boldsymbol{\beta})-\langle\nabla\mathcal{L}_n(\boldsymbol{\beta}),\boldsymbol{\delta}\rangle\geq\underline{c}\|\mathbf{X}\boldsymbol{\delta}\|_2^2/n,$$
$$\mathcal{L}_n(\boldsymbol{\beta})-\mathcal{L}_n(\boldsymbol{\beta}+\boldsymbol{\delta})-\langle\nabla\mathcal{L}_n(\boldsymbol{\beta}+\boldsymbol{\delta}),-\boldsymbol{\delta}\rangle\geq\underline{c}\|\mathbf{X}\boldsymbol{\delta}\|_2^2/n.$$

Summing up the above two inequalities yields the desired result in part (1).

For the second part, let $\hat{\boldsymbol{\delta}}=\hat{\boldsymbol{\beta}}^{\text{oracle}}-\boldsymbol{\beta}^*$. By definition of $\hat{\boldsymbol{\beta}}^{\text{oracle}}$, we have $\hat{\boldsymbol{\delta}}_{A^c}=\mathbf{0}$ and $\nabla_A\mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})=\mathbf{0}$. Now by part (1) we have

$$\begin{aligned}2\underline{c}\rho_{\min}\|\hat{\boldsymbol{\delta}}\|_2^2&=2\underline{c}\rho_{\min}\|\hat{\boldsymbol{\delta}}_A\|_2^2\leq 2\underline{c}\hat{\boldsymbol{\delta}}_A^{\mathrm{T}}(\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A/n)\hat{\boldsymbol{\delta}}_A=2\underline{c}\|\mathbf{X}\hat{\boldsymbol{\delta}}\|_2^2/n\\&\leq\langle\nabla\mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})-\nabla\mathcal{L}_n(\boldsymbol{\beta}^*),\hat{\boldsymbol{\delta}}\rangle=\langle-\nabla_A\mathcal{L}_n(\boldsymbol{\beta}^*),\hat{\boldsymbol{\delta}}_A\rangle\\&\leq\|\nabla_A\mathcal{L}_n(\boldsymbol{\beta}^*)\|_2\|\hat{\boldsymbol{\delta}}_A\|_2=\|n^{-1}\mathbf{X}_A^{\mathrm{T}}\boldsymbol{\zeta}\|_2\|\hat{\boldsymbol{\delta}}\|_2,\end{aligned}$$

which implies that $2\underline{c}\rho_{\min}\|\hat{\boldsymbol{\beta}}^{\text{oracle}}-\boldsymbol{\beta}^*\|_2\leq\|n^{-1}\mathbf{X}_A^{\mathrm{T}}\boldsymbol{\zeta}\|_2$. The result of part (2) then follows. □

PROOF OF THEOREM 1. Let $\hat{\boldsymbol{\delta}}=\hat{\boldsymbol{\beta}}^{\text{lasso}}-\boldsymbol{\beta}^*$ and $z_\infty^*=\|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty$. Note that $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ satisfies the Karush–Kuhn–Tucker (KKT) condition

$$\nabla\mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\text{lasso}})+\mathbf{g}=\mathbf{0},$$

where $g_j=\lambda_{\text{lasso}}\operatorname{sgn}(\hat{\beta}_j^{\text{lasso}})$ if $\hat{\beta}_j^{\text{lasso}}\neq 0$ and $g_j\in[-\lambda_{\text{lasso}},\lambda_{\text{lasso}}]$ if $\hat{\beta}_j^{\text{lasso}}=0$. It follows that $\hat{\beta}_j^{\text{lasso}}g_j=\lambda_{\text{lasso}}|\hat{\beta}_j^{\text{lasso}}|$, $\forall j$. Since $\boldsymbol{\beta}_{A^c}^*=\mathbf{0}$, we have $\hat{\boldsymbol{\delta}}_{A^c}=\hat{\boldsymbol{\beta}}_{A^c}^{\text{lasso}}$. By Lemma 4 and Hölder's inequality, we get

$$\begin{aligned}0&\leq 2\underline{c}\|\mathbf{X}\hat{\boldsymbol{\delta}}\|_2^2/n\leq\langle\nabla\mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\text{lasso}})-\nabla\mathcal{L}_n(\boldsymbol{\beta}^*),\hat{\boldsymbol{\delta}}\rangle=\langle-\mathbf{g}-\nabla\mathcal{L}_n(\boldsymbol{\beta}^*),\hat{\boldsymbol{\delta}}\rangle\\&=\langle\hat{\boldsymbol{\delta}}_A,-\mathbf{g}_A-\nabla_A\mathcal{L}_n(\boldsymbol{\beta}^*)\rangle+\langle\hat{\boldsymbol{\beta}}_{A^c}^{\text{lasso}},-\mathbf{g}_{A^c}-\nabla_{A^c}\mathcal{L}_n(\boldsymbol{\beta}^*)\rangle\qquad(6.1)\\&\leq(z_\infty^*+\lambda_{\text{lasso}})\|\hat{\boldsymbol{\delta}}_A\|_1+(z_\infty^*-\lambda_{\text{lasso}})\|\hat{\boldsymbol{\delta}}_{A^c}\|_1.\end{aligned}$$

Under the event $\mathcal{E}=\{z_\infty^*\leq 2^{-1}\lambda_{\text{lasso}}\}$, from (6.1) we get

$$\|\hat{\boldsymbol{\delta}}_{A^c}\|_1\leq\frac{z_\infty^*+\lambda_{\text{lasso}}}{z_\infty^*-\lambda_{\text{lasso}}}\|\hat{\boldsymbol{\delta}}_A\|_1\leq 3\|\hat{\boldsymbol{\delta}}_A\|_1,$$

which implies that $\hat{\boldsymbol{\delta}}\in\mathscr{C}$. Now under $\mathcal{E}$, by condition (C3), it follows from (6.1) that

$$2\underline{c}\kappa\|\hat{\boldsymbol{\delta}}\|_2^2\leq(3/2)\lambda_{\text{lasso}}\|\hat{\boldsymbol{\delta}}_A\|_1\leq(3/2)\lambda_{\text{lasso}}s^{1/2}\|\hat{\boldsymbol{\delta}}_A\|_2\leq(3/2)\lambda_{\text{lasso}}s^{1/2}\|\hat{\boldsymbol{\delta}}\|_2,$$

and similarly by condition (C4) and (6.1) we get

$$2\underline{c}\varrho\|\hat{\boldsymbol{\delta}}\|_\infty\leq 2\underline{c}\|\mathbf{X}\hat{\boldsymbol{\delta}}\|_2^2/(n\|\hat{\boldsymbol{\delta}}_A\|_1)\leq(3/2)\lambda_{\text{lasso}}.$$

Thus, we have

$$\begin{aligned}&\mathrm{P}\big(\|\hat{\boldsymbol{\delta}}\|_2 \le 3s^{1/2}\lambda_{\mathrm{lasso}}(4\kappa\underline{c})^{-1} \cap \|\hat{\boldsymbol{\delta}}\|_\infty \le 3\lambda_{\mathrm{lasso}}(4\varrho\underline{c})^{-1}\big)\\ &\quad\ge \mathrm{P}\big(z^*_\infty \le 2^{-1}\lambda_{\mathrm{lasso}}\big) \ge 1 - \mathrm{P}\big(\big\|n^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\zeta}\big\|_\infty \ge 2^{-1}\lambda_{\mathrm{lasso}}\big).\end{aligned}$$

Note that $\zeta_i = \Psi'_\tau(\varepsilon_i) = 2\tau\varepsilon_i^+ - 2(1-\tau)\varepsilon_i^-$. It follows from Lemma 3 and $\mathscr{E}^\tau(\varepsilon_i) = 0$ that ζ_i are i.i.d. mean zero sub-Gaussian random variables. Now by the union bound argument and Lemma 3 again

$$\mathrm{P}\big(\big\|n^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\zeta}\big\|_\infty \ge 2^{-1}\lambda_{\mathrm{lasso}}\big) \le 2p\exp\bigg(-\frac{Cn\lambda^2_{\mathrm{lasso}}}{4K_0^2M_0^2}\bigg) = 1 - p_1^{\mathrm{ALS}}.$$

This completes the proof. □

LEMMA 5. *Under the assumptions of Theorem 2, the probability that the LLA algorithm (Algorithm 2) initialized by* $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ *converges to* $\hat{\boldsymbol{\beta}}^{\mathrm{oracle}}$ *after two iterations is at least* $1 - p_1 - p_2 - p_3$, *where*

$$\begin{aligned}p_1 &= \mathrm{P}\big(\big\|\hat{\boldsymbol{\beta}}^{\mathrm{lasso}} - \boldsymbol{\beta}^*\big\|_\infty > a_0\lambda\big),\\ p_2 &= \mathrm{P}\big(\big\|\nabla_{A^{\mathrm{c}}}\mathcal{L}_n\big(\hat{\boldsymbol{\beta}}^{\mathrm{oracle}}\big)\big\|_\infty \ge a_1\lambda\big),\\ p_3 &= \mathrm{P}\Big(\min_{j\in A}|\hat{\beta}_j^{\mathrm{oracle}}| < a\lambda\Big).\end{aligned}$$

PROOF. The convexity of $\mathcal{L}_n(\boldsymbol{\beta})$ follows from Lemma 1. Let $\mathcal{S} = \{\boldsymbol{\beta} \in \mathbb{R}^p: \boldsymbol{\beta}_{A^{\mathrm{c}}} = \mathbf{0}\}$. Note that $\hat{\boldsymbol{\beta}}^{\mathrm{oracle}} \in \mathcal{S}$. For any $\boldsymbol{\beta} \in \mathcal{S}$, let $\bar{\mathcal{L}}_n(\boldsymbol{\beta}_A) \equiv n^{-1}\sum_{i=1}^n \Psi_\tau(y_i - \mathbf{x}_{iA}^{\mathrm{T}}\boldsymbol{\beta}_A) = \mathcal{L}_n(\boldsymbol{\beta})$. Then $\nabla\bar{\mathcal{L}}_n(\boldsymbol{\beta}_A) = -n^{-1}\sum_{i=1}^n \mathbf{x}_{iA}\Psi'_\tau(y_i - \mathbf{x}_{iA}^{\mathrm{T}}\boldsymbol{\beta}_A)$. Now for any $\boldsymbol{\beta}$ and $\boldsymbol{\beta}' \in \mathcal{S}$, by Lemma 1 again, we get

$$\bar{\mathcal{L}}_n(\boldsymbol{\beta}_A) \ge \bar{\mathcal{L}}_n(\boldsymbol{\beta}'_A) + \langle\nabla\bar{\mathcal{L}}_n(\boldsymbol{\beta}'_A), \boldsymbol{\beta}_A - \boldsymbol{\beta}'_A\rangle + \underline{c}(\boldsymbol{\beta}_A - \boldsymbol{\beta}'_A)^{\mathrm{T}}\frac{\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A}{n}(\boldsymbol{\beta}_A - \boldsymbol{\beta}'_A).$$

Since $\mathbf{X}_A$ is of full column rank by assumption, we can see that $\bar{\mathcal{L}}_n(\boldsymbol{\beta}_A)$ is strongly convex with respect to $\boldsymbol{\beta}_A$ and, therefore, $\hat{\boldsymbol{\beta}}^{\mathrm{oracle}}$ is the unique solution of problem (2.10) with $\nabla\bar{\mathcal{L}}_n(\hat{\boldsymbol{\beta}}_A^{\mathrm{oracle}}) = \mathbf{0}$. The lemma then follows from Theorems 1 and 2 in Fan, Xue and Zou (2014). □

PROOF OF THEOREM 2. Let $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}}^{\mathrm{lasso}} - \boldsymbol{\beta}^*$. Assume both (C3) and (C4) hold. The other cases where either (C3) or (C4) holds are similar. From Lemma 5 and Theorem 1, we immediately get

$$\begin{aligned}p_1 &\le \mathrm{P}\big(\|\hat{\boldsymbol{\delta}}\|_\infty > \big[3s^{1/2}\lambda_{\mathrm{lasso}}(4\kappa\underline{c})^{-1}\big] \wedge \big[3\lambda_{\mathrm{lasso}}(4\varrho\underline{c})^{-1}\big]\big)\\ &\le \mathrm{P}\big(\|\hat{\boldsymbol{\delta}}\|_2 > 3s^{1/2}\lambda_{\mathrm{lasso}}(4\kappa\underline{c})^{-1}\big) \vee \mathrm{P}\big(\|\hat{\boldsymbol{\delta}}\|_\infty > 3\lambda_{\mathrm{lasso}}(4\varrho\underline{c})^{-1}\big) \le p_1^{\mathrm{ALS}}.\end{aligned}$$

To derive the bound for p_2, by the triangular inequality, it suffices to show bounds for $\mathrm{P}(\|\nabla_{A^c}\mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \geq 2^{-1}a_1\lambda)$ and $\mathrm{P}(\|\nabla_{A^c}\mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathrm{oracle}}) - \nabla_{A^c}\mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \geq 2^{-1}a_1\lambda)$. By the union bound argument and Lemma 3,

$$\begin{aligned}\mathrm{P}\big(\|\nabla_{A^c}\mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \geq 2^{-1}a_1\lambda\big) &= \mathrm{P}\big(\|-n^{-1}X_{A^c}^{\mathrm{T}}\boldsymbol{\zeta}\|_\infty \geq 2^{-1}a_1\lambda\big)\\ &\leq 2(p-s)\exp\Big(-\frac{Ca_1^2n\lambda^2}{4M_0^2K_0^2}\Big).\end{aligned}$$

Let $\mathbf{d} = (d_i, i = 1, \ldots, n)^{\mathrm{T}}$ with $d_i = \Psi_\tau'(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{\mathrm{oracle}}) - \Psi_\tau'(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*)$. By Cauchy–Schwarz inequality and Lemma 2, we get

$$\begin{aligned}&\|\nabla_{A^c}\mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathrm{oracle}}) - \nabla_{A^c}\mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty\\ &\quad= n^{-1}\max_{j\in A^c}\Big|\sum_{i=1}^n d_i x_{ij}\Big| \leq n^{-1}\max_{j\in A^c}\big(\|\mathbf{d}\|_2\|X_j\|_2\big)\\ &\quad\leq (2\bar{c}M_0)\big[(\hat{\boldsymbol{\beta}}_A^{\mathrm{oracle}} - \boldsymbol{\beta}_A^*)^{\mathrm{T}}\big(n^{-1}\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A\big)(\hat{\boldsymbol{\beta}}_A^{\mathrm{oracle}} - \boldsymbol{\beta}_A^*)\big]^{1/2}\\ &\quad\leq \big(2\bar{c}\rho_{\max}^{1/2}M_0\big)\|\hat{\boldsymbol{\beta}}^{\mathrm{oracle}} - \boldsymbol{\beta}^*\|_2.\end{aligned}$$

It follows from Lemma 4 and Lemma 3 that

$$\begin{aligned}&\mathrm{P}\big(\|\nabla_{A^c}\mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathrm{oracle}}) - \nabla_{A^c}\mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \geq 2^{-1}a_1\lambda\big)\\ &\quad\leq \mathrm{P}\Big(\|\hat{\boldsymbol{\beta}}^{\mathrm{oracle}} - \boldsymbol{\beta}^*\|_2 \geq \frac{a_1\lambda}{4\bar{c}\rho_{\max}^{1/2}M_0}\Big) \leq \mathrm{P}\big(\|n^{-1}\mathbf{X}_A^{\mathrm{T}}\boldsymbol{\zeta}\|_2 \geq Q_1\lambda\big)\\ &\quad\leq \Gamma(Q_1\lambda; n, s, K_0, M_0, \rho_{\max}, \nu_0).\end{aligned}$$

This establishes the desired upper bound for p_2. To show the upper bound for p_3, let $R = \min_{j\in A}|\beta_j^*| - a\lambda$ and observe that

$$\begin{aligned}p_3 &= \mathrm{P}\Big(\min_{j\in A}|\hat{\beta}_j^{\mathrm{oracle}}| < a\lambda\Big) \leq \mathrm{P}\big(\|\hat{\boldsymbol{\beta}}^{\mathrm{oracle}} - \boldsymbol{\beta}^*\|_\infty > R\big)\\ &\leq \mathrm{P}\big(\|\hat{\boldsymbol{\beta}}^{\mathrm{oracle}} - \boldsymbol{\beta}^*\|_2 > R\big) \leq \mathrm{P}\big(\|n^{-1}\mathbf{X}_A^{\mathrm{T}}\boldsymbol{\zeta}\|_2 \geq 2\underline{c}\rho_{\min}R\big).\end{aligned}$$

Similarly, by Lemma 3 we obtain

$$\mathrm{P}\big(\|n^{-1}\mathbf{X}_A^{\mathrm{T}}\boldsymbol{\zeta}\|_2 \geq 2\underline{c}\rho_{\min}R\big) \leq \Gamma(2\underline{c}\rho_{\min}R; n, s, K_0, M_0, \rho_{\max}, \nu_0),$$

which completes the proof. □

Let us now prove the results for the COSALES estimation. To simplify notation, let $\boldsymbol{\varpi} = (\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\varphi}^{\mathrm{T}})^{\mathrm{T}}$. It follows that $\mathrm{supp}(\boldsymbol{\varpi}^*) = A_0$. Let $\lambda_{\mathrm{lasso}} = \lambda_1^{\mathrm{lasso}} \wedge \lambda_2^{\mathrm{lasso}}$ and $\Lambda_{\mathrm{lasso}} = \lambda_1^{\mathrm{lasso}} \vee \lambda_2^{\mathrm{lasso}}$. We first present a lemma to facilitate the proofs.

LEMMA 6. *Let* $\boldsymbol{\varepsilon} = (\varepsilon_i, 1 \le i \le n)^{\mathrm{T}}$ *and* $\boldsymbol{\eta} = (\eta_i, 1 \le i \le n)^{\mathrm{T}}$, *where* $\eta_i = \Psi'_\tau(\varepsilon_i - e_\tau)$. *Also, let* $\mathbf{W} = \operatorname{diag}\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\omega}^*, 1 \le i \le n\}$.

(1) *For* $\boldsymbol{\varpi}, \boldsymbol{\delta} \in \mathbb{R}^{2p}$, $\langle \nabla S_n(\boldsymbol{\varpi} + \boldsymbol{\delta}) - \nabla S_n(\boldsymbol{\varpi}), \boldsymbol{\delta}\rangle \ge n^{-1}c_0\|(\mathbf{I}_2 \otimes \mathbf{X})\boldsymbol{\delta}\|_2^2$, *where* $\mathbf{I}_2$ *is a* 2×2 *identity matrix and* $c_0 = 2^{-1}[(1 + 4\underline{c}) - (1 + 16\underline{c}^2)^{1/2}] > 0$.

(2) *For* $d > 0$, $\mathrm{P}(\|\hat{\boldsymbol{\varpi}}^{\mathrm{oracle}} - \boldsymbol{\varpi}^*\|_2 > d) \le \mathrm{P}(\|\nabla_{A_0} S_n(\boldsymbol{\varpi}^*)\|_2 \ge c_0\phi_{\min} d)$, *where*

$$\nabla_{A_0} S_n(\boldsymbol{\varpi}^*) = -n^{-1}\begin{pmatrix} \mathbf{X}_{A_1}^{\mathrm{T}}\mathbf{W}(\boldsymbol{\varepsilon} + \boldsymbol{\eta}) \\ \mathbf{X}_{A_2}^{\mathrm{T}}\mathbf{W}\boldsymbol{\eta} \end{pmatrix}.$$

PROOF. The first part follows directly from the strong convexity of the (asymmetric) squared error loss. Specifically, note that since c_0 is the smaller eigenvalue of the 2×2 matrix $\left(\begin{smallmatrix} 1+2\underline{c} & 2\underline{c} \\ 2\underline{c} & 2\underline{c} \end{smallmatrix}\right)$, we have

$$\begin{aligned} S_n(\boldsymbol{\varpi} + \boldsymbol{\delta}) - S_n(\boldsymbol{\varpi}) - \langle \nabla S_n(\boldsymbol{\varpi}), \boldsymbol{\delta}\rangle &\ge \frac{1}{2n}\boldsymbol{\delta}^{\mathrm{T}}\left[\begin{pmatrix} 1 + 2\underline{c} & 2\underline{c} \\ 2\underline{c} & 2\underline{c} \end{pmatrix} \otimes (\mathbf{X}^{\mathrm{T}}\mathbf{X})\right]\boldsymbol{\delta} \\ &\ge (2n)^{-1}c_0\|(\mathbf{I}_2 \otimes \mathbf{X})\boldsymbol{\delta}\|_2^2. \end{aligned}$$

Similarly, $S_n(\boldsymbol{\varpi}) - S_n(\boldsymbol{\varpi} + \boldsymbol{\delta}) - \langle \nabla S_n(\boldsymbol{\varpi} + \boldsymbol{\delta}), -\boldsymbol{\delta}\rangle \ge (2n)^{-1}c_0\|(\mathbf{I}_2 \otimes \mathbf{X})\boldsymbol{\delta}\|_2^2$. Result (1) then follows by summing up the above two inequalities.

Let $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\varpi}}^{\mathrm{oracle}} - \boldsymbol{\varpi}^*$. Note that $\hat{\boldsymbol{\delta}}_{A_0^{\mathrm{c}}} = \mathbf{0}$ and $\nabla_{A_0} S_n(\hat{\boldsymbol{\varpi}}^{\mathrm{oracle}}) = \mathbf{0}$. From result (1), we have

$$\begin{aligned} c_0\phi_{\min}\|\hat{\boldsymbol{\delta}}\|_2^2 = c_0\phi_{\min}\|\hat{\boldsymbol{\delta}}_{A_0}\|_2^2 &\le n^{-1}c_0\|(\mathbf{I}_2 \otimes \mathbf{X})\hat{\boldsymbol{\delta}}\|_2^2 \\ &\le \langle \nabla S_n(\hat{\boldsymbol{\varpi}}^{\mathrm{oracle}}) - \nabla S_n(\boldsymbol{\varpi}^*), \hat{\boldsymbol{\delta}}\rangle = \langle -\nabla_{A_0} S_n(\boldsymbol{\varpi}^*), \hat{\boldsymbol{\delta}}_{A_0}\rangle \\ &\le \|\nabla_{A_0} S_n(\boldsymbol{\varpi}^*)\|_2^2\|\hat{\boldsymbol{\delta}}\|_2^2. \end{aligned}$$

Result (2) follows immediately. □

PROOF OF THEOREM 3. Let $\hat{\boldsymbol{\delta}}_1 = \hat{\boldsymbol{\gamma}}^{\mathrm{lasso}} - \boldsymbol{\gamma}^*$, $\hat{\boldsymbol{\delta}}_2 = \hat{\boldsymbol{\varphi}}^{\mathrm{lasso}} - \boldsymbol{\varphi}^*$, $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\delta}}_1^{\mathrm{T}}, \hat{\boldsymbol{\delta}}_2^{\mathrm{T}})^{\mathrm{T}}$, $z_{1\infty}^* = \|\partial S_n(\boldsymbol{\varpi}^*)/\partial\boldsymbol{\gamma}\|_\infty$, and $z_{2\infty}^* = \|\partial S_n(\boldsymbol{\varpi}^*)/\partial\boldsymbol{\varphi}\|_\infty$. By Lemma 6 and similar arguments in the proof of Theorem 1, it can be shown that

$$\begin{aligned} 0 \le n^{-1}c_0\|(\mathbf{I}_2 \otimes \mathbf{X})\hat{\boldsymbol{\delta}}\|_2^2 &\le \langle \nabla S_n(\hat{\boldsymbol{\varpi}}^{\mathrm{lasso}}) - \nabla S_n(\boldsymbol{\varpi}^*), \hat{\boldsymbol{\delta}}\rangle \\ &\le (z_{1\infty}^* + \lambda_1^{\mathrm{lasso}})\|\hat{\boldsymbol{\delta}}_{1A_1}\|_1 + (z_{1\infty}^* - \lambda_1^{\mathrm{lasso}})\|\hat{\boldsymbol{\delta}}_{1A_1^{\mathrm{c}}}\|_1 \\ &\quad + (z_{2\infty}^* + \lambda_2^{\mathrm{lasso}})\|\hat{\boldsymbol{\delta}}_{2A_2}\|_1 + (z_{2\infty}^* - \lambda_2^{\mathrm{lasso}})\|\hat{\boldsymbol{\delta}}_{2A_2^{\mathrm{c}}}\|_1. \end{aligned} \tag{6.2}$$

Under events $\mathcal{E}_1 = \{z_{1\infty}^* \le 2^{-1}\lambda_1^{\mathrm{lasso}}\}$ and $\mathcal{E}_2 = \{z_{2\infty}^* \le 2^{-1}\lambda_2^{\mathrm{lasso}}\}$, it follows from (6.2) that

$$\begin{aligned} 2^{-1}\lambda_{\mathrm{lasso}}\|\hat{\boldsymbol{\delta}}_{A_0^{\mathrm{c}}}\|_1 &\le 2^{-1}\lambda_1^{\mathrm{lasso}}\|\hat{\boldsymbol{\delta}}_{1A_1^{\mathrm{c}}}\|_1 + 2^{-1}\lambda_2^{\mathrm{lasso}}\|\hat{\boldsymbol{\delta}}_{2A_2^{\mathrm{c}}}\|_1 \\ &\le (3/2)\lambda_1^{\mathrm{lasso}}\|\hat{\boldsymbol{\delta}}_{1A_1}\|_1 + (3/2)\lambda_2^{\mathrm{lasso}}\|\hat{\boldsymbol{\delta}}_{2A_2}\|_1 \le (3/2)\Lambda_{\mathrm{lasso}}\|\hat{\boldsymbol{\delta}}_{A_0}\|_1, \end{aligned}$$

which implies that $\hat{\boldsymbol{\delta}} \in \mathscr{C}_{3\check{M}}$. Now under conditions (C4′)–(C5′) we have from (6.2) that

$$c_0\bar{\kappa}\|\hat{\boldsymbol{\delta}}\|_2^2 \le n^{-1}c_0\|(\mathbf{I}_2 \otimes \mathbf{X})\hat{\boldsymbol{\delta}}\|_2^2 \le (3/2)\Lambda_{\text{lasso}}\|\hat{\boldsymbol{\delta}}_{A_0}\|_1$$
$$\le (3/2)\Lambda_{\text{lasso}}(s_1+s_2)^{1/2}\|\hat{\boldsymbol{\delta}}\|_2$$

and that

$$c_0\bar{\varrho}\|\hat{\boldsymbol{\delta}}\|_\infty\|\hat{\boldsymbol{\delta}}_{A_0}\|_1 \le n^{-1}c_0\|(\mathbf{I}_2 \otimes \mathbf{X})\hat{\boldsymbol{\delta}}\|_2^2 \le (3/2)\Lambda_{\text{lasso}}\|\hat{\boldsymbol{\delta}}_{A_0}\|_1.$$

It follows that under events $\mathcal{E}_1$ and $\mathcal{E}_2$, we have $\|\hat{\boldsymbol{\delta}}\|_2 \le 3(s_1+s_2)^{1/2}\Lambda_{\text{lasso}}(2\bar{\kappa}c_0)^{-1}$ and $\|\hat{\boldsymbol{\delta}}\|_\infty \le 3\Lambda_{\text{lasso}}(2\bar{\varrho}c_0)^{-1}$. Recall that in Lemma 6 ε_i and $\eta_i = \Psi_\tau'(\varepsilon_i - e_\tau)$ are both mean zero sub-Gaussian random variables with $K_1 = \|\varepsilon_i\|_{\text{SG}}$ and $K_2 = \|\eta_i\|_{\text{SG}}$. It follows that $\varepsilon_i + \eta_i$ is also sub-Gaussian, and moreover, $\|\varepsilon_i + \eta_i\|_{\text{SG}} \le K_1 + K_2$. Since $M_1 = \|\mathbf{X}\boldsymbol{\omega}^*\|_\infty$, we have

$$\begin{aligned}
&\mathrm{P}\big(\|\hat{\boldsymbol{\delta}}\|_2 \le 3(s_1+s_2)^{1/2}\Lambda_{\text{lasso}}(2\bar{\kappa}c_0)^{-1} \cap \|\hat{\boldsymbol{\delta}}\|_\infty \le 3\Lambda_{\text{lasso}}(2\bar{\varrho}c_0)^{-1}\big)\\
&\quad \ge \mathrm{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - \mathrm{P}(\mathcal{E}_1^{\mathrm{c}}) - \mathrm{P}(\mathcal{E}_2^{\mathrm{c}})\\
&\quad = 1 - \mathrm{P}\big(\|n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(\boldsymbol{\varepsilon}+\boldsymbol{\eta})\|_\infty > 2^{-1}\lambda_1^{\text{lasso}}\big)\\
&\qquad - \mathrm{P}\big(\|n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\boldsymbol{\eta}\|_\infty > 2^{-1}\lambda_2^{\text{lasso}}\big)\\
&\quad \ge 1 - 2p\exp\Big(-\frac{Cn(\lambda_1^{\text{lasso}})^2}{4M_0^2M_1^2(K_1+K_2)^2}\Big) - 2p\exp\Big(-\frac{Cn(\lambda_2^{\text{lasso}})^2}{4M_0^2M_1^2K_2^2}\Big).
\end{aligned}$$

Theorem 3 then follows. □

The proof of Theorem 4 relies on the following lemma.

LEMMA 7. *Under assumptions of Theorem 4, the LLA algorithm (Algorithm 4) initialized by* $\hat{\boldsymbol{\gamma}}^{\text{lasso}}$ *and* $\hat{\boldsymbol{\varphi}}^{\text{lasso}}$ *converges to the oracle estimators* $\hat{\boldsymbol{\gamma}}^{\text{oracle}}$ *and* $\hat{\boldsymbol{\varphi}}^{\text{oracle}}$ *in two iterations with probability at least* $1 - \pi_1 - \pi_2 - \pi_3$, *where*

$$\pi_1 = \mathrm{P}\big(\|\hat{\boldsymbol{\gamma}}^{\text{lasso}} - \boldsymbol{\gamma}^*\|_\infty > a_0\lambda_1, \|\hat{\boldsymbol{\varphi}}^{\text{lasso}} - \boldsymbol{\varphi}^*\|_\infty > a_0\lambda_2\big),$$
$$\pi_2 = \mathrm{P}\big(\|\partial S_n(\hat{\boldsymbol{\varpi}}^{\text{oracle}})/\partial\boldsymbol{\gamma}_{A_1^{\mathrm{c}}}\|_\infty \ge a_1\lambda_1, \|\partial S_n(\hat{\boldsymbol{\varpi}}^{\text{oracle}})/\partial\boldsymbol{\varphi}_{A_2^{\mathrm{c}}}\|_\infty \ge a_1\lambda_2\big),$$
$$\pi_3 = \mathrm{P}\Big(\min_{j\in A_1}|\hat{\gamma}_j^{\text{oracle}}| < a\lambda_1, \min_{j\in A_2}|\hat{\varphi}_j^{\text{oracle}}| < a\lambda_2\Big).$$

PROOF. The convexity of $S_n(\boldsymbol{\gamma}, \boldsymbol{\varphi})$ follows immediately from Lemma 1,

$$\begin{aligned}
S_n(\boldsymbol{\gamma}, \boldsymbol{\varphi}) \ge{}& S_n(\boldsymbol{\gamma}', \boldsymbol{\varphi}') + \langle\nabla_{\boldsymbol{\gamma}} S_n(\boldsymbol{\gamma}', \boldsymbol{\varphi}'), \boldsymbol{\gamma} - \boldsymbol{\gamma}'\rangle + \langle\nabla_{\boldsymbol{\varphi}} S_n(\boldsymbol{\gamma}', \boldsymbol{\varphi}'), \boldsymbol{\varphi} - \boldsymbol{\varphi}'\rangle\\
&+ 2^{-1}\begin{pmatrix}\boldsymbol{\gamma}-\boldsymbol{\gamma}'\\ \boldsymbol{\varphi}-\boldsymbol{\varphi}'\end{pmatrix}^{\mathrm{T}}\left[\begin{pmatrix}1+2\underline{c} & 2\underline{c}\\ 2\underline{c} & 2\underline{c}\end{pmatrix}\otimes(n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})\right]\begin{pmatrix}\boldsymbol{\gamma}-\boldsymbol{\gamma}'\\ \boldsymbol{\varphi}-\boldsymbol{\varphi}'\end{pmatrix}.
\end{aligned}$$

Restrict $S_n(\boldsymbol{\gamma},\boldsymbol{\varphi})$ to the set $\mathcal{S}=\{\boldsymbol{\gamma},\boldsymbol{\varphi}\in\mathbb{R}^p\colon \boldsymbol{\gamma}_{A_1^c}=\mathbf{0},\boldsymbol{\varphi}_{A_2^c}=\mathbf{0}\}$ and define for any $(\boldsymbol{\gamma},\boldsymbol{\varphi})\in\mathcal{S}$

$$\breve{S}_n(\boldsymbol{\gamma}_{A_1},\boldsymbol{\varphi}_{A_2})=n^{-1}\sum_{i=1}^n\bigl\{\Psi_{0.5}\bigl(y_i-\mathbf{x}_{iA_1}^{\mathrm{T}}\boldsymbol{\gamma}_{A_1}\bigr)+\Psi_\tau\bigl(y_i-\mathbf{x}_{iA_1}^{\mathrm{T}}\boldsymbol{\gamma}_{A_1}-\mathbf{x}_{iA_2}^{\mathrm{T}}\boldsymbol{\varphi}_{A_2}\bigr)\bigr\}.$$

It follows immediately that for any $(\boldsymbol{\gamma},\boldsymbol{\varphi})$, $(\boldsymbol{\gamma}',\boldsymbol{\varphi}')\in\mathcal{S}$,

$$\begin{aligned}\breve{S}_n(\boldsymbol{\gamma}_{A_1},\boldsymbol{\varphi}_{A_2})\geq{}&\breve{S}_n(\boldsymbol{\gamma}'_{A_1},\boldsymbol{\varphi}'_{A_2})+\langle\nabla_{\boldsymbol{\gamma}_{A_1}}\breve{S}_n(\boldsymbol{\gamma}'_{A_1},\boldsymbol{\varphi}'_{A_2}),\boldsymbol{\gamma}_{A_1}-\boldsymbol{\gamma}'_{A_1}\rangle\\&+\langle\nabla_{\boldsymbol{\varphi}_{A_2}}\breve{S}_n(\boldsymbol{\gamma}'_{A_1},\boldsymbol{\varphi}'_{A_2}),\boldsymbol{\varphi}_{A_2}-\boldsymbol{\varphi}'_{A_2}\rangle\\&+2^{-1}c_0(\boldsymbol{\gamma}_{A_1}-\boldsymbol{\gamma}'_{A_1})^{\mathrm{T}}(n^{-1}\mathbf{X}_{A_1}^{\mathrm{T}}\mathbf{X}_{A_1})(\boldsymbol{\gamma}_{A_1}-\boldsymbol{\gamma}'_{A_1})\\&+2^{-1}c_0(\boldsymbol{\varphi}_{A_2}-\boldsymbol{\varphi}'_{A_2})^{\mathrm{T}}(n^{-1}\mathbf{X}_{A_2}^{\mathrm{T}}\mathbf{X}_{A_2})(\boldsymbol{\varphi}_{A_2}-\boldsymbol{\varphi}'_{A_2}),\end{aligned}$$

where $c_0=2^{-1}[(1+4\underline{c})-(1+16\underline{c}^2)^{1/2}]$. Since both $\mathbf{X}_{A_1}$ and $\mathbf{X}_{A_2}$ are of full column ranks by assumption, we can see that $\breve{S}_n(\boldsymbol{\gamma}_{A_1},\boldsymbol{\varphi}_{A_2})$ is strongly convex and thus the oracle estimators $\hat{\boldsymbol{\gamma}}^{\mathrm{oracle}}$ and $\hat{\boldsymbol{\varphi}}^{\mathrm{oracle}}$ are the unique solution of problem (4.3).

Let $\mathcal{E}_1$ be the event that $\|\hat{\boldsymbol{\gamma}}^{\mathrm{lasso}}-\boldsymbol{\gamma}^*\|_\infty\leq a_0\lambda_1$ and $\|\hat{\boldsymbol{\varphi}}^{\mathrm{lasso}}-\boldsymbol{\varphi}^*\|_\infty\leq a_0\lambda_2$. Under $\mathcal{E}_1$ and Assumption (A0′), on one hand we have $\min_{j\in A_1}|\hat{\gamma}_j^{\mathrm{lasso}}|\geq\min_{j\in A_1}|\gamma_j^*|-\|\hat{\boldsymbol{\gamma}}^{\mathrm{lasso}}-\boldsymbol{\gamma}^*\|_\infty>a\lambda_1$, implying that $p'_{\lambda_1}(|\hat{\gamma}_j^{\mathrm{lasso}}|)=0$ for $j\in A_1$. On the other hand, we have $\|\hat{\boldsymbol{\gamma}}_{A^c}^{\mathrm{lasso}}\|_\infty\leq\|\hat{\boldsymbol{\gamma}}^{\mathrm{lasso}}-\boldsymbol{\gamma}^*\|_\infty\leq a_2\lambda_1$, indicating that $p'_{\lambda_1}(|\hat{\gamma}_j^{\mathrm{lasso}}|)\geq a_1\lambda_1$ for $j\in A_1^c$. Similarly, we can show that $p'_{\lambda_2}(|\hat{\varphi}_j^{\mathrm{lasso}}|)=0$ for $j\in A_2$ and $p'_{\lambda_2}(|\hat{\varphi}_j^{\mathrm{lasso}}|)\geq a_1\lambda_2$ for $j\in A_2^c$.

Let $\hat{\boldsymbol{\gamma}}^1$ and $\hat{\boldsymbol{\varphi}}^1$ be the update after the first iteration of the LLA algorithm. Then under $\mathcal{E}1$, $\hat{\boldsymbol{\gamma}}^1$ and $\hat{\boldsymbol{\varphi}}^1$ are minimizers of

$$\mathcal{Q}_n(\boldsymbol{\gamma},\boldsymbol{\varphi})=S_n(\boldsymbol{\gamma},\boldsymbol{\varphi})+\sum_{j\in A_1^c}p'_{\lambda_1}\bigl(|\hat{\gamma}_j^{\mathrm{lasso}}|\bigr)|\gamma_j|+\sum_{j\in A_2^c}p'_{\lambda_2}\bigl(|\hat{\varphi}_j^{\mathrm{lasso}}|\bigr)|\varphi_j|.$$

By definition of the oracle estimators, $\partial S_n(\hat{\boldsymbol{\gamma}}^{\mathrm{oracle}},\hat{\boldsymbol{\varphi}}^{\mathrm{oracle}})/\partial\gamma_j=0$ for $j\in A_1$ and $\partial S_n(\hat{\boldsymbol{\gamma}}^{\mathrm{oracle}},\hat{\boldsymbol{\varphi}}^{\mathrm{oracle}})/\partial\varphi_j=0$ for $j\in A_2^c$. Also, $\hat{\boldsymbol{\gamma}}_{A_1^c}^{\mathrm{oracle}}=\mathbf{0}$ and $\hat{\boldsymbol{\varphi}}_{A_2^c}^{\mathrm{oracle}}=\mathbf{0}$. Now let $\mathcal{E}_2$ be the event that $\max_{j\in A_1^c}|\partial\mathcal{L}(\hat{\boldsymbol{\gamma}}^{\mathrm{oracle}},\hat{\boldsymbol{\varphi}}^{\mathrm{oracle}})/\partial\gamma_j|<a_1\lambda_1$ and that $\max_{j\in A_2^c}|\partial\mathcal{L}(\hat{\boldsymbol{\gamma}}^{\mathrm{oracle}},\hat{\boldsymbol{\varphi}}^{\mathrm{oracle}})/\partial\varphi_j|<a_1\lambda_2$. It follows from the convexity of $S_n(\boldsymbol{\gamma},\boldsymbol{\varphi})$ that

$$\begin{aligned}&\mathcal{Q}_n(\boldsymbol{\gamma},\boldsymbol{\varphi})-\mathcal{Q}_n\bigl(\hat{\boldsymbol{\gamma}}^{\mathrm{oracle}},\hat{\boldsymbol{\varphi}}^{\mathrm{oracle}}\bigr)\\&\quad\geq\sum_{j\in A_1^c}\frac{\partial}{\partial\gamma_j}S_n\bigl(\hat{\boldsymbol{\gamma}}^{\mathrm{oracle}},\hat{\boldsymbol{\varphi}}^{\mathrm{oracle}}\bigr)\gamma_j+\sum_{j\in A_2^c}\frac{\partial}{\partial\varphi_j}S_n\bigl(\hat{\boldsymbol{\gamma}}^{\mathrm{oracle}},\hat{\boldsymbol{\varphi}}^{\mathrm{oracle}}\bigr)\varphi_j\\&\qquad+\sum_{j\in A_1^c}p'_{\lambda_1}\bigl(|\hat{\gamma}_j^{\mathrm{lasso}}|\bigr)|\gamma_j|+\sum_{j\in A_2^c}p'_{\lambda_2}\bigl(|\hat{\varphi}_j^{\mathrm{lasso}}|\bigr)|\varphi_j|.\end{aligned}$$

Under $\mathcal{E}_2$, this implies that $Q_n(\boldsymbol{\gamma},\boldsymbol{\varphi}) \geq Q_n(\hat{\boldsymbol{\gamma}}^{\text{oracle}}, \hat{\boldsymbol{\varphi}}^{\text{oracle}})$ for any $\boldsymbol{\gamma} \in \mathbb{R}^p$ and $\boldsymbol{\varphi} \in \mathbb{R}^p$. The strict inequality holds unless $\gamma_j = 0$ for all $j \in A_1^{\text{c}}$ and $\varphi_j = 0$ for all $j \in A_2^{\text{c}}$. By the uniqueness of the oracle estimators, we must have $\hat{\boldsymbol{\gamma}}^1 = \hat{\boldsymbol{\gamma}}^{\text{oracle}}$ and $\hat{\boldsymbol{\varphi}}^1 = \hat{\boldsymbol{\varphi}}^{\text{oracle}}$.

Let $\mathcal{E}_3$ be the event that $\min_{j\in A_1} |\hat{\gamma}_j^{\text{oracle}}| \geq a\lambda_1$ and $\min_{j\in A_2} |\hat{\varphi}_j^{\text{oracle}}| \geq a\lambda_2$. Once the oracle estimators are obtained after the first iteration, under $\mathcal{E}_3$, we can see that $p'_{\lambda_1}(|\hat{\gamma}_j^{\text{oracle}}|) = 0$ for $j \in A_1$, $p'_{\lambda_1}(|\hat{\gamma}_j^{\text{oracle}}|) \geq a_1\lambda_1$ for $j \in A_1^{\text{c}}$ and $p'_{\lambda_2}(|\hat{\varphi}_j^{\text{oracle}}|) = 0$ for $j \in A_2$, $p'_{\lambda_2}(|\hat{\varphi}_j^{\text{oracle}}|) \geq a_1\lambda_2$ for $j \in A_2^{\text{c}}$. By similar arguments, it can be shown that the second iteration of the LLA algorithm will still yield the oracle estimators, which means the algorithm converges to the oracle estimators hereafter. This completes the proof. □

PROOF OF THEOREM 4. Let $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\varpi}}^{\text{lasso}} - \boldsymbol{\varpi}^*$. Assume both (C4′) and (C5′) hold. The other cases where either (C4′) or (C5′) holds are similar. It follows from Theorem 3 that

$$\begin{aligned}\pi_1 &\leq \mathrm{P}\big(\|\hat{\boldsymbol{\delta}}\|_\infty > a_0\lambda\big) \leq \mathrm{P}\big(\|\hat{\boldsymbol{\delta}}\|_\infty > 3\Lambda_{\text{lasso}}(2c_0)^{-1}\big[\big(s^{1/2}\bar{\kappa}^{-1}\big) \wedge \bar{\varrho}^{-1}\big]\big)\\ &\leq \mathrm{P}\big(\|\hat{\boldsymbol{\delta}}\|_2 > 3s^{1/2}\Lambda_{\text{lasso}}(2c_0\bar{\kappa})^{-1}\big) \vee \mathrm{P}\big(\|\hat{\boldsymbol{\delta}}\|_\infty > 3\Lambda_{\text{lasso}}(2c_0\bar{\varrho})^{-1}\big) \leq \pi_1^{\text{ALS}}.\end{aligned}$$

Next, note that $\pi_2 \leq \mathrm{P}(\|\nabla_{A_0^{\text{c}}} S_n(\hat{\boldsymbol{\varpi}}^{\text{oracle}})\|_\infty \geq a_1\lambda)$. By the triangular inequality, it suffices to show upper bounds for respectively $\mathrm{P}(\|\nabla_{A_0^{\text{c}}} S_n(\boldsymbol{\varpi}^*)\|_\infty \geq 2^{-1}a_1\lambda)$ and $\mathrm{P}(\|\nabla_{A_0^{\text{c}}} S_n(\hat{\boldsymbol{\varpi}}^{\text{oracle}}) - \nabla_{A_0^{\text{c}}} S_n(\boldsymbol{\varpi}^*)\|_\infty \geq 2^{-1}a_1\lambda)$. First, by the union bound argument we have

$$\begin{aligned}&\mathrm{P}\big(\big\|\nabla_{A_0^{\text{c}}} S_n\big(\boldsymbol{\varpi}^*\big)\big\|_\infty \geq 2^{-1}a_1\lambda\big)\\ &\quad\leq \mathrm{P}\big(\big\|n^{-1}\mathbf{X}_{A_1^{\text{c}}}^{\mathrm{T}}\mathbf{W}(\boldsymbol{\varepsilon}+\boldsymbol{\eta})\big\|_\infty \geq 2^{-1}a_1\lambda\big) + \mathrm{P}\big(\big\|n^{-1}\mathbf{X}_{A_2^{\text{c}}}^{\mathrm{T}}\mathbf{W}\boldsymbol{\eta}\big\|_\infty \geq 2^{-1}a_1\lambda\big)\\ &\quad\leq 2(p-s_1)\exp\bigg(-\frac{Ca_1^2n\lambda^2}{4M_0^2M_1^2(K_1+K_2)^2}\bigg) + 2(p-s_2)\exp\bigg(-\frac{Ca_1^2n\lambda^2}{4M_0^2M_1^2K_2^2}\bigg).\end{aligned}$$

Now let $\bar{d}_i = \Psi'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\gamma}}^{\text{oracle}} - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\varphi}}^{\text{oracle}}) - \Psi'_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}^* - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\varphi}^*)$ and set $\bar{\mathbf{d}} = (\bar{d}_i, 1 \leq i \leq n)^{\mathrm{T}}$. It follows that

$$\begin{aligned}&\big\|\nabla_{A_0^{\text{c}}} S_n\big(\hat{\boldsymbol{\varpi}}^{\text{oracle}}\big) - \nabla_{A_0^{\text{c}}} S_n\big(\boldsymbol{\varpi}^*\big)\big\|_\infty\\ &\quad\leq M_0\big(\big\|\mathbf{X}\big(\hat{\boldsymbol{\gamma}}^{\text{oracle}} - \boldsymbol{\gamma}^*\big)\big\|_2 + \|\bar{\mathbf{d}}\|_2\big)/\sqrt{n}\\ &\quad\leq M_0\big[(1+2\bar{c})\big\|\mathbf{X}_{A_1}\big(\hat{\boldsymbol{\gamma}}_{A_1}^{\text{oracle}} - \boldsymbol{\gamma}_{A_1}^*\big)\big\|_2\\ &\qquad + (2\bar{c})\big\|\mathbf{X}_{A_2}\big(\hat{\boldsymbol{\varphi}}_{A_2}^{\text{oracle}} - \boldsymbol{\varphi}_{A_2}^*\big)\big\|_2\big]/\sqrt{n}\\ &\quad\leq (1+2\bar{c})M_0\phi_{\max}^{1/2}\big\|\hat{\boldsymbol{\varpi}}^{\text{oracle}} - \boldsymbol{\varpi}^*\big\|_2.\end{aligned}$$

By Lemma 6 and Lemma 3, we get

$$
\begin{aligned}
&\mathrm{P}\big(\big\|\nabla_{A_0^c} S_n\big(\hat{\boldsymbol{\varpi}}^{\text{oracle}}\big) - \nabla_{A_0^c} S_n\big(\boldsymbol{\varpi}^*\big)\big\|_\infty \geq 2^{-1}a_1\lambda\big)\\
&\quad\leq \mathrm{P}\bigg(\big\|\hat{\boldsymbol{\varpi}}^{\text{oracle}} - \boldsymbol{\varpi}^*\big\|_2 \geq \frac{a_1\lambda}{2(1+2\underline{c})M_0\phi_{\max}^{1/2}}\bigg)\\
&\quad\leq \mathrm{P}\bigg(\bigg\|\frac{1}{n}\begin{pmatrix}\mathbf{X}_{A_1}^{\mathrm{T}}\mathbf{W}(\boldsymbol{\varepsilon}+\boldsymbol{\eta})\\ \mathbf{X}_{A_2}^{\mathrm{T}}\mathbf{W}\boldsymbol{\eta}\end{pmatrix}\bigg\|_2 \geq Q_2\lambda\bigg)\\
&\quad\leq \mathrm{P}\big(\big\|n^{-1}\mathbf{X}_{A_1}^{\mathrm{T}}\mathbf{W}(\boldsymbol{\varepsilon}+\boldsymbol{\eta})\big\|_2 \geq 2^{-1}Q_2\lambda\big) + \mathrm{P}\big(\big\|n^{-1}\mathbf{X}_{A_2}^{\mathrm{T}}\mathbf{W}\boldsymbol{\eta}\big\|_2 \geq 2^{-1}Q_2\lambda\big)\\
&\quad\leq \Gamma\big(2^{-1}Q_2\lambda; n, s_1, K_1+K_2, M_0M_1, M_1^2\rho_{1\bullet\max}, \nu_1\big)\\
&\qquad + \Gamma\big(2^{-1}Q_2\lambda; n, s_2, K_2, M_0M_1, M_1^2\rho_{2\bullet\max}, \nu_2\big).
\end{aligned}
$$

This completes the upper bound for π_2. To derive the upper bound for π_3, note that by Assumption (A0′) we have $\min_{j\in A_1}|\gamma_j^*| \geq (a+1)\lambda_1$ and $\min_{j\in A_2}|\varphi_j^*| \geq (a+1)\lambda_2$. Observe that $\min_{j\in A_1}|\hat{\gamma}_j^{\text{oracle}}| \geq \min_{j\in A_1}|\gamma_j^*| - \|\hat{\boldsymbol{\gamma}}^{\text{oracle}} - \boldsymbol{\gamma}^*\|_\infty$ and $\min_{j\in A_2}|\hat{\varphi}_j^{\text{oracle}}| \geq \min_{j\in A_1}|\varphi_j^*| - \|\hat{\boldsymbol{\varphi}}^{\text{oracle}} - \boldsymbol{\varphi}^*\|_\infty$, and it follows that

$$
\begin{aligned}
\pi_3 &\leq \mathrm{P}\big(\big\|\hat{\boldsymbol{\varpi}}^{\text{oracle}} - \boldsymbol{\varpi}^*\big\|_\infty > \bar{R}\big) \leq \mathrm{P}\big(\big\|\hat{\boldsymbol{\varpi}}^{\text{oracle}} - \boldsymbol{\varpi}^*\big\|_2 > \bar{R}\big)\\
&\leq \mathrm{P}\bigg(\bigg\|\frac{1}{n}\begin{pmatrix}\mathbf{X}_{A_1}^{\mathrm{T}}\mathbf{W}(\boldsymbol{\varepsilon}+\boldsymbol{\eta})\\ \mathbf{X}_{A_2}^{\mathrm{T}}\mathbf{W}\boldsymbol{\eta}\end{pmatrix}\bigg\|_2 \geq c_0\phi_{\min}\bar{R}\bigg)\\
&\leq \mathrm{P}\bigg(\big\|n^{-1}\mathbf{X}_{A_1}^{\mathrm{T}}\mathbf{W}(\boldsymbol{\varepsilon}+\boldsymbol{\eta})\big\|_2 \geq \frac{1}{2}c_0\phi_{\min}\bar{R}\bigg) + \mathrm{P}\bigg(\big\|n^{-1}\mathbf{X}_{A_2}^{\mathrm{T}}\mathbf{W}\boldsymbol{\eta}\big\|_2 \geq \frac{1}{2}c_0\phi_{\min}\bar{R}\bigg)\\
&\leq \Gamma\big(2^{-1}c_0\phi_{\min}\bar{R}; n, s_1, K_1+K_2, M_0M_1, M_1^2\rho_{1\bullet\max}, \nu_1\big)\\
&\quad + \Gamma\big(2^{-1}c_0\phi_{\min}\bar{R}; n, s_2, K_2, M_0M_1, M_1^2\rho_{2\bullet\max}, \nu_2\big). \qquad \square
\end{aligned}
$$

Acknowledgments. The authors sincerely thank the Editor, Associate Editor, and three referees for their helpful comments and suggestions that led to substantial improvement of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "High-dimensional generalizations of asymmetric least squares regression and their applications" (DOI: 10.1214/15-AOS1431SUPP; .pdf). The supplementary material includes the iteration complexity analysis of the SALES algorithm.

REFERENCES

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

CHAMBERS, R. and TZAVIDIS, N. (2006). M-quantile models for small area estimation. *Biometrika* **93** 255–268. MR2278081

DAYE, Z. J., CHEN, J. and LI, H. (2012). High-dimensional heteroscedastic regression with an application to eQTL data analysis. *Biometrics* **68** 316–326. MR2909888

EFRON, B. (1991). Regression percentiles using asymmetric squared error loss. *Statist. Sinica* **1** 93–125. MR1101317

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

EILERS, P. H. and BOELENS, H. F. (2005). Baseline correction with asymmetric least squares smoothing. Leiden Univ. Medical Centre Report.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. MR2849368

FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GU, Y. and ZOU, H. (2015). Supplement to "High-dimensional generalizations of asymmetric least squares regression and their applications." DOI:10.1214/15-AOS1431SUPP.

HUANG, J. and ZHANG, C. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *J. Mach. Learn. Res.* **13** 1839–1864. MR2956344

KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644

KOENKER, R. and BASSETT, G. JR. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* **50** 43–61. MR0640165

KOENKER, R. and ZHAO, Q. S. (1994). L-estimation for linear heteroscedastic models. *J. Nonparametr. Stat.* **3** 223–235. MR1291546

KUAN, C., YEH, J. and HSU, Y. (2009). Assessing value at risk with CARE, the conditional autoregressive expectile models. *J. Econometrics* **150** 261–270. MR2535521

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

NEWEY, W. K. and POWELL, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55** 819–847. MR0906565

PARIKH, N. and BOYD, S. (2013). Proximal algorithms. *Found. Trends Optim.* **1** 123–231.

RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. MR3125258

SALVATI, N., TZAVIDIS, N., PRATESI, M. and CHAMBERS, R. (2012). Small area estimation via M-quantile geographically weighted regression. *TEST* **21** 1–28. MR2912969

SCHEETZ, T. E., KIM, K. A., SWIDERSKI, R. E., PHILP, A. R., BRAUN, T. A., KNUDTSON, K. L., DORRANCE, A. M., DIBONA, G. F., HUANG, J., CASAVANT, T. L. et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natl. Acad. Sci. USA* **103** 14429–14434.

TAYLOR, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *J. Financ. Econom.* **6** 231–252.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. MR1835069

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. Preprint. Available at arXiv:1011.3027v7.

WANG, L., KIM, Y. and LI, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Ann. Statist.* **41** 2505–2536. MR3127873

WANG, L., WU, Y. and LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.* **107** 214–222. MR2949353

XIE, S., ZHOU, Y. and WAN, A. T. K. (2014). A varying-coefficient expectile model for estimating value at risk. *J. Bus. Econom. Statist.* **32** 576–592. MR3272888

YANG, Y. and ZOU, H. (2013). An efficient algorithm for computing the HHSVM and its generalizations. *J. Comput. Graph. Statist.* **22** 396–415. MR3173721

YE, F. and ZHANG, C. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

ZHANG, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA
USA
E-MAIL: guxxx192@umn.edu
zouxx019@umn.edu