Orthogonalizing Penalized Regression

Shifeng Xiong[1], Bin Dai[2], and Peter Z. G. Qian[2]

[1] Academy of Mathematics and Systems Science

Chinese Academy of Sciences, Beijing 100190

[2] Department of Statistics

University of Wisconsin-Madison, Madison, WI 53706

Abstract

Since the penalized likelihood function of the smoothly clipped absolute deviation (SCAD) penalty is highly non-linear and has many local optima, finding a local solution to achieve the so-called oracle property is an open problem. We propose an iterative algorithm, called the OEM algorithm, to fill this gap. The development of the algorithm draws direct impetus from a missing-data problem arising in design of experiments with an orthogonal complete matrix. In each iteration, the algorithm imputes the missing data based on the current estimates of the parameters and updates a closed-form solution associated with the complete data. By introducing a procedure called active orthogonization, we make the algorithm broadly applicable to problems with arbitrary regression matrices. In addition to the SCAD penalty, the proposed algorithm works for other penalties like the MCP, lasso and nonnegative garrote. Convergence and convergence rate of the algorithm are examined. The algorithm has several unique theoretical properties. For the SCAD and MCP penalties, an OEM sequence can achieve the oracle property after sufficient iterations. For various penalties, an OEM sequence converges to a point having grouping coherence for fully aliased regression matrices. For computing the ordinary least squares estimator with a singular regression matrix, an OEM sequence converges to the Moore-Penrose generalized inverse-based least squares estimator.

KEY WORDS: Design of experiments; MCP; Missing data; Optimization; Oracle property; Orthogonal design; SCAD; The EM algorithm; The Lasso.

[2]Corresponding author: Peter Z. G. Qian. Email: peterq@stat.wisc.edu

1 INTRODUCTION

Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty to achieve simultaneous estimation and variable selection. Consider a linear model

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

where $\boldsymbol{X} = (x_{ij})$ is the $n \times p$ regression matrix, $\mathbf{Y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients and the distribution of the vector of random error $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is $N(\mathbf{0}_n,\ \sigma^2 \boldsymbol{I}_n)$ with $\mathbf{0}_n$ being the nth zero vector and $\boldsymbol{I}_n$ being the $n \times n$ identity matrix. Throughout, let $\|\cdot\|$ denote the Euclidean norm. A *regularized least squares estimator* of $\boldsymbol{\beta}$ with this penalty is given by solving

$$\min_{\boldsymbol{\beta}} \left[\|\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2\sum_{j=1}^{p} P_\lambda(|\beta_j|) \right], \tag{2}$$

where for $\theta > 0$,

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

$a > 2$, $\lambda > 0$ is the tuning parameter and I is the indicator function. In order to apply the penalty P_λ equally on all the variables, $\boldsymbol{X}$ can be standardized so that

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

Both theory and computation of the estimator in (2) have been actively studied. On the theoretical side, Fan and Li (2001) introduced an important concept, called the *oracle* property. An estimator of $\boldsymbol{\beta}$ having this property can not only select the correct submodel asymptotically, but also estimate the nonzero coefficients as efficiently as if the correct submodel were known in advance. On the computational side, existing algorithms for solving this optimization problem include local quadratic approximation (Fan and Li 2001; Hunter and Li 2005), local linear approximation (Zou and Li 2008), the coordinate descent algorithm (Tseng 2001;

Tseng and Yun 2009; Breheny and Huang 2010; Mazumder, Friedman, and Hastie 2010) and the minimization by iterative soft thresholding (MIST) algorithm (Schifano, Strawderman, and Wells 2010), among others.

Departing from the existing work, we study the SCAD penalty from a new perspective, targeting on the *interface* between theory and computing. Fan and Li (2001) proved that there exists a local solution to (2) with the oracle property. From the optimization viewpoint, (2) can have many local minima (Huo and Chen 2010) and it is very challenging to find one of them to achieve the oracle property. To the best of our knowledge, no theoretical results are available to show that any existing algorithm can provide such a local minimum. We propose an iterative algorithm, called orthogonalizing EM (OEM), to fill this gap. We will show in Section 4 that the OEM solution to (2) can indeed achieve the oracle property under regularity conditions. OEM draws its direct impetus from a missing data problem with a complete orthogonal design arising in design of experiments. Throughout, a matrix is orthogonal if its columns are orthogonal. In each iteration, the algorithm imputes the missing data based on the current estimate of $\boldsymbol{\beta}$ and updates a closed-form solution to (2) associated with the complete data. Much beyond this orthogonal design formulation, the OEM algorithm applies to general data structures by *actively orthogonalizing* arbitrary regression matrices.

Though the inspiration of the OEM algorithm stems from the SCAD penalty, it, not surprisingly, works for the general penalized regression problem:

$$\min_{\boldsymbol{\beta}\in\Theta} \left[\|\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + P(\boldsymbol{\beta};\lambda)\right], \tag{5}$$

where $\boldsymbol{\beta} \in \Theta$, Θ is a subset of $\mathbb{R}^p$ and λ is the vector of tuning parameters. Besides the SCAD penalty, choices for $P(\boldsymbol{\beta};\lambda)$ include the ridge regression (Hoerl and Kennard 1970), the nonnegative garrote (Breiman 1995), the lasso (Tibshirani 1996) and the MCP (Zhang 2010). Algorithms for solving the problem in (5) include those developed in Fu (1998), Grandvalet (1998), Osborne, Presnell, and Turlach (2000), the LARS algorithm introduced in Efron, Hastie, Johnstone, and Tibshirani (2004) and the coordinate descent algorithm

(Tseng 2001; Friedman, Hastie, Hofling and Tibshirani 2007; Wu and Lange 2008; Tseng and Yun 2009), and are available in R packages like lars (Hastie and Efron 2011), glmnet (Friedman, Hastie, and Tibshirani 2011) and scout (Witten and Tibshirani 2011).

In addition to achieving the oracle property for the SCAD and MCP penalties, the OEM algorithm has several other unique theoretical features. 1. *Having grouping coherence*: An estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ in (1) is said to have grouping coherence if it has the same coefficient for full aliased columns in $\boldsymbol{X}$ (Zou and Hastie 2005). For the lasso, SCAD and MCP, an OEM sequence converges to a point having grouping coherence. 2. *Convergence in singular case*: When $\boldsymbol{X}$ in (1) is singular, the ordinary least squares estimator given by (5) without any penalty is not unique. For this singular case, an OEM solution, or essentially the Healy-Westmacott estimator (Healy and Westmacott 1956), converges to the Moore-Penrose generalized inverse-based least squares estimator.

The remainder of the article will unfold as follows. Section 2 derives the OEM algorithm for a missing data problem with a complete orthogonal design. Section 3 significantly broadens the applicability of the algorithm by introducing an idea for actively expanding any regression matrix to an orthogonal matrix. Section 4 establishes the oracle property of the OEM solution for the SCAD and MCP. Section 5 provides convergence properties of the OEM algorithm. Section 6 shows that for a regression matrix with full aliased columns, an OEM sequence for the lasso, SCAD or MCP converges to a solution with grouping coherency and illustrates how to use the OEM algorithm to compute the ordinary least squares estimator for a singular regression matrix. Section 7 provides some discussion.

2 THE OEM ALGORITHM AS A PENALIZED HEALY-WESTMACOTT PROCEDURE

Orthogonal designs are widely used in science and engineering. Such designs have been intensively studied in different branches of statistics including design of experiments, information theory (MacWilliams and Sloane 1977), liner models, sampling survey and computer

experiments. Popular classes of orthogonal designs include orthogonal arrays (Hedayat, Sloane, and Stufken 1999), orthogonal main-effect plans (Addelman 1962; Wu and Hamada 2009) and orthogonal Latin hypercube designs (Ye 1998; Steinberg and Lin 2006; Bingham, Sitter, and Tang 2009; Lin, Mukerjee, and Tang 2009; Pang, Liu, and Lin, 2009; Sun, Liu, and Lin 2009; Lin, Bingham, Sitter, and Tang 2010) from the computer experiments literature (Santner, Williams, and Notz 2003; Fang, Li, and Sudjianto 2005). In this section, we motivate the OEM algorithm by using a missing data problem with an orthogonal complete design. Suppose that the matrix $\boldsymbol{X}$ in (1) for this problem is a submatrix of an $m \times p$ *complete* orthogonal matrix

$$\boldsymbol{X}_c = (\boldsymbol{X}' \ \boldsymbol{\Delta}')', \tag{6}$$

where $\boldsymbol{\Delta}$ is the $(m-n) \times p$ *missing* matrix. Let

$$\mathbf{Y}_c = (\mathbf{Y}', \ \mathbf{Y}'_{\text{miss}})' \tag{7}$$

define the vector of complete observations with $\mathbf{Y}_{\text{miss}}$ corresponding to $\boldsymbol{\Delta}$. If $\mathbf{Y}_{\text{miss}}$ were observable, then the ordinary least square estimator of $\boldsymbol{\beta}$ based on the complete data ($\boldsymbol{X}_c$, $\mathbf{Y}_c$) has a closed form as $\boldsymbol{X}_c$ is orthogonal. In light of this fact, Healy and Westmacott (1956) proposed an iterative procedure to compute the ordinary least squares estimator $\boldsymbol{\beta}_{OLS}$ of $\boldsymbol{\beta}$. In each iteration, their procedure imputes the values of $\mathbf{Y}_{\text{miss}}$ and updates the closed-form ordinary least squares estimator associated with the complete data. The OEM algorithm follows the same idea but solves (2) with the SCAD penalty. If $\mathbf{Y}_{\text{miss}}$ were observable, then $\boldsymbol{X}$ in (2) and $\mathbf{Y}$ can be replaced by $\boldsymbol{X}_c$ and $\mathbf{Y}_c$, yielding a closed-form solution to (2). Much beyond this orthogonal design formulation, we will significantly broaden the applicability of the algorithm in Section 3 by introducing an idea, called active orthogonalization, to actively expand any regression matrix into an orthogonal matrix.

Define

$$\boldsymbol{A} = \boldsymbol{\Delta}'\boldsymbol{\Delta}. \tag{8}$$

Let $(d_1, \ldots, d_p)$ denote the diagonal elements of $\boldsymbol{X}'_c\boldsymbol{X}_c$. The OEM algorithm for solving the

optimization problem in (2) proceeds as follows. Let $\boldsymbol{\beta}^{(0)}$ be an initial estimate of $\boldsymbol{\beta}$. For $k = 0, 1, \ldots$, impute $\mathbf{Y}_{\text{miss}}$ as $\mathbf{Y}_I = \boldsymbol{\Delta}\boldsymbol{\beta}^{(k)}$, let $\mathbf{Y}_c = (\mathbf{Y}',\ \mathbf{Y}_I')'$, and solve

$$\boldsymbol{\beta}^{(k+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\|\mathbf{Y}_c - \boldsymbol{X}_c\boldsymbol{\beta}\|^2 + 2\sum_{j=1}^{p} P_\lambda(|\beta_j|) \right] \tag{9}$$

until $\{\boldsymbol{\beta}^{(k)}\}$ converges. Letting

$$\mathbf{u} = (u_1, \ldots, u_p)' = \boldsymbol{X}'\mathbf{Y} + \boldsymbol{A}\boldsymbol{\beta}^{(k)}, \tag{10}$$

(9) becomes

$$\boldsymbol{\beta}^{(k+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\sum_{j=1}^{p} (d_j\beta_j^2 - 2u_j\beta_j) + 2\sum_{j=1}^{p} P_\lambda(|\beta_j|) \right], \tag{11}$$

which is *separable* in the dimensions of $\boldsymbol{\beta}$. If $\boldsymbol{X}$ in (1) is standardized as in (4) with $d_j \geqslant 1$ for all j, (11) has a closed-form

$$\beta_j^{(k+1)} = \begin{cases} \operatorname{sign}(u_j)\big(|u_j| - \lambda\big)_+/d_j, & \text{when } |u_j| \leqslant (d_j+1)\lambda, \\ \operatorname{sign}(u_j)\big[(a-1)|u_j| - a\lambda\big]/\big[(a-1)d_j - 1\big], & \text{when } (d_j+1)\lambda < |u_j| \leqslant a\lambda d_j, \\ u_j/d_j, & \text{when } |u_j| > a\lambda d_j. \end{cases} \tag{12}$$

As pointed out in Dempster, Laird, and Rubin (1977) that the Healy-Westmacott procedure is essentially an EM algorithm, OEM is an EM algorithm as well. The complete data $\mathbf{Y}_c = (\mathbf{Y}',\ \mathbf{Y}'_{\text{miss}})'$ in (7) follow a regression model $\mathbf{Y}_c = \boldsymbol{X}_c\boldsymbol{\beta} + \boldsymbol{\varepsilon}_c$, where $\boldsymbol{\varepsilon}_c$ is from $N(\mathbf{0}_m,\ \boldsymbol{I}_m)$. Let $\boldsymbol{\beta}_{\text{SCAD}}$ be a solution to (2), where $\boldsymbol{\beta}_{\text{SCAD}} = \operatorname{argmax}_{\boldsymbol{\beta}} L(\boldsymbol{\beta} \mid \mathbf{Y})$ and the penalized likelihood function $L(\boldsymbol{\beta} \mid \mathbf{Y})$ is

$$(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\|\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\right) \exp\left[-\sum_{j=1}^{p} P_\lambda(|\beta_j|)\right].$$

Given $\boldsymbol{\beta}^{(k)}$, the E-step of the OEM algorithm for the SCAD is

$$\begin{aligned} & E\big[\log\{L(\boldsymbol{\beta}|\mathbf{Y}_c)\} \mid \mathbf{Y}, \boldsymbol{\beta}^{(k)}\big] \\ = \; & -C\big\{\|\mathbf{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + E\big(\|\mathbf{Y}_{\text{miss}}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \mid \boldsymbol{\beta}^{(k)}\big) + 2\sum_{j=1}^{p} P_\lambda(|\beta_j|)\big\} \\ = \; & -C\big\{n + \|\mathbf{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \|\boldsymbol{\Delta}\boldsymbol{\beta}^{(k)}-\boldsymbol{\Delta}\boldsymbol{\beta}\|^2 + 2\sum_{j=1}^{p} P_\lambda(|\beta_j|)\big\} \end{aligned}$$

for some constant $C > 0$. Define

$$Q_{\text{SCAD}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(k)}) = \|\mathbf{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \|\boldsymbol{\Delta}\boldsymbol{\beta}^{(k)}-\boldsymbol{\Delta}\boldsymbol{\beta}\|^2 + 2\sum_{j=1}^{p} P_\lambda(|\beta_j|). \tag{13}$$

The M-step of the OEM algorithm for the SCAD is

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} Q_{\text{SCAD}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(k)}),$$

which is equivalent to (11).

For the general penalized regression problem in (5), the M-step of the OEM algorithm becomes

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}\in\Theta} Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(k)}), \tag{14}$$

where Q replaces Q_{SCAD} in (13) for the corresponding penalty function. If Θ and P in (5) are *decomposable* as $\Theta = \prod_{j=1}^{p}\Theta_j$ and $P(\boldsymbol{\beta};\lambda) = \sum_{j=1}^{p} P_j(\beta_j;\lambda)$, similarly to (11), (14) reduces to p one-dimensional problems

$$\beta_j^{(k+1)} = \underset{\beta_j\in\Theta_j}{\arg\min}\, \big[d_j\beta_j^2 - 2u_j\beta_j + P_j(\beta_j;\lambda)\big], \text{ for } j = 1,\ldots,p, \tag{15}$$

with $\mathbf{u} = (u_1,\ldots,u_p)'$ defined in the same way as in (10). This shortcut applies to the following penalties:

1. The lasso (Tibshirani 1996), where $\Theta_j = \mathbb{R}$,

$$P_j(\beta_j; \lambda) = 2\lambda|\beta_j|, \tag{16}$$

and (15) becomes

$$\beta_j^{(k+1)} = \text{sign}(u_j)\left(\frac{|u_j| - \lambda}{d_j}\right)_+. \tag{17}$$

Here, for $a \in \mathbb{R}$, $(a)_+$ denotes $\max\{a, 0\}$.

2. The nonnegative garrote (Breiman 1995), where $\Theta_j = \{x : x\hat{\beta}_j \geqslant 0\}$, $P_j(\beta_j; \lambda) = 2\lambda\beta_j/\hat{\beta}_j$, $\hat{\beta}_j$ is the ordinary least squares estimator of β_j, and (15) becomes

$$\beta_j^{(k+1)} = \left(\frac{u_j\hat{\beta}_j - \lambda}{d_j\hat{\beta}_j^2}\right)_+ \hat{\beta}_j.$$

3. The elastic-net (Zou and Hastie 2005), where $\Theta_j = \mathbb{R}$,

$$P_j(\beta_j; \lambda) = 2\lambda_1|\beta_j| + \lambda_2\beta_j^2. \tag{18}$$

and (15) becomes

$$\beta_j^{(k+1)} = \text{sign}(u_j)\left(\frac{|u_j| - \lambda_1}{d_j + \lambda_2}\right)_+. \tag{19}$$

5. The MCP (Zhang 2010), where $\Theta_j = \mathbb{R}$, $P_j(\beta_j; \lambda) = 2P_\lambda(|\beta_j|)$, and

$$P'_\lambda(\theta) = (\lambda - \theta/a)I(\theta \leqslant a\lambda) \tag{20}$$

with $a > 1$ and $\theta > 0$. If $\boldsymbol{X}$ in (1) is standardized as in (4) with $d_j \geqslant 1$ for all j, (15) becomes

$$\beta_j^{(k+1)} = \begin{cases} \text{sign}(u_j)a\big(|u_j| - \lambda\big)_+/(ad_j - 1), & \text{when } |u_j| \leqslant a\lambda d_j, \\ u_j/d_j, & \text{when } |u_j| > a\lambda d_j \end{cases} \tag{21}$$

6. The "Berhu" penalty (Owen 2006), where $\Theta_j = \mathbb{R}$, $P_j(\beta_j;\lambda) = 2\lambda\big\{|\beta_j| I(|\beta_j| < \delta) + (\beta_j^2+\delta^2) I(|\beta_j| \geqslant \delta)/(2\delta)\big\}$ for some $\delta > 0$, and (15) becomes

$$\beta_j^{(k+1)} = \begin{cases} \text{sign}(u_j)\big(|u_j| - \lambda\big)_+/d_j, & \text{when } |u_j| < \lambda + d_j\delta, \\ u_j\delta/(\lambda + d_j\delta), & \text{when } |u_j| \geqslant \lambda + d_j\delta. \end{cases}$$

Obviously, if the penalty on $\boldsymbol{\beta}$ disappears, the OEM algorithm reduces to the Healy-Westmacott procedure. Quite interestingly, Theorem 6 in Section 5 shows that, for the same $\boldsymbol{X}$ and $\mathbf{Y}$ in (1), the OEM algorithm for the elastic-net and lasso numerically converges faster than the Healy-Westmacott procedure.

Example 1. For the model in (1), let the complete matrix $\boldsymbol{X}_c$ be a fractional factorial design from Xu (2009) with 4096 runs in 30 factors. Clearly, $\boldsymbol{X}_c$ is an orthogonal design. Let $\boldsymbol{X}$ in (1) be the submatrix of $\boldsymbol{X}_c$ consisting of the first 3000 rows and let $\mathbf{Y}$ be generated with $\sigma = 1$ and

$$\beta_j = (-1)^j \exp\big[-2(j-1)/20\big] \text{ for } j = 1, \ldots, p. \tag{22}$$

Here, $p = 30$ and $n = 3000$. Assume the response values corresponding to the last 1096 rows of $\boldsymbol{X}_c$ are missing. We used the OEM algorithm to solve the optimization problem in (2) with an initial value $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ and a criterion to stop when relative changes in all coefficients are less than 10^{-6}. For $\lambda = 1$ and $a = 3.7$ in (3), Figure 1 plots values of the objective function in (2) of the OEM sequence against iteration numbers, where the algorithm converges at iteration 13.

3 THE GENERAL FORMULATION WITH ACTIVE ORTHOGONALIZATION

The OEM algorithm in Section 2 was derived for a missing-data problem where $\boldsymbol{X}$ in (1) is imbedded in a *pre-specified* orthogonal matrix. We drop this assumption in this section

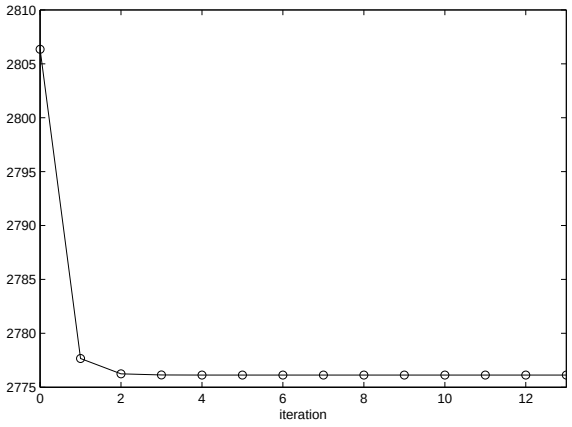

Figure 1. Values of the objective function of an OEM sequence for the SCAD against iterations for Example 1.

and further develop the algorithm for general data structures by introducing a procedure to *actively* expand an arbitrary matrix to an orthogonal matrix. The general idea of augmenting extra data has been used for EM problems before. For example, for a covariance estimation problem in Rubin and Szatrowski (1982), extra data are added elaborately to make the maximum likelihood estimator of the expanded patterned covariance matrices have an explicit form. To facilitate the use of the OEM algorithm in Section 2, the contribution here is to develop a scheme to *orthogonalize* the matrix $\boldsymbol{X}$ with an arbitrary structure.

Take $\boldsymbol{S}$ to be a $p \times p$ diagonal matrix with non-zero diagonal elements $s_1, \ldots, s_p$. Define

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{S}^{-1}. \tag{23}$$

The eigenvalue decomposition of $\boldsymbol{Z}'\boldsymbol{Z}$ (Wilkinson 1965) is

$$\boldsymbol{V}'\boldsymbol{\Gamma}\boldsymbol{V},$$

where $\boldsymbol{V}$ is an orthogonal matrix and $\boldsymbol{\Gamma}$ is a diagonal matrix whose diagonal elements, $\gamma_1 \geqslant \cdots \geqslant \gamma_p$, are the nonnegative eigenvalues of $\boldsymbol{Z}'\boldsymbol{Z}$. Let

$$t = \#\{j : \ \gamma_j = \gamma_1, \ j = 1, \ldots, p\} \tag{24}$$

denote the number of the γ_j equal to γ_1. For example, if $\gamma_1 = \gamma_2$ and $\gamma_1 > \gamma_j$ for $j = 3, \ldots, p$, then $t = 2$. Define

$$\boldsymbol{B} = \mathrm{diag}(\gamma_1 - \gamma_{t+1}, \ldots, \gamma_1 - \gamma_p) \tag{25}$$

and

$$\boldsymbol{\Delta} = \boldsymbol{B}^{1/2}\boldsymbol{V}_1\boldsymbol{S}, \tag{26}$$

where $\boldsymbol{V}_1$ is the submatrix of $\boldsymbol{V}$ consisting of the last $p - t$ rows. Let $\boldsymbol{X}_c$ be the augmented matrix of $\boldsymbol{\Delta}$ and $\boldsymbol{X}$.

Lemma 1. The matrix $\boldsymbol{X}_c$ constructed above is orthogonal.

Proof. Note that

$$\boldsymbol{X}_c'\boldsymbol{X}_c = \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}'\boldsymbol{\Delta},$$

which, by plugging (25) and (26), becomes

$$\boldsymbol{S}[\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{V}'(\gamma_1\boldsymbol{I}_p - \boldsymbol{\Gamma})\boldsymbol{V}]\boldsymbol{S} = \gamma_1\boldsymbol{S}^2. \tag{27}$$

Now, because

$$\gamma_1\boldsymbol{I}_p - \boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{pmatrix},$$

(27) is orthogonal, which completes the proof. □

The underlying geometry of the active orthogolization in Lemma 1 can be described as follows. For a vector $\mathbf{x} \in \mathbb{R}^m$, let $P_\omega\mathbf{x}$ denote its projection onto a subspace ω of $\mathbb{R}^m$. This lemma implies that for the column vectors of $\boldsymbol{X}$ in (1), $\mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathbb{R}^n$, there exists a set of mutually orthogonal vectors $\mathbf{x}_{c1}, \ldots, \mathbf{x}_{cp} \in \mathbb{R}^{n+p-t}$, essentially the column vectors of $\boldsymbol{X}_c$ in (6), satisfy the condition that $P_{\mathbb{R}^n}\mathbf{x}_{ci} = \mathbf{x}_i$, for $j = 1, \ldots, p$. Proposition 1 makes this precise.

Proposition 1. Let ω be an n-dimensional subspace of $\mathbb{R}^m$ with $n \leqslant m$. If $p \leqslant m - n + 1$, then for any p vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p \in \omega$, there exist p vectors $\mathbf{x}_{c1}, \ldots, \mathbf{x}_{cp} \in \mathbb{R}^m$ such that $P_\omega\mathbf{x}_{ci} = \mathbf{x}_i$ for $j = 1, \ldots, p$ and $\mathbf{x}_{ci}'\mathbf{x}_{cj} = 0$ for $i \neq j$.

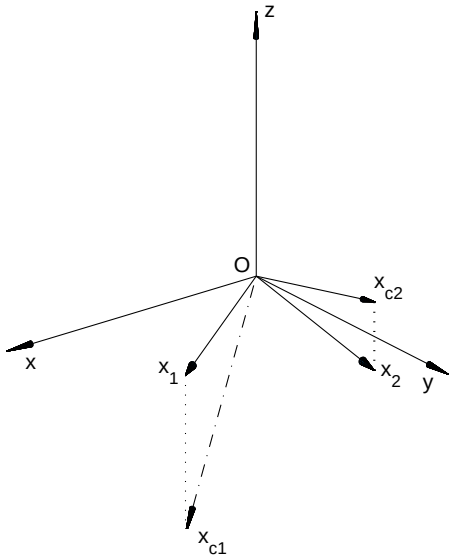

Figure 2. Expand two two-dimensional vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ to two three-dimensional vectors $\mathbf{x}_{c1}$ and $\mathbf{x}_{c2}$ with $\mathbf{x}_{c1}'\mathbf{x}_{c2} = 0$.

Figure 2 expands two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^2$ to two orthogonal vectors $\mathbf{x}_{c1}$ and $\mathbf{x}_{c2}$ in $\mathbb{R}^3$.

Now, if $\boldsymbol{X}_c$ from Lemma 1 is treated as the complete matrix defined in (6), the OEM algorithm in Section 2 follows through immediately.

When using the OEM algorithm to solve (5), in (10) instead of computing $\boldsymbol{\Delta}$ in (26), one may compute $\boldsymbol{A} = \boldsymbol{\Delta}'\boldsymbol{\Delta}$ and the diagonal entries $d_1, \ldots, d_p$ of $\boldsymbol{X}_c'\boldsymbol{X}_c$. Note that

$$\boldsymbol{A} = \gamma_1 \boldsymbol{S}^2 - \boldsymbol{X}'\boldsymbol{X} \tag{28}$$

and

$$d_j = \gamma_1 s_j^2 \text{ for } j = 1, \ldots, p, \tag{29}$$

where $\boldsymbol{S}$ and $\boldsymbol{Z}$ are defined in (23) and γ_1 is the largest eigenvalue of $\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{S}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{S}^{-1}$. One way to compute γ_1 is to use the power method (Wilkinson 1965) described below. Given a nonzero initial vector $\mathbf{a}^{(0)} \in \mathbb{R}^p$, let $\gamma_1^{(0)} = \|\mathbf{a}^{(0)}\|$. For $k = 0, 1, ...,$ compute $\mathbf{a}^{(k+1)} = \boldsymbol{X}'\boldsymbol{X}\mathbf{a}^{(k)}/\gamma_1^{(k)}$ and $\gamma_1^{(k+1)} = \|\mathbf{a}^{(k+1)}\|$ until convergence. If $\mathbf{a}^{(0)}$ is not an eigenvector

of any γ_j that does not equal γ_1, then $\gamma_1^{(k)}$ converges to γ_1. For t defined in (24), the convergence rate is linear (Watkins 2002) specified by

$$\lim_{k\to\infty} \frac{\|\gamma_1^{(k+1)} - \gamma_1\|}{\|\gamma_1^{(k)} - \gamma_1\|} = \frac{\gamma_{t+1}}{\gamma_1}.$$

An easy way to make $\boldsymbol{A} = \boldsymbol{\Delta}'\boldsymbol{\Delta}$ in (28) positive definite is to replace $\boldsymbol{B}$ in (25) by

$$\boldsymbol{B} = \text{diag}(d - \gamma_{t+1}, \ldots, d - \gamma_p)$$

with $d \geqslant \gamma_1$, which changes (28) and (29) to

$$\boldsymbol{A} = d\boldsymbol{S}^2 - \boldsymbol{X}'\boldsymbol{X} \tag{30}$$

and

$$d_j = ds_j^2, \text{ for } j = 1, \ldots, p, \tag{31}$$

respectively. If $d > \gamma_1$, then $\boldsymbol{A} = \boldsymbol{\Delta}'\boldsymbol{\Delta}$ is positive definite.

Remark 1. The matrix $\boldsymbol{S}$ in (23) can be chosen flexibly. One possibility is to use $\boldsymbol{S} = \boldsymbol{I}_p$ so that

$$\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}'\boldsymbol{\Delta} = d\boldsymbol{I}_p \tag{32}$$

with $d \geqslant \gamma_1$, and $\boldsymbol{X}_c/\sqrt{d}$ is standardized as in (4).

Example 2. Suppose that $\boldsymbol{X}$ in (1) is orthogonal. Take

$$\boldsymbol{S} = \text{diag}\left[(\sum_{i=1}^{n} x_{i1}^2)^{1/2}, \ldots, (\sum_{i=1}^{n} x_{ip}^2)^{1/2}\right]. \tag{33}$$

Since $t = p$, $\boldsymbol{\Delta}$ in (26) is empty. This result indicates the active orthogonalization procedure will not overshoot: if $\boldsymbol{X}$ is orthogonal already, it adds no row.

Example 3. Let

$$\boldsymbol{X} = \begin{pmatrix} 0 & 0 & 3/2 \\ -4/3 & -2/3 & 1/6 \\ 2/3 & 4/3 & 1/6 \\ -2/3 & 2/3 & -7/6 \end{pmatrix}.$$

If $\boldsymbol{S} = \boldsymbol{I}_3$, (26) gives $\boldsymbol{\Delta} = (-2/\sqrt{3},\ 2/\sqrt{3},\ 1/\sqrt{3})$.

Example 4. Consider a two-level design in three factors, A, B and C:

$$\begin{pmatrix} -1 & -1 & -1 \\ -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}.$$

The regression matrix including all main effects and two-way interactions is

$$\boldsymbol{X} = \begin{pmatrix} -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 \end{pmatrix},$$

where the last three columns for the interactions are fully aliased with the first three columns for the main effects. For $\boldsymbol{S} = \boldsymbol{I}_3$, (26) gives

$$\boldsymbol{\Delta} = \begin{pmatrix} 0 & -2 & 0 & 0 & -2 & 0 \\ 0 & 0 & -2 & -2 & 0 & 0 \\ -2 & 0 & 0 & 0 & 0 & -2 \end{pmatrix}.$$

The elements in $\boldsymbol{\Delta}$ are chosen flexibly, not restricted to ± 1.

Example 5. Consider a 1000×10 random matrix $\boldsymbol{X} = (x_{ij})$ with entries independently

drawn from the uniform distribution on $[0, 1)$. Using $\boldsymbol{S}$ in (33), (26) gives

$$\boldsymbol{\Delta} = \begin{pmatrix} -7.99 & 16.06 & -6.39 & -18.26 & 12.91 & -8.67 & 7.56 & 34.08 & -17.04 & -11.81 \\ 26.83 & -12.09 & 7.91 & 1.02 & -22.75 & -6.90 & -19.98 & 26.10 & -0.86 & 0.88 \\ -4.01 & 1.48 & 9.51 & -21.99 & 19.46 & -10.27 & -25.12 & -3.39 & 7.29 & 27.90 \\ 21.77 & 10.72 & -0.61 & -6.46 & 28.00 & 1.28 & -6.86 & -7.04 & 11.13 & -30.64 \\ -15.78 & 5.60 & -15.26 & -7.67 & -9.76 & 23.93 & -14.71 & 12.25 & 29.45 & -7.89 \\ 16.34 & 10.61 & -41.82 & 11.82 & 6.49 & -7.38 & -6.14 & -1.82 & -1.86 & 13.09 \\ -8.15 & 24.97 & 12.11 & 24.35 & 3.66 & -2.59 & -27.84 & -3.45 & -9.40 & -13.72 \\ -5.35 & -21.70 & -4.16 & 7.42 & 13.98 & 29.84 & -10.26 & 7.60 & -25.13 & 7.78 \\ -19.62 & -22.43 & -2.61 & 22.58 & 11.80 & -22.08 & 1.25 & 15.87 & 14.94 & 0.31 \end{pmatrix}.$$

Only nine rows need to be added to make this large $\boldsymbol{X}$ matrix orthogonal.

4 ACHIEVING THE ORACLE PROPERTY WITH NONCONVEX PENALTIES

Fan and Li (2001) introduced an important concept called the oracle property and showed that there exists one local minimum of (2) with this property. However, because the optimization problem in (2) has an exponential number of local optima (Huo and Ni 2007; Huo and Chen 2010), no theoretical results in the literature claim that an existing algorithm can provide such a local minimum. In this section, we prove that an OEM sequence for the SCAD and MCP can indeed achieve this property. The theoretical results in this and the following sections work for the OEM algorithm in both Sections 2 and 3.

First, we describe the oracle concept. A penalized least squares estimator of $\boldsymbol{\beta}$ in (1) has this property if it can not only select the correct submodel asymptotically, but also estimate the nonzero coefficients $\boldsymbol{\beta}_1$ in (34) as efficiently as if the correct submodel were known in advance. Suppose that the number of nonzero coefficients of $\boldsymbol{\beta}$ in (1) is p_1 (with $p_1 \leqslant p$) and

partition $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')', \tag{34}$$

where $\boldsymbol{\beta}_2 = \mathbf{0}$ and no component of $\boldsymbol{\beta}_1$ is zero. Divide the regression matrix $\boldsymbol{X}$ in (1) to $(\boldsymbol{X}_1\ \boldsymbol{X}_2)$ with $\boldsymbol{X}_1$ corresponding to $\boldsymbol{\beta}_1$. If all the variables that influence the response in (1) are known, an *oracle* estimator can be given as $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')'$ with $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$, where

$$\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \sim N(\mathbf{0},\ \sigma^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}).$$

We need several assumptions.

Assumption 1. As $n \to \infty$,

$$\frac{\boldsymbol{X}'\boldsymbol{X}}{n} \to \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_2 \end{pmatrix},$$

where $\boldsymbol{\Sigma}$ is positive definite and $\boldsymbol{\Sigma}_1$ is $p_1 \times p_1$. Furthermore, $\boldsymbol{X}/\sqrt{n}$ is standardized such that each entry on the diagonal of $\boldsymbol{X}'\boldsymbol{X}/n$ is 1, and $\boldsymbol{X}'\boldsymbol{X}/n + \boldsymbol{\Delta}'\boldsymbol{\Delta} = d\boldsymbol{I}_p$ with $d \geqslant \gamma_1$, where $d = O(1)$ and γ_1 is the largest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}/n$.

Assumption 2. The tuning parameter $\lambda = \lambda_n$ in (3) satisfies the condition that, as $n \to \infty$, $\lambda_n/n \to 0$ and $\lambda_n/\sqrt{n} \to \infty$.

Let $\{\boldsymbol{\beta}^{(k)},\ k = 0, 1, \ldots, \}$ be the OEM sequence from (12) for the SCAD with a fixed $a > 2$ in (3). We need an assumption on $k = k_n$. Let η be the largest eigenvalue of $\boldsymbol{I}_{p_1} - \boldsymbol{X}_1'\boldsymbol{X}_1/(nd)$. Under Assumption 1, η tends to a limit lying between 0 and 1 as $n \to \infty$.

Assumption 3. As $n \to \infty$, $\eta^{k_n}\lambda_n/\sqrt{n} \to 0$ and $k_n^2 \exp(-c(\lambda_n/\sqrt{n})^2) \to 0$ for any $c > 0$.

One choice for k_n to satisfy Assumption 3 is

$$k_n = \left(\frac{\lambda_n}{\sqrt{n}}\right)^{\nu} \text{ for some } \nu > 0.$$

Under Assumption 2, $k_n \to \infty$ as $n \to \infty$.

Under the above assumptions, Theorem 1 shows that $\boldsymbol{\beta}^{(k)} = (\boldsymbol{\beta}_1^{(k)'}, \boldsymbol{\beta}_2^{(k)'})'$ can achieve the oracle property.

Theorem 1. Suppose that Assumption 1-3 hold. If $\boldsymbol{\beta}^{(0)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{Y}$, then as $n \to \infty$,
(i) $\mathrm{P}(\boldsymbol{\beta}_2^{(k)} = \mathbf{0}) \to 1$;
(ii) $\sqrt{n}(\boldsymbol{\beta}_1^{(k)} - \boldsymbol{\beta}_1) \to N(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}_1^{-1})$ in distribution.

The proof of Theorem 1 is deferred to the Appendix.

Remark 2. From (60) in the proof of Theorem 1, for $k = 1, 2, \ldots$, $\boldsymbol{\beta}^{(k)}$ is *consistent* in variable selection. That is, $\mathrm{P}(\beta_j^{(k)} \neq 0 \text{ for } j = 1, \ldots, p_1) \to 1$ and $\mathrm{P}(\boldsymbol{\beta}_2^{(k)} = \mathbf{0}) \to 1$ as $n \to \infty$.

Remark 3. The proof of Theorem 1 uses the convergence rates of $\mathrm{P}(A_k)$, $\mathrm{P}(B_k)$ and $\mathrm{P}(C_k^h)$. If an OEM sequence satisfies the condition that $\beta_j^{(k+1)} = 0$ when $|u_j| < \lambda$ and $\beta_j^{(k+1)} = u_j/d$ when $|u_j| > c\lambda$ for some $c = O(1)$, then $\mathrm{P}(A_{k+1}) = \mathrm{P}(|u_j| < \lambda)$ and $\mathrm{P}(B_{k+1}) = \mathrm{P}(|u_j| > c\lambda)$. Since an OEM sequence for the MCP satisfies the above condition, an argument very similar to the proof in the Appendix shows that the convergence rates of $\mathrm{P}(A_k)$, $\mathrm{P}(B_k)$ and $\mathrm{P}(C_k^h)$ for the MCP are the same as those with the SCAD. Thus, under Assumption 1-3, Theorem 1 holds for the MCP with a fixed $a > 1$ in (20).

Huo and Chen (2010) showed that, for the SCAD penalty, solving the global minimum of (5) leads to an NP-hard problem. Theorem 1 indicates that as far as the oracle property is concerned, the local solution given by OEM will suffice.

5 CONVERGENCE OF THE OEM ALGORITHM

In this section, we derive theoretical results on convergence properties of the OEM algorithm and compare the convergence rates of OEM for the ordinary least squares estimator and the elastic-net and lasso. The general penalty in (5) is considered here. Our derivations employs the main tool in Wu (1983) in conjunction with special properties of the penalties mentioned in Section 2.

We make several assumptions for Θ and $P(\boldsymbol{\beta};\lambda)$ in (5).

Assumption 4. The parameter space Θ is a closed convex subset of $\mathbb{R}^p$.

Assumption 5. For a fixed λ, the penalty $P(\boldsymbol{\beta};\lambda) \to +\infty$ as $\|\boldsymbol{\beta}\| \to +\infty$.

Assumption 6. For a fixed λ, the penalty $P(\boldsymbol{\beta};\lambda)$ is continuous with respect to $\boldsymbol{\beta} \in \Theta$.

All penalties discussed in Section 2 satisfy these assumptions. The assumptions cover the case in which the iterative sequence $\{\boldsymbol{\beta}^{(k)}\}$ defined in (14) may fall on the boundary of Θ (Nettleton 1999), like the nonnegative garrote (Breiman 1995) and the nonnegative lasso (Efron et al. 2004). The bridge penalty (Frank and Friedman 1993) in (37) also satisfies the above assumptions.

For the model in (1), denote the objective function in (5) by

$$l(\boldsymbol{\beta}) = \|\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + P(\boldsymbol{\beta};\lambda). \tag{35}$$

For some penalizes like the bridge, it may be numerically infeasible to perform the M-step in (14). For this situation, following the generalized EM algorithm in Dempster, Laird, and Rubin (1977), we define a *generalized OEM* algorithm to be an iterative scheme

$$\boldsymbol{\beta}^{(k)} \to \boldsymbol{\beta}^{(k+1)} \in \mathcal{M}(\boldsymbol{\beta}^{(k)}), \tag{36}$$

where $\boldsymbol{\beta} \to \mathcal{M}(\boldsymbol{\beta}) \subset \Theta$ is a point-to-set map such that

$$Q(\boldsymbol{\phi} \mid \boldsymbol{\beta}) \leqslant Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}), \quad \text{for all } \boldsymbol{\phi} \in \mathcal{M}(\boldsymbol{\beta}).$$

Here, Q is given by replacing the SCAD with $P(\boldsymbol{\beta};\lambda)$ in (13). The OEM sequence defined by (14) is a special case of (36). For example, the generalized OEM algorithm can be used for the bridge penalty, where $\Theta_j = \mathbb{R}$ and

$$P_j(\beta_j;\lambda) = \lambda|\beta_j|^a \tag{37}$$

for some $a \in (0, 1)$ in (5). Since the solution to (15) with the bridge penalty has no closed form, one may use one-dimensional search to compute $\beta_j^{(k+1)}$ that satisfies (36). By Assumption 1, $\{\boldsymbol{\beta} \in \Theta : l(\boldsymbol{\beta}) \leqslant l(\boldsymbol{\beta}^{(0)})\}$ is compact for any $l(\boldsymbol{\beta}^{(0)}) > -\infty$. By Assumption 6, $\mathcal{M}$ is a closed point-to-set map (Zangwill 1969; Wu 1983).

The objective functions of the EM algorithms in the literature like those discussed in Wu (1983), Green (1990) and McLachlan and Krishnan (2008) are typically continuously differentiable. This condition does not hold for the objective function in (5) with the lasso and other penalties. A more general definition of stationary points is needed here. We call $\boldsymbol{\beta} \in \Theta$ a stationary point of l if

$$\liminf_{t \to 0_+} \frac{l\big((1-t)\boldsymbol{\beta} + t\boldsymbol{\phi}\big) - l(\boldsymbol{\beta})}{t} \geqslant 0 \quad \text{for all } \boldsymbol{\phi} \in \Theta.$$

Let S denote the set of stationary points of l. Analogous to Theorem 1 in Wu (1983) on the global convergence of the EM algorithm, we have the following result.

Theorem 2. Let $\{\boldsymbol{\beta}^{(k)}\}$ be a generalized OEM sequence generated by (36). Suppose that

$$l(\boldsymbol{\beta}^{(k+1)}) < l(\boldsymbol{\beta}^{(k)}) \quad \text{for all} \quad \boldsymbol{\beta}^{(k)} \in \Theta \setminus S. \tag{38}$$

Then all limit points of $\{\boldsymbol{\beta}^{(k)}\}$ are elements of S and $l(\boldsymbol{\beta}^{(k)})$ converges monotonically to $l^* = l(\boldsymbol{\beta}^*)$ for some $\boldsymbol{\beta}^* \in S$.

Theorem 3. If $\boldsymbol{\beta}^*$ is a local minimum of $Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*)$, then $\boldsymbol{\beta}^* \in S$.

This theorem follows from the fact that $l(\boldsymbol{\beta}) - Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*)$ is differentiable and

$$\frac{\partial\big[l(\boldsymbol{\beta}) - Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*)\big]}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = 0.$$

Remark 4. By Theorem 3, if $\boldsymbol{\beta}^{(k)} \notin S$, then $\boldsymbol{\beta}^{(k)}$ cannot be a local minimum of $Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(k)})$. Thus, there exists at least one point $\boldsymbol{\beta}^{(k+1)} \in \mathcal{M}(\boldsymbol{\beta}^{(k)})$ such that $Q(\boldsymbol{\beta}^{(k+1)} \mid \boldsymbol{\beta}^{(k)}) < Q(\boldsymbol{\beta}^{(k)} \mid \boldsymbol{\beta}^{(k)})$ and therefore satisfies the condition in (38). As a special case, an OEM sequence generated by (14) satisfies (38) in Theorem 2.

Next, we consider the convergence of a generalized OEM sequence $\{\boldsymbol{\beta}^{(k)}\}$ in (36). By Theorem 3, such results will automatically hold for an OEM sequence as well. If the penalty function $P(\boldsymbol{\beta};\lambda)$ is convex and $l(\boldsymbol{\beta})$ has a unique minimum, Theorem 4 shows that $\{\boldsymbol{\beta}^{(k)}\}$ converges to the global minimum.

Theorem 4. Let $\{\boldsymbol{\beta}^{(k)}\}$ be defined in Theorem 2. Suppose that $l(\boldsymbol{\beta})$ in (35) is a convex function on Θ with a unique minimum $\boldsymbol{\beta}^*$ and that (38) holds for $\{\boldsymbol{\beta}^{(k)}\}$. Then $\boldsymbol{\beta}^{(k)} \to \boldsymbol{\beta}^*$ as $k \to \infty$.

Proof. We only need to show that $S = \{\boldsymbol{\beta}^*\}$. For $\boldsymbol{\phi} \in \Theta$ with $\boldsymbol{\phi} \neq \boldsymbol{\beta}^*$ and $t > 0$, we have

$$\frac{l\big((1-t)\boldsymbol{\phi} + t\boldsymbol{\beta}^*\big) - l(\boldsymbol{\beta}^*)}{t} \leqslant \frac{tl(\boldsymbol{\beta}^*) + (1-t)l(\boldsymbol{\phi}) - l(\boldsymbol{\phi})}{t} = l(\boldsymbol{\beta}^*) - l(\boldsymbol{\phi}) < 0.$$

This implies $\boldsymbol{\phi} \notin S$. □

Theorem 5 discusses the convergence of an OEM sequence $\{\boldsymbol{\beta}^{(k)}\}$ for more general penalties. For $a \in \mathbb{R}$, define $S(a) = \{\boldsymbol{\phi} \in S : \ l(\boldsymbol{\phi}) = a\}$. From Theorem 2, all limit points of an OEM sequence are in $S(l^*)$, where l^* is the limit of $l(\boldsymbol{\beta}^{(k)})$ in Theorem 2. Theorem 5 states that the limit point is unique under certain conditions.

Theorem 5. Let $\{\boldsymbol{\beta}^{(k)}\}$ be a generalized OEM sequence generated by (36) with $\boldsymbol{\Delta}'\boldsymbol{\Delta} > 0$. If (38) holds, then all limit points of $\{\boldsymbol{\beta}^{(k)}\}$ are in a connected and compact subset of $S(l^*)$. In particular, if the set $S(l^*)$ is discrete in that its only connected components are singletons, then $\boldsymbol{\beta}^{(k)}$ converges to some $\boldsymbol{\beta}^*$ in $S(l^*)$ as $k \to \infty$.

Proof. Note that $Q(\boldsymbol{\beta}^{(k+1)} \mid \boldsymbol{\beta}^{(k)}) = l(\boldsymbol{\beta}^{(k+1)}) + \|\boldsymbol{\Delta}\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\Delta}\boldsymbol{\beta}^{(k)}\|^2 \leqslant Q(\boldsymbol{\beta}^{(k)} \mid \boldsymbol{\beta}^{(k)}) = l(\boldsymbol{\beta}^{(k)})$. By Theorem 2, $\|\boldsymbol{\Delta}\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\Delta}\boldsymbol{\beta}^{(k)}\|^2 \leqslant l(\boldsymbol{\beta}^{(k)}) - l(\boldsymbol{\beta}^{(k+1)}) \to 0$ as $k \to \infty$. Thus, $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\| \to 0$. This theorem now follows immediately from Theorem 5 of Wu (1983). □

Since the bridge, SCAD and MCP penalties all satisfy the condition that $S(l^*)$ is discrete, an OEM sequence for any of them converges to the stationary points of l. Theorem 5 is obtained under the condition $\boldsymbol{\Delta}'\boldsymbol{\Delta} > 0$. Since the error $\boldsymbol{\varepsilon}$ in (1) has a continuous distribution,

it is easy to show that Theorem 5 holds with probability one if $\boldsymbol{\Delta}'\boldsymbol{\Delta}$ is singular when d defined in (30) and (31) equals γ_1.

We now derive the convergence rate of the OEM sequence in (14). Following Dempster, Laird, and Rubin (1977), write

$$\boldsymbol{\beta}^{(k+1)} = \mathbf{M}(\boldsymbol{\beta}^{(k)}),$$

where the map $\mathbf{M}(\boldsymbol{\beta}) = (M_1(\boldsymbol{\beta}), \ldots, M_p(\boldsymbol{\beta}))'$ is defined by (14). We capture the convergence rate of the OEM sequence $\{\boldsymbol{\beta}^{(k)}\}$ through $\mathbf{M}$. Assume that (32) holds for $d \geqslant \gamma_1$, where γ_1 is the largest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$. For the active orthogolization in (30) and (31), taking $\boldsymbol{S} = \boldsymbol{I}_p$ satisfies this assumption; see Remark 1.

Let $\boldsymbol{\beta}^*$ be the limit of the OEM sequence $\{\boldsymbol{\beta}^{(k)}\}$. As in Meng (1994), we call

$$R = \limsup_{k\to\infty} \frac{\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^*\|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} = \limsup_{k\to\infty} \frac{\|\mathbf{M}(\boldsymbol{\beta}_k) - \mathbf{M}(\boldsymbol{\beta}^*)\|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} \tag{39}$$

the global rate of convergence for the OEM sequence. If there is no penalty in (5), i.e., computing the ordinary least squares estimator, the global rate of convergence R in (39) becomes the largest eigenvalue of $\boldsymbol{J}(\boldsymbol{\beta}^*)$, denoted by R_0, where $\boldsymbol{J}(\boldsymbol{\phi})$ is the $p \times p$ Jacobian matrix for $\mathbf{M}(\boldsymbol{\phi})$ having (i,j)th entry $\partial M_i(\boldsymbol{\phi})/\partial \boldsymbol{\phi}_j$. If (32) holds, then $\boldsymbol{J}(\boldsymbol{\beta}^*) = \boldsymbol{A}/d$ with $\boldsymbol{A} = \boldsymbol{\Delta}'\boldsymbol{\Delta}$. Thus,

$$R_0 = \frac{d - \gamma_p}{d}. \tag{40}$$

For (5), the penalty function $P(\boldsymbol{\beta}; \lambda)$ typically is not sufficiently smooth and R in (39) does not have an analytic form. Theorem 6 gives an upper bound of R_{net}, the value of R for the elastic-net penalty in (18) with $\lambda_1, \lambda_2 \geqslant 0$.

Theorem 6. For $\boldsymbol{\Delta}$ from (6), if (32) holds, then $R_{\text{NET}} \leqslant R_0$.

Proof. Let $\mathbf{x}_j$ denote the jth column of $\boldsymbol{X}$ and $\mathbf{a}_j$ denote the jth column of $\boldsymbol{A} = \boldsymbol{\Delta}'\boldsymbol{\Delta}$, respectively. For an OEM sequence for the elastic-net, by (19),

$$M_j(\boldsymbol{\beta}) = f(\mathbf{x}_j'\mathbf{Y} + \mathbf{a}_j'\boldsymbol{\beta}), \text{ for } j = 1, \ldots, p,$$

where

$$f(u) = \text{sign}(u) \left(\frac{|u| - \lambda_1}{d + \lambda_2} \right)_+ .$$

For $j = 1, \ldots, p$, observe that

$$\begin{aligned} \frac{|M_j(\boldsymbol{\beta}^{(k)}) - M_j(\boldsymbol{\beta}^*)|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} &= \frac{|f(\mathbf{x}_j'\mathbf{Y} + \mathbf{a}_j'\boldsymbol{\beta}^{(k)}) - f(\mathbf{x}_j'\mathbf{Y} + \mathbf{a}_j'\boldsymbol{\beta}^*)|}{|(\mathbf{x}_j'\mathbf{Y} + \mathbf{a}_j'\boldsymbol{\beta}^{(k)}) - (\mathbf{x}_j'\mathbf{Y} + \mathbf{a}_j'\boldsymbol{\beta}^*)|} \\ &\quad \cdot \frac{|(\mathbf{x}_j'\mathbf{Y} + \mathbf{a}_j'\boldsymbol{\beta}^{(k)}) - (\mathbf{x}_j'\mathbf{Y} + \mathbf{a}_j'\boldsymbol{\beta}^*)|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} \\ &\leqslant \frac{1}{d} \cdot \frac{|\mathbf{a}_j'(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*)|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|}. \end{aligned}$$

Thus,

$$\frac{\|M(\boldsymbol{\beta}^{(k)}) - M(\boldsymbol{\beta}^*)\|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} \leqslant \frac{1}{d} \cdot \frac{\|A(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*)\|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} \leqslant \frac{d - \gamma_p}{d}.$$

This completes the proof. □

Remark 5. Theorem 6 indicates that, for the same $\boldsymbol{X}$ and $\mathbf{Y}$ in (1), the OEM solution for the elastic-net converges faster than its counterpart for the ordinary least squares. Since the lasso is a special case of the elastic-net with $\lambda_2 = 0$ in (18), this theorem holds for the lasso as well.

Remark 6. From (40) and Theorem 6, the convergence rate of the OEM algorithm depends on the ratio of the smallest eigenvalue, γ_p, and the largest eigenvalue, γ_1, of $\boldsymbol{X}'\boldsymbol{X}$. This rate is the fastest when $\gamma_1 = \gamma_p$, i.e., if $\boldsymbol{X}$ is orthogonal and standardized. This result suggests that OEM converges faster if $\boldsymbol{X}$ is orthogonal or nearly orthogonal like from a supersaturated design or a nearly orthogonal Latin hypercube design (Owen 1994; Tang 1998). This result is in agreement with the recent finding in the design of experiments community that the use of orthogonal or nearly orthogonal designs can significantly improve the accuracy of penalized variable selection methods (Phoa, Pan, and Xu 2009; Deng, Lin, and Qian 2010; Zhu 2011).

Example 6. We generate $\boldsymbol{X}$ from p dimensional Gaussian distribution $N(\mathbf{0}, \boldsymbol{V})$ with n independent observations, where the (i, j)th entry of $\boldsymbol{V}$ is 1 for $i = j$ and ρ for $i \neq j$. Values

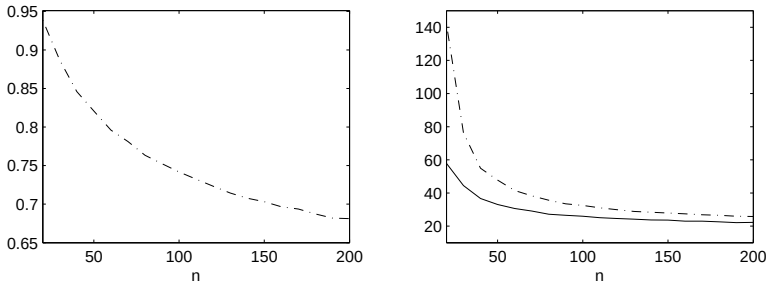

Figure 3. (Left) the average values of R_0 in (40) against increasing n for Example 6; (right) the average iteration numbers against increasing n for Example 6, where the dashed and solid lines denote the ordinary least squares estimator and the lasso, respectively.

of $\mathbf{Y}$ and $\boldsymbol{\beta}$ are generated by (1) and (22). The same setup was used in Friedman, Hastie, and Tibshirani (2009). For $p = 10,\ \rho = 0.1,\ \lambda = 0.5$ and increasing n, the left panel of Figure 3 depicts the average values of R_0 in (40) against increasing n and the right panel of the figure depicts the average iteration numbers against increasing n, with the dashed and solid lines corresponding to the ordinary least squares estimator and the lasso, respectively. Quite strikingly, this figure indicates that OEM requires *fewer* iterations as n becomes larger, which makes OEM particulary attractive for situations with massive data (SAMSI 2012). It is important to note that here the OEM sequence for the lasso requires fewer iteration than its counterpart for the ordinary least squares, empirically validating Theorem 6.

6 POSSESSING GROUPING COHERENCE

In this section, we consider the convergence of the OEM algorithm when the regression matrix $\boldsymbol{X}$ in (1) is singular due to fully aliased columns or other conditions. Let $\boldsymbol{X}$ be standardized as in (4) with columns $\mathbf{x}_1, \ldots, \mathbf{x}_p$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are fully aliased, i.e., $|\mathbf{x}_i| = |\mathbf{x}_2|$, then the objective function in (5) for the lasso is not strictly convex and has many minima (Zou and Hastie 2005). Data with fully aliasing structures commonly appear in observational studies and various classes of experimental designs like supersaturated designs (Wu 1993; Lin 1993; Tang and Wu 1997; Li and Lin 2002; Xu and Wu 2005) and factorial designs (Dey

and Mukerjee 1999; Mukerjee and Wu 2006).

Zou and Hastie (2005) states that if some columns of $\boldsymbol{X}$ are identical, it is desirable to have grouping coherence by assigning the same coefficient to them. Definition 1 makes this precise.

Definition 1. An estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ of $\boldsymbol{\beta}$ in (1) has *grouping coherence* if $\mathbf{x}_i = \mathbf{x}_j$ implies $\hat{\beta}_i = \hat{\beta}_j$ and $\mathbf{x}_i = -\mathbf{x}_j$ implies $\hat{\beta}_i = -\hat{\beta}_j$.

Let $\mathbf{0}_p$ denote the zero vector in $\mathbb{R}^p$. Let $\mathbf{e}_{ij}^+$ be the vector obtained by replacing the ith and jth entries of $\mathbf{0}_p$ with 1. Let $\mathbf{e}_{ij}^-$ be the vector obtained by replacing the ith and jth entries of $\mathbf{0}_p$ with 1 and -1, respectively. Let $\mathcal{E}$ denote the set of all $\mathbf{e}_{ij}^+$ and $\mathbf{e}_{ij}^-$. By Definition 1, an estimator $\hat{\boldsymbol{\beta}}$ has grouping coherence if and only if for any $\boldsymbol{\alpha} \in \mathcal{E}$ with $\boldsymbol{X}\boldsymbol{\alpha} = \mathbf{0}$, $\boldsymbol{\alpha}'\hat{\boldsymbol{\beta}} = 0$.

Lemma 2. Suppose that (32) holds. For the OEM sequence $\{\boldsymbol{\beta}^{(k)}\}$ of the lasso, SCAD or MCP, if $\boldsymbol{X}\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\alpha}'\boldsymbol{\beta}^{(k)} = 0$ for $\boldsymbol{\alpha} \in \mathcal{E}$, then $\boldsymbol{\alpha}'\boldsymbol{\beta}^{(k+1)} = 0$.

Proof. For $\mathbf{u}$ defined in (10), we have that $\boldsymbol{\alpha}'\mathbf{u} = \boldsymbol{\alpha}'\boldsymbol{X}'\mathbf{Y} + \boldsymbol{\alpha}'(d\boldsymbol{I}_p - \boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta}^{(k)} = 0$ for any $\boldsymbol{\alpha} \in \mathcal{E}$ with $\boldsymbol{X}\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\alpha}'\boldsymbol{\beta}^{(k)} = 0$. Then by (17), (12) and (21), an OEM sequence of the lasso, SCAD or MCP satisfies the condition that if $\boldsymbol{\alpha}'\mathbf{u} = 0$, then $\boldsymbol{\alpha}'\boldsymbol{\beta}^{(k+1)} = 0$ for $\boldsymbol{\alpha} \in \mathcal{E}$. This completes the proof. □

Remark 7. Lemma 2 implies that, for $k = 1, 2, \ldots$, $\boldsymbol{\beta}^{(k)}$ has grouping coherence if $\boldsymbol{\beta}^{(0)}$ has grouping coherence. Thus, if $\{\boldsymbol{\beta}^{(k)}\}$ converges, then its limit has grouping coherence. By Theorem 5, if $d > \lambda_1$ in (32), then an OEM sequence for the SCAD or MCP converges to a point with grouping coherence.

When $\boldsymbol{X}$ in (1) has fully aliased columns, the objective function in (5) for the lasso has many minima and hence the condition in Theorem 4 does not hold. Theorem 7 shows that, even with full aliasing, an OEM sequence (17) for the lasso converges to a point with grouping coherence.

Theorem 7. Suppose that (32) holds. If $\boldsymbol{\beta}^{(0)}$ has grouping coherence, then as $k \to \infty$, the OEM sequence $\{\boldsymbol{\beta}^{(k)}\}$ of the lasso converges to a limit that has grouping coherence.

Proof. Partition the matrix $\boldsymbol{X}$ in (1) as $(\boldsymbol{X}_1\ \boldsymbol{X}_2)$, where no two columns of $\boldsymbol{X}_2$ are fully aliased and any column of $\boldsymbol{X}_1$ is fully aliased with at least one column of $\boldsymbol{X}_2$. Let J denote the number of columns in $\boldsymbol{X}_1$. Partition $\boldsymbol{\beta}$ as $(\boldsymbol{\beta}_1',\ \boldsymbol{\beta}_2')'$ and $\boldsymbol{\beta}^{(k)}$ as $(\boldsymbol{\beta}_1^{(k)'},\ \boldsymbol{\beta}_2^{(k)'})'$, corresponding to $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively. For $j = 1, \ldots, p$, let

$$\omega(j) = \#\{i = 1, \ldots, p : |\mathbf{x}_i| = |\mathbf{x}_j|\}.$$

By Lemma 2, for $j = 1, \ldots, J$, $\beta_j^{(k)} = \beta_{j'}^{(k)}$ if $\mathbf{x}_j = \mathbf{x}_{j'}$ and $\beta_j^{(k)} = -\beta_{j'}^{(k)}$ otherwise, where $j' \in \{J+1, \ldots, p\}$. It follows that $\{\boldsymbol{\beta}_2^{(k)}\}$ can be viewed as an OEM sequence for solving

$$\min_{\boldsymbol{\theta}} \|\mathbf{Y} - \tilde{\boldsymbol{X}}\boldsymbol{\theta}\|^2 + 2\sum_{j=1}^{p-J} |\theta_j|, \tag{41}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p-J})'$, and the columns of $\tilde{\boldsymbol{X}}$ are $\omega(J+1)\mathbf{x}_{J+1}, \ldots, \omega(p)\mathbf{x}_p$. Because the objective function in (41) is strictly convex, by Theorem 4, $\{\boldsymbol{\beta}_2^{(k)}\}$ converges to a limit with grouping coherence. This completes the proof. □

Example 7. Consider $\boldsymbol{X}$ in Example 4. Let $\mathbf{Y} = (2, 1, -4, 1.5)'$. Using an initial point $\boldsymbol{\beta}^{(0)} = 0$, the OEM sequence of the lasso with $\lambda = 1$ converges to

$$\hat{\boldsymbol{\beta}} = (-0.5625,\ 0.4375,\ -0.6875,\ 0.6875,\ -0.4375,\ 0.5625)',$$

which has grouping coherence. For the same data and the same initial point, the coordinate descent sequence converges to $2(-0.5625,\ 0.4375,\ -0.6875,\ 0,\ 0,\ 0)'$, which does not have grouping coherence.

Theorem 7 shows that, if the initial point has grouping coherence, then every limit point of the OEM sequence for the lasso inherits this property. It is now tempting to ask whether such a result holds for the OEM sequence with $\lambda = 0$ in (16), i.e., the Healy and Westmacott procedure. Since full aliasing is just one possible culprit for making the matrix $\boldsymbol{X}$ in (1) lack full column rank and hence $\boldsymbol{X}'\boldsymbol{X}$ become singular, Theorem 8 provides an answer to this question for the general singular situation.

Let r denote the rank of $\boldsymbol{X}$. When $r < p$, the singular value decomposition (Wilkinson 1965) of $\boldsymbol{X}$ is

$$\boldsymbol{X} = \boldsymbol{U}' \begin{pmatrix} \boldsymbol{\Gamma}_0^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{V},$$

where $\boldsymbol{U}$ is an $n \times n$ matirx, $\boldsymbol{V}$ is a $p \times p$ orthogonal matrices and $\boldsymbol{\Gamma}_0$ is a diagonal matrix whose diagonal elements are the positive eigenvalues, $\gamma_1 \geqslant \cdots \geqslant \gamma_r$, of $\boldsymbol{X}'\boldsymbol{X}$. Define

$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}'\boldsymbol{X})^+ \boldsymbol{X}'\mathbf{Y}, \tag{42}$$

where $^+$ denotes the Moore-Penrose generalized inverse (Ben-Israel and Greville 2003).

Theorem 8. Suppose that $\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}'\boldsymbol{\Delta} = \gamma_1 \boldsymbol{I}_p$. If $\boldsymbol{\beta}^{(0)}$ lies in the linear space spanned by the first r columns of $\boldsymbol{V}'$, then as $k \to \infty$, for the OEM sequence $\{\boldsymbol{\beta}^{(k)}\}$ of the ordinary least squares, $\boldsymbol{\beta}^{(k)} \to \hat{\boldsymbol{\beta}}^*$.

Proof. Denote $\boldsymbol{D} = \boldsymbol{I}_p - \gamma_1^{-1}\boldsymbol{X}'\boldsymbol{X}$. Note that $\boldsymbol{\beta}^{(k+1)} = \gamma_1^{-1}\boldsymbol{X}'\mathbf{Y} + \boldsymbol{D}\boldsymbol{\beta}^{(k)}$. By induction,

$$\begin{aligned}
\boldsymbol{\beta}^{(k)} &= \gamma_1^{-1}(\boldsymbol{I}_p + \boldsymbol{D} + \cdots + \boldsymbol{D}^{k-1})\boldsymbol{X}'\mathbf{Y} + \boldsymbol{D}^k\boldsymbol{\beta}^{(0)} \\
&= \gamma_1^{-1}\boldsymbol{V}' \left\{ \boldsymbol{I}_p + \begin{pmatrix} \boldsymbol{I}_r - \gamma_1^{-1}\boldsymbol{\Gamma}_0 & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{I}_{p-r} \end{pmatrix} + \cdots + \begin{pmatrix} (\boldsymbol{I}_r - \gamma_1^{-1}\boldsymbol{\Gamma}_0)^{k-1} & \mathbf{0} \\ \mathbf{0} & (-1)^{k-1}\boldsymbol{I}_{p-r} \end{pmatrix} \right\} \boldsymbol{V} \\
&\quad \cdot \boldsymbol{V}' \begin{pmatrix} \boldsymbol{\Gamma}_0^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{U}\mathbf{Y} + \boldsymbol{D}^k\boldsymbol{\beta}^{(0)} \\
&= \gamma_1^{-1}\boldsymbol{V}' \begin{pmatrix} \{\boldsymbol{I}_r + (\boldsymbol{I}_r - \gamma_1^{-1}\boldsymbol{\Gamma}_0) + \cdots + (\boldsymbol{I}_r - \gamma_1^{-1}\boldsymbol{\Gamma}_0)^{k-1}\}\boldsymbol{\Gamma}_0^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{U}\mathbf{Y} + \boldsymbol{D}^k\boldsymbol{\beta}^{(0)}.
\end{aligned}$$

As $k \to \infty$, we have that

$$\boldsymbol{D}^k \to \boldsymbol{V}' \begin{pmatrix} \mathbf{0} & \\ & \boldsymbol{I}_{p-r} \end{pmatrix} \boldsymbol{V}$$

and $\boldsymbol{D}^k\boldsymbol{\beta}^{(0)} \to 0$, which implies that

$$\boldsymbol{\beta}^{(k)} \to \boldsymbol{V}' \begin{pmatrix} \boldsymbol{\Gamma}_0^{-1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{U}\mathbf{Y} = \hat{\boldsymbol{\beta}}^*.$$

This completes the proof. □

The condition $\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}'\boldsymbol{\Delta} = \gamma_1 \boldsymbol{I}_p$ holds if $\boldsymbol{S} = \boldsymbol{I}_p$ in (26).

Remark 8. Computing the Moore-Penrose generalized inverse $\hat{\boldsymbol{\beta}}^*$ in (42) is a difficult problem. Theorem 8 says that the OEM algorithm provides an efficient solution to this problem. When $r < p$, the limiting matrix $\hat{\boldsymbol{\beta}}^*$ given by an OEM sequence has the following properties. First, it has the minimal Euclidean norm among the least squares estimators $(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\mathbf{Y}$ (Ben-Israel and Greville 2003). Second, its model error has a simple form, $E\big[(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})\big] = r\sigma^2$. Third, $\boldsymbol{X}\boldsymbol{\alpha} = \mathbf{0}$ implies $\boldsymbol{\alpha}'\hat{\boldsymbol{\beta}}^* = 0$ for any vector $\boldsymbol{\alpha}$, which immediately implies that $\hat{\boldsymbol{\beta}}^*$ has grouping coherence.

Example 8. Use the same data and the same initial point in Example 7. The OEM sequence of the ordinary least squares converges to

$$\hat{\boldsymbol{\beta}}^* = (-0.6875,\ 0.5625,\ -0.8125,\ 0.8125,\ -0.5625,\ 0.6875)',$$

which has grouping coherence.

7 DISCUSSION

For the regularized least squares method with the SCAD penalty, finding a local solution to achieve the oracle property is a well-known open problem. We have proposed an algorithm, called the OEM algorithm, to fill this gap. For the SCAD and MCP penalties, this algorithm can provide a local solution with the oracle property. The discovery of the algorithm is quite accidental, drawing direct impetus from a missing-data problem arising in design of experiments. The introduction of the active orthogonization procedure in Section 3

makes the algorithm applicable to very general data structures from observational studies and experiments. Recent years have witnessed an explosive interest in both the theoretical and computational aspects of penalized methods. Our introduction of an algorithm that not only has desirable numerical convergence properties but also possesses an important theoretical property suggests a new interface between these two aspects. Subsequent work in this direction is expected. The active orthogonization idea is general and may have potential applications beyond the scope of the OEM algorithm, such as other EM algorithms (Meng and Rubin 1991; Meng and Rubin 1993; Meng 2007), data augmentation (Tanner and Wong 1987), Markov chain Monte Carlo (Liu 2001), smoothing splines (Wahba and Luo 1997; Luo 1998), mesh-free methods in numerical analysis (Fasshauer 2007) and parallel computing (Kumar, Grama, Gupta, and Karypis 2003). This procedure may also has intriguing mathematical connection with complementary theory in design of experiments (Tang and Wu 1996; Chen and Hedayat 1996; Xu and Wu 2005).

The result on the oracle property in Section 4 uses the assumption that the sample size n goes to infinity. This result is appealing for practical situations with massive data (SAMSI 2012), such as the data deluge in astronomy, the Internet and marketing (the Economist 2010), large-scale industrial experiments (Xu 2009) and modern simulations in engineering (NAE 2008), to just name a few. For applications like micro-array and image analysis, one might be interested in extending the result to the small n and large p case, like in Fan and Peng (2004). Such an extension, however, poses significant challenges. Even for a fixed p, the penalized likelihood function for the SCAD can have a large number of local minima (Huo and Chen 2010). When p goes to infinity, that number can be prohibitively large, which makes it very difficult to sort out a local minima with the oracle property.

In addition to achieving the oracle property for nonconvex penalties, an OEM sequence has other unique theoretical properties, including convergence to a point having grouping coherence for the lasso, SCAD or MCP and convergence to the Moore-Penrose generalized inverse-based least squares estimator for singular regression matrices. These theoretical results together with the active orthogonization scheme form the main contribution of the article.

A computer package for distributing the OEM algorithm to the general audience is under development and will be released. We now remark on the acceleration issue and directions for future work. The algorithm can be speeded up by using various methods from the EM literature (McLachlan and Krishnan 2008). For example, following the idea in Varadhan and Roland (2008), one can replace the OEM iteration in (14) by

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - 2\gamma\mathbf{r} + \gamma^2\mathbf{v},$$

where $\mathbf{r} = \mathbf{M}(\boldsymbol{\beta}^{(k)}) - \boldsymbol{\beta}^{(k)}$, $\mathbf{v} = \mathbf{M}(\mathbf{M}(\boldsymbol{\beta}^{(k)})) - \mathbf{M}(\boldsymbol{\beta}^{(k)}) - r$ and $\gamma = -\|\mathbf{r}\|/\|\mathbf{v}\|$. This scheme is found to lead to significant reduction of the running time in several examples. For problems with very large p, one may consider a hybrid algorithm to combine the OEM and coordinate descent ideas. It partitions $\boldsymbol{\beta}$ in (1) into G groups and in each iteration, it minimizes the objective function l in (35) by using the OEM algorithm with respect to one group while holding the other groups fixed. Here are some details. Group $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_G')'$. For $k = 0, 1, \ldots$, solve

$$\boldsymbol{\beta}_g^{(k+1)} = \arg\min_{\boldsymbol{\beta}_g} l(\boldsymbol{\beta}_1^{(k+1)}, \ldots, \boldsymbol{\beta}_{g-1}^{(k+1)}, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{g+1}^{(k)}, \ldots, \boldsymbol{\beta}_G^{(k)}) \text{ for } g = 1, \ldots, G \qquad (43)$$

by OEM until convergence. Note that (43) has a much lower dimension than the iteration in (14). For $G = 1$, the hybrid algorithm reduces to the OEM algorithm and for $G = p$, it becomes the coordinate descent algorithm. Theoretical properties of this hybrid algorithm will be studied and reported elsewhere.

Extension of the OEM algorithm can be made by imposing special structures on regression matrices, such as grouped variables (Yuan and Lin 2006; Zhou and Zhu 2008; Huang, Ma, Xie, and Zhang 2009; Zhao, Rocha, and Yu 2009; Wang, Chen, and Li 2009; Xiong 2010), mixtures (Khalili and Chen 2007) and heredity constraints (Yuan, Joseph, and Lin 2007; Yuan, Joseph, and Zou 2009; Choi, Li, and Zhu 2010), among many other possibilities.

APPENDIX: PROOF OF THEOREM 1

Proof. We first give some definitions and notation. Let Φ be the cumulative distribution function of the standard normal random variable. For $a > 2$ and λ in (3) and $d \geqslant \gamma_1$ in Assumption 1, define

$$s(u;\lambda) = \begin{cases} \text{sign}(u)\big(|u| - \lambda\big)_+/d, & \text{when } |u| \leqslant (d+1)\lambda, \\ \text{sign}(u)\big\{(a-1)|u| - a\lambda\big\}/\big\{(a-1)d - 1\big\}, & \text{when } (d+1)\lambda < |u| \leqslant ad\lambda, \\ u/d, & \text{when } |u| > ad\lambda, \end{cases}$$

Let $\mathbf{s}(\mathbf{u};\lambda) = \big[s(u_1;\lambda),\ldots,s(u_p;\lambda)\big].'$ The OEM sequence from (12) satisfies the condition that $\boldsymbol{\beta}^{(k+1)} = \mathbf{s}(\mathbf{u}^{(k)};\lambda_n/n)$, where

$$\mathbf{u}^{(k)} = (\mathbf{u}_1^{(k)'}, \mathbf{u}_2^{(k)'})' = \frac{\boldsymbol{X}'\mathbf{Y}}{n} + \left(d\boldsymbol{I}_d - \frac{\boldsymbol{X}'\mathbf{Y}}{n}\right)\boldsymbol{\beta}^{(k)}. \tag{44}$$

For $k = 1, 2, \ldots$, define two sequences of events $A_k = \{\boldsymbol{\beta}_2^{(k)} = \mathbf{0}\}$ and $B_k = \{\boldsymbol{\beta}_1^{(k)} = \mathbf{u}_1^{(k-1)}/d\}$. For $h > 0$ and $k = 0, 1, \ldots$, let $C_k^h = \{\|\boldsymbol{\beta}_1^{(k)} - \boldsymbol{\beta}_1\| \leqslant h\lambda_n/n\}$. The flow of the proof is to first show that $\mathrm{P}(A_k)$, $\mathrm{P}(B_k)$ and $\mathrm{P}(C_k^h)$ all tend to one at exponential rates as n goes to infinity, thus establishing Theorem 1 (i), then show that $\mathrm{P}(\cap_{i=1}^k A_i)$ and $\mathrm{P}(\cap_{i=1}^{k+1} B_i)$ tend to one and finally establish Theorem 1(ii) by noting that the asymptotic normality of $\boldsymbol{\beta}_1^{(k)}$ follows when $\cap_{i=1}^k A_i$ and $\cap_{i=1}^{k+1} B_i$ both occur.

Step 1. Let $\boldsymbol{G} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ with columns $\mathbf{g}_1, \ldots, \mathbf{g}_p$. Let $\boldsymbol{G}_1$ denote its submatrix with the first p_1 columns. Let $\tau \geqslant 0$ denote the largest eigenvalue of $\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{X}_2'\boldsymbol{X}_1/n^2$. Define

$$h_A = \begin{cases} 1/(2\tau), & \text{when } \tau > 0, \\ +\infty, & \text{when } \tau = 0. \end{cases} \tag{45}$$

Let $(\mathbf{v}_1, \ldots, \mathbf{v}_{p_1}) = d\boldsymbol{I}_{p_1} - n^{-1}\boldsymbol{X}_1'\boldsymbol{X}_1$ and $b = \max\{\|\mathbf{v}_1\|, \ldots, \|\mathbf{v}_{p_1}\|\}$. Define

$$h_B = \frac{ad}{b}, \tag{46}$$

with a and d given in (3) and Assumption 1, respectively. For $h > 0$, define

$$h_C = \frac{h}{2\eta}, \tag{47}$$

where η, used in Assumption 3, is the largest eigenvalue of $\boldsymbol{I}_{p_1} - \boldsymbol{X}_1'\boldsymbol{X}_1/(nd)$.

For C_0^h, since $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta} + \boldsymbol{G}'\boldsymbol{\varepsilon}$, we have that

$$\begin{aligned}
\mathrm{P}(C_0^h) &= \mathrm{P}(\|\boldsymbol{G}_1'\boldsymbol{\varepsilon}\| \leqslant h\lambda_n/n) \\
&\geqslant \mathrm{P}\big(|\mathbf{g}_j'\boldsymbol{\varepsilon}| \leqslant h\lambda_n/(n\sqrt{p_1}) \text{ for } j = 1, \ldots .p_1\big) \\
&\geqslant 1 - \sum_{j=1}^{p_1}\Big[1 - \mathrm{P}\big(|\mathbf{g}_j'\boldsymbol{\varepsilon}| \leqslant h\lambda_n/(n\sqrt{p_1})\big)\Big] \\
&= 1 - 2\sum_{j=1}^{p_1}\left[1 - \Phi\left(\frac{h\lambda_n}{n\sqrt{p_1}\sigma\|\mathbf{g}_j\|}\right)\right].
\end{aligned} \tag{48}$$

For A_1, note that

$$\begin{aligned}
\mathrm{P}(A_1) &= \mathrm{P}\big(|\beta_j^{(0)}| \leqslant \lambda_n/(nd) \text{ for } j = p_1 + 1, \ldots, p\big) \\
&\geqslant 1 - \sum_{j=p_1+1}^{p}\Big[1 - \mathrm{P}\big(|\mathbf{g}_j'\boldsymbol{\varepsilon}| \leqslant \lambda_n/(nd)\big)\Big] \\
&= 1 - 2\sum_{j=p_1+1}^{p}\left[1 - \Phi\left(\frac{\lambda_n}{nd\sigma\|\mathbf{g}_j\|}\right)\right].
\end{aligned} \tag{49}$$

For B_1, note that

$$\begin{aligned}
\mathrm{P}(B_1) &= \mathrm{P}\big(|\beta_j^{(0)}| \geqslant a\lambda_n/n \text{ for } j = 1, \ldots, p_1\big) \\
&\geqslant 1 - \sum_{j=1}^{p_1}\Big[1 - \mathrm{P}\big(|\beta_j + \mathbf{g}_j'\boldsymbol{\varepsilon}| \geqslant a\lambda_n/n\big)\Big] \\
&= 1 - \sum_{j=1}^{p_1}\left[\Phi\left(\frac{n\beta_j + a\lambda_n}{n\sigma\|\mathbf{g}_j\|}\right) - \Phi\left(\frac{n\beta_j - a\lambda_n}{n\sigma\|\mathbf{g}_j\|}\right)\right].
\end{aligned} \tag{50}$$

For any $h > 0$, by (48),

$$\begin{aligned} \mathrm{P}(C_1^h) &\geqslant \mathrm{P}(C_1^h \cap B_1) = \mathrm{P}(C_0^h \cap B_1) \\ &\geqslant 1 - 2\sum_{j=1}^{p_1}\left[1 - \Phi\left(\frac{h\lambda_n}{n\sqrt{p_1}\sigma\|\mathbf{g}_j\|}\right)\right] \\ &\quad - \sum_{j=1}^{p_1}\left[\Phi\left(\frac{n\beta_j + a\lambda_n}{n\sigma\|\mathbf{g}_j\|}\right) - \Phi\left(\frac{n\beta_j - a\lambda_n}{n\sigma\|\mathbf{g}_j\|}\right)\right]. \end{aligned} \tag{51}$$

Next, consider A_k, B_k and C_k^h, for $k = 2, 3, \ldots$. If A_{k-1} occurs, then by (44),

$$\mathbf{u}^{(k-1)} = \frac{\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{n} + \frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{n} + \begin{pmatrix} d\boldsymbol{I}_{p_1} - \boldsymbol{X}_1'\boldsymbol{X}_1/n & -\boldsymbol{X}_1'\boldsymbol{X}_2/n \\ -\boldsymbol{X}_2'\boldsymbol{X}_1/n & d\boldsymbol{I}_{p-p_1} - \boldsymbol{X}_2'\boldsymbol{X}_2/n \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1^{(k-1)} \\ \mathbf{0} \end{pmatrix}.$$

Thus,

$$\mathbf{u}_1^{(k-1)} = d\boldsymbol{\beta}_1^{(k-1)} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{n}[\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{(k-1)}] + \frac{\boldsymbol{X}_1'\boldsymbol{\varepsilon}}{n}, \tag{52}$$

$$\mathbf{u}_2^{(k-1)} = \frac{\boldsymbol{X}_2'\boldsymbol{X}_1}{n}[\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{(k-1)}] + \frac{\boldsymbol{X}_2'\boldsymbol{\varepsilon}}{n}. \tag{53}$$

By (53),

$$\begin{aligned} \mathrm{P}(A_k) &\geqslant \mathrm{P}(A_k \cap A_{k-1} \cap C_{k-1}^{h_A}) \\ &= \mathrm{P}\big(\{|u_j^{(k-1)}| \leqslant \lambda_n/n \text{ for } j = p_1+1, \ldots, p\} \cap A_{k-1} \cap C_{k-1}^{h_A}\big) \\ &\geqslant \mathrm{P}\big(\{\|\mathbf{u}_2^{(k-1)}\| \leqslant \lambda_n/n\} \cap A_{k-1} \cap C_{k-1}^{h_A}\big) \\ &= \mathrm{P}\big(\{\|n^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{(k-1)}) + n^{-1}\boldsymbol{X}_2'\boldsymbol{\varepsilon}\| \leqslant \lambda_n/n\} \cap A_{k-1} \cap C_{k-1}^{h_A}\big) \\ &\geqslant \mathrm{P}\big(\{\|n^{-1}\boldsymbol{X}_2'\boldsymbol{\varepsilon}\| \leqslant \lambda_n/n - \tau\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{(k-1)}\|\} \cap A_{k-1} \cap C_{k-1}^{h_A}\big) \\ &= \mathrm{P}\big(\{\|n^{-1}\boldsymbol{X}_2'\boldsymbol{\varepsilon}\| \leqslant \lambda_n/(2n)\|\} \cap A_{k-1} \cap C_{k-1}^{h_A}\big) \\ &\geqslant \mathrm{P}\big(\{|\mathbf{x}_j'\boldsymbol{\varepsilon}| \leqslant \lambda_n/(2\sqrt{p-p_1}) \text{ for } j = p_1+1, \ldots, p\} \cap A_{k-1} \cap C_{k-1}^{h_A}\big) \\ &\geqslant 1 - 2\sum_{j=p_1+1}^{p}\left[1 - \Phi\left(\frac{\lambda_n}{2\sqrt{p-p_1}\sigma\|\mathbf{x}_j\|}\right)\right] \\ &\quad - \big[1 - \mathrm{P}(A_{k-1})\big] - \big[1 - \mathrm{P}(C_{k-1}^{h_A})\big], \end{aligned} \tag{54}$$

where h_A is defined in (45).

By (52),

$$\begin{aligned}
\mathrm{P}(B_k) &\geqslant \mathrm{P}(B_k \cap A_{k-1} \cap C_{k-1}^{h_B}) \\
&= \mathrm{P}\big(\{|d\beta_j + \mathbf{u}_j'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{(k-1)}) + n^{-1}\mathbf{x}_j'\boldsymbol{\varepsilon}| \geqslant ad\lambda_n/n \text{ for } j = 1, \ldots, p_1\} \cap A_{k-1} \cap C_{k-1}^{h_B}\big) \\
&\geqslant \mathrm{P}\big(\{|d\beta_j + n^{-1}\mathbf{x}_j'\boldsymbol{\varepsilon}| \geqslant |\mathbf{u}_j'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{(k-1)})| + ad\lambda_n/n \text{ for } j = 1, \ldots, p_1\} \cap A_{k-1} \cap C_{k-1}^{h_B}\big) \\
&\geqslant \mathrm{P}\big(\{|d\beta_j + n^{-1}\mathbf{x}_j'\boldsymbol{\varepsilon}| \geqslant b\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{(k-1)}\| + ad\lambda_n/n \text{ for } j = 1, \ldots, p_1\} \cap A_{k-1} \cap C_{k-1}^{h_B}\big) \\
&\geqslant \mathrm{P}\big(\{|d\beta_j + n^{-1}\mathbf{x}_j'\boldsymbol{\varepsilon}| \geqslant 2ad\lambda_n/n \text{ for } j = 1, \ldots, p_1\} \cap A_{k-1} \cap C_{k-1}^{h_B}\big) \\
&\geqslant 1 - \sum_{j=1}^{p_1}\left[\Phi\left(\frac{nd\beta_j + 2ad\lambda_n}{\sigma\|\mathbf{x}_j\|}\right) - \Phi\left(\frac{nd\beta_j - 2ad\lambda_n}{\sigma\|\mathbf{x}_j\|}\right)\right] - \big[1 - \mathrm{P}(A_{k-1})\big] \\
&\quad -\big[1 - \mathrm{P}(C_{k-1}^{h_B})\big], \qquad (55)
\end{aligned}$$

where h_B is defined in (46).

For any $h > 0$, we have that

$$\begin{aligned}
\mathrm{P}(C_k^h) &\geqslant \mathrm{P}(C_k^h \cap B_k \cap A_{k-1} \cap C_{k-1}^{h_C}) \\
&= \mathrm{P}\big(\{\|(\boldsymbol{I}_{p_1} - \boldsymbol{X}_1'\boldsymbol{X}_1/(nd))(\boldsymbol{\beta}_1^{(k-1)} - \boldsymbol{\beta}_1) + \boldsymbol{X}_1'\boldsymbol{\varepsilon}/(nd)\| \leqslant h\lambda_n/n\} \cap B_k \cap A_{k-1} \cap C_{k-1}^{h_C}\big) \\
&\geqslant \mathrm{P}\big(\{\eta\|\boldsymbol{\beta}_1^{(k-1)} - \boldsymbol{\beta}_1\| + \|\boldsymbol{X}_1'\boldsymbol{\varepsilon}/(nd)\| \leqslant h\lambda_n/n\} \cap B_k \cap A_{k-1} \cap C_{k-1}^{h_C}\big) \\
&\geqslant \mathrm{P}\big(\{\|\boldsymbol{X}_1'\boldsymbol{\varepsilon}/(nd)\| \leqslant h\lambda_n/(2n)\} \cap B_k \cap A_{k-1} \cap C_{k-1}^{h_C}\big) \\
&\geqslant \mathrm{P}\big(\{|\mathbf{x}_j'\boldsymbol{\varepsilon}/(nd)| \leqslant h\lambda_n/(2n\sqrt{p_1}) \text{ for } j = 1, \ldots, p_1\} \cap B_k \cap A_{k-1} \cap C_{k-1}^{h_C}\big) \\
&\geqslant 1 - 2\sum_{j=1}^{p_1}\left[1 - \Phi\left(\frac{h\lambda_n}{2n\sqrt{p_1}\sigma\|\mathbf{x}_j\|}\right)\right] - \big[1 - \mathrm{P}(A_{k-1})\big] - \big[1 - \mathrm{P}(B_k)\big] \\
&\quad -\big[1 - \mathrm{P}(C_{k-1}^{h_C})\big], \qquad (56)
\end{aligned}$$

where h_C is defined in (47).

Since $1-\Phi(x)=o\big(\exp(-x^2/2)\big)$ as $x\to+\infty$, (49), (50) and (51) imply that

$$\begin{aligned}
1-P(A_1)&=o\big(\exp[-c_1(\lambda_n/\sqrt{n})^2]\big),\\
1-P(B_1)&=o\big(\exp(-c_2 n)\big)=o\big(\exp[-c_2(\lambda_n/\sqrt{n})^2]\big),\\
1-P(C_1^h)&=o\big(\exp[-c_3(\lambda_n/\sqrt{n})^2]\big),
\end{aligned}$$

where $c_1,\ c_2$ and c_3 are positive constants. By induction, it now follows from (54), (55) and (56) that

$$\begin{aligned}
1-P(A_k) &= \big[1-P(A_{k-1})\big]+\big[1-\mathrm{P}(C_{k-1}^{h_A})\big]+o\big(\exp[-c_4(\lambda_n/\sqrt{n})^2]\big)\\
&= o\big(k\exp[-c_4(\lambda_n/\sqrt{n})^2]\big), \qquad (57)
\end{aligned}$$

where c_4 and c_5 are positive constants. Similarly,

$$1-P(B_k)=o\big(k\exp[-c_6(\lambda_n/\sqrt{n})^2]\big), \qquad (58)$$

$$1-P(C_k^h)=o\big(k\exp[-c_7(\lambda_n/\sqrt{n})^2]\big), \qquad (59)$$

where c_6 and c_7 are positive constants. By (57), a sufficient condition for $\mathrm{P}(A_k)\to 1$ is

$$k\exp\big[-c(\lambda_n/\sqrt{n})^2\big]\to 0 \qquad (60)$$

for any $c>0$, which is covered by Assumption 3.

Step 2. Consider the asymptotic normality of $\boldsymbol{\beta}_1^{(k)}$. When the events $A_{k-1},\ A_k,\ B_k$ and B_{k+1} all occur, by (52),

$$\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{n}(\boldsymbol{\beta}_1^{(k)}-\boldsymbol{\beta}_1)=\frac{\boldsymbol{X}_1'\boldsymbol{\varepsilon}}{n}+d(\boldsymbol{\beta}_1^{(k)}-\boldsymbol{\beta}_1^{(k+1)}), \qquad (61)$$

$$\|\boldsymbol{\beta}_1^{(k+1)}-\boldsymbol{\beta}_1^{(k)}\|=\left\|\left(\boldsymbol{I}_{p_1}-\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{nd}\right)(\boldsymbol{\beta}_1^{(k)}-\boldsymbol{\beta}_1^{(k-1)})\right\|\leqslant\eta\|\boldsymbol{\beta}_1^{(k)}-\boldsymbol{\beta}_1^{(k-1)}\|. \qquad (62)$$

When the events $C_1^{1/2},\ C_2^{1/2},\ A^{(k)} = \bigcap_{i=1}^{k} A_i$ and $B^{(k+1)} = \bigcap_{i=1}^{k+1} B_i$ all occur, by (62),

$$\|\boldsymbol{\beta}_1^{(k+1)} - \boldsymbol{\beta}_1^{(k)}\| \leqslant \eta^{k-1}\|\boldsymbol{\beta}_1^{(2)} - \boldsymbol{\beta}_1^{(1)}\| \leqslant \eta^{k-1}\lambda_n/n. \tag{63}$$

For any $\boldsymbol{\alpha} \in \mathbb{R}^p$ with $\boldsymbol{\alpha} \neq \mathbf{0}$, by (61) and (63),

$$\begin{aligned} & |\sqrt{n}\boldsymbol{\alpha}'(\boldsymbol{\beta}_1^{(k)} - \boldsymbol{\beta}_1) - \sqrt{n}\boldsymbol{\alpha}'(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\varepsilon}| \\ = \ & dn^{3/2}|\boldsymbol{\alpha}'(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}(\boldsymbol{\beta}_1^{(k)} - \boldsymbol{\beta}_1^{(k+1)})| \\ \leqslant \ & n^{-1/2}\lambda_n\eta^{k-1}d\|\boldsymbol{\alpha}\|\|n(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\|. \end{aligned}$$

For any $x \in \mathbb{R}$, note that

$$\begin{aligned} & \left|\mathrm{P}\big(\sqrt{n}\boldsymbol{\alpha}'(\boldsymbol{\beta}_1^{(k)} - \boldsymbol{\beta}_1) \leqslant x\big) - \Phi\left(\frac{x}{\sigma(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\alpha})^{1/2}}\right)\right| \\ \leqslant \ & \left|\mathrm{P}\big(\{\sqrt{n}\boldsymbol{\alpha}'(\boldsymbol{\beta}_1^{(k)} - \boldsymbol{\beta}_1) \leqslant x\} \cap C_1^{1/2} \cap C_2^{1/2} \cap A^{(k)} \cap B^{(k+1)}\big) - \Phi\left(\frac{x}{\sigma(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\alpha})^{1/2}}\right)\right| \\ & + \big[1 - \mathrm{P}(C_1^{1/2})\big] + \big[1 - \mathrm{P}(C_1^{1/2})\big] + \big[1 - \mathrm{P}(A^{(k)})\big] + \big[1 - \mathrm{P}(B^{(k+1)})\big] \\ \leqslant \ & \left|\mathrm{P}\big(\sqrt{n}\boldsymbol{\alpha}'(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\varepsilon} \leqslant x - n^{-1/2}\lambda_n\eta^{k-1}d\|\boldsymbol{\alpha}\|\|n(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\|\big) - \Phi\left(\frac{x}{\sigma(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\alpha})^{1/2}}\right)\right| \\ & + \left|\mathrm{P}\big(\sqrt{n}\boldsymbol{\alpha}'(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\varepsilon} \leqslant x + n^{-1/2}\lambda_n\eta^{k-1}d\|\boldsymbol{\alpha}\|\|n(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\|\big) - \Phi\left(\frac{x}{\sigma(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\alpha})^{1/2}}\right)\right| \\ & + 3\big[1 - \mathrm{P}(C_1^{1/2})\big] + 3\big[1 - \mathrm{P}(C_1^{1/2})\big] + 3\big[1 - \mathrm{P}(A^{(k)})\big] + 3\big[1 - \mathrm{P}(B^{(k+1)})\big] \\ = \ & \left|\Phi\left(\frac{x - n^{-1/2}\lambda_n\eta^{k-1}d\|\boldsymbol{\alpha}\|\|n(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\|}{\sigma[\boldsymbol{\alpha}'n(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{\alpha}]^{1/2}}\right) - \Phi\left(\frac{x}{\sigma(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\alpha})^{1/2}}\right)\right| \\ & + \left|\Phi\left(\frac{x + n^{-1/2}\lambda_n\eta^{k-1}d\|\boldsymbol{\alpha}\|\|n(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\|}{\sigma[\boldsymbol{\alpha}'n(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{\alpha}]^{1/2}}\right) - \Phi\left(\frac{x}{\sigma(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\alpha})^{1/2}}\right)\right| \\ & + 6\big[1 - \mathrm{P}(C_1^{1/2})\big] + 3\big[1 - \mathrm{P}(A^{(k)})\big] + 3\big[1 - \mathrm{P}(B^{(k+1)})\big]. \end{aligned} \tag{64}$$

Now, under Assumption 3, by (57), (58) and (59), (64) converges to zero as $n \to \infty$. This completes the proof. □

ACKNOWLEDGEMENTS

Xiong is supported by grant 10801130 of the National Natural Science Foundation of China. Dai is partially support by Grace Wahba through NIH grant R01 EY009946, ONR grant N00014-09-1-0655 and NSF grant DMS-0906818. Qian is partially supported by NSF grant CMMI 0969616, an IBM Faculty Award and an NSF Faculty Early Career Development Award DMS 1055214. The authors thank Xiao-Li Meng and Grace Wahba for comments and suggestions.

REFERENCES

Addelman, S. (1962), "Orthogonal Main-Effect Plans for Asymmetrical Factorial Experiments," *Technometrics*, 4, 21-46.

Ben-Israel, A. and Greville, T. N. E. (2003), *Generalized Inverses, Theory and Applications*, 2nd ed., New York: Springer.

Bertsekas, D. P. (1999), *Nonlinear Programming*, 2nd ed., Belmont, MA: Athena Scientific.

Bingham, D., Sitter, R. R. and Tang, B. (2009). "Orthogonal and Nearly Orthogonal Designs for Computer Experiments," *Biometrika*, 96, 51–65.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

Breheny, P. and Huang, J. (2011), "Coordinate Descent Algorithms for Nonconvex Penalized Regression, With Applications to Biological Feature Selection," *The Annals of Applied Statistics*, in press.

Chen, H. and Hedayat, A. S. (1996), "2^{n-l} designs with Weak Minimum Aberration," *The Annals of Statistics*, 24, 2536–2548.

Choi, N. H., Li W. and Zhu J. (2010), "Variable Selection With the Strong Heredity Constraint and Its Oracle Property," *Journal of the American Statistical Association*, 105, 354–364.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–38.

Deng, X., Lin, C. D. and Qian, P. Z. G. (2010), "Designs for the Lasso," *Technical Report*.

Dey, A. and Mukerjee, R. (1999), *Fractional Factorial Plans*, New York: Wiley.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–451.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Peng, H. (2004), "On Non-concave Penalized Likelihood With Diverging Number of Parameters," *The Annals of Statistics*, 32, 928–961.

Fang, K. T., Li, R. and Sudjianto, A. (2005), *Design and Modeling for Computer Experiments*, New York: Chapman and Hall/CRC Press.

Fasshauer, G. F. (2007), *Meshfree Approximation Methods With MATLAB*, Singapore: World Scientific Publishing Company.

Frank, L. E. and Friedman, J. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, **35**, 109–135.

Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1, 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2009), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22.

Friedman, J., Hastie, T. and Tibshirani, R. (2011), "Glmnet," R package.

Fu, W. J. (1998), "Penalized Regressions: The Bridge versus the LASSO," *Journal of Computational and Graphical Statistics*, 7, 397–416.

Grandvalet, Y. (1998), "Least Absolute Shrinkage Is Equivalent to Quadratic Penalization," *In: Niklasson, L., Bodén, M., Ziemske, T. (eds.), ICANN'98. Vol. 1 of Perspectives in Neural Computing, Springer,* 201–206.

Green, P. J. (1990), "On Use of the EM Algorithm for Penalized Likelihood Estimation," *Journal of the Royal Statistical Society, Ser. B*, 52, 443–452.

Hastie, T. and Efron, B. (2011), "Lars," R package.

Healy, M. J. R. and Westmacott, M. H. (1956), "Missing Values in Experiments Analysed on Automatic Computers," *Journal of the Royal Statistical Society, Ser. C*, 5, 203–206.

Hedayat, A. S., Sloane, N. J. A. and Stufken, J. (1999), *Orthogonal Arrays: Theory and Applications*, New York: Springer.

Hoerl, A. E. and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Huang, J., Ma, S., Xie, H. and Zhang, C-H. (2009), "A Group Bridge Approach for Variable Selection," *Biometrika*, 96, 339–355.

Hunter, D. R. and Li, R. (2005), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, 33, 1617–1642.

Huo, X. and Chen, J. (2010), "Complexity of Penalized Likelihood Estimation," *Journal of Statistical Computation and Simulation*, 80, 747–759 .

Huo, X. and Ni, X. L. (2007), "When Do Stepwise Algorithms Meet Subset Selection Criteria?" *The Annals of Statistics*, 35, 870–887.

Khalili, A. and Chen, J. (2007), "Variable Selection in Finite Mixture of Regression Models," *Journal of the American Statistical Association*, 102, 1025–1038.

Kumar, V., Grama, A., Gupta, A. and Karypis, G. (2003), *Introduction to Parallel Computing*, 2nd ed., Boston, MA: Addison Wesley.

Lin, C. D., Bingham, D., Sitter, R. R. and Tang, B. (2010), "A New and Flexible Method for Constructing Designs for Computer Experiments," *The Annals of Statistics*, 38, 1460–1477.

Lin, C. D., Mukerjee, R. and Tang, B. (2009), "Construction of Orthogonal and Nearly Orthogonal Latin Hypercubes," *Biometrika*, 96, 243–247.

Li, R. and Lin, D. K. J. (2002), "Data Analysis in Supersaturated Designs," *Statistics and Probability Letters*, 59, 135–144.

Lin, D. K. J. (1993), "A New Class of Supersaturated Designs," *Technometrics*, 35, 28–31.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer.

Luo, Z. (1998), "Backfitting in Smoothing Spline ANOVA," *The Annals of Statistics*, 26, 1733–1759.

MacWilliams, F. J. and Sloane, N. J. A. (1977), *The Theory of Error Correcting Codes*, New York: North-Holland.

Mazumder, R., Friedman, J. and Hastie, T. (2010), "SparseNet: Coordinate Descent with Non-Convex Penalties," *Technical Report*.

McLachlan, G. and Krishnan, T. (2008), *The EM Algotithm and Extensions*, 2nd ed., New York: Wiley.

Meng, X. L. (1994), "On the Rate of Convergence of the ECM Algorithm," *The Annals of Statistics*, 22, 326–339.

Meng, X. L. (2007), "Thirty Years of EM and Much More," *Statistica Sinica*, 17, 839–840.

Meng, X. L. and Rubin, D. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909.

Meng, X. L. and Rubin, D. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

Mukerjee, R. and Wu, C. F. J. (2006), *A Modern Theory of Factorial Design*, New York: Springer.

NAE (2008), "Grand Challenges for Engineering," *Technical report*, The National Academy of Engineering.

Nettleton, D. (1999), "Convergence Properties of the EM Algorithm in Constrained Parameter Spaces," *Canadian Journal of Statistics*, 27, 639–648.

Osborne, M. R., Presnell, B. and Turlach, B. (2000), "On the LASSO and Its Dual," *Journal of Computational and Graphical Statistics*, 9, 319–337.

Owen, A. B. (1994), "Controlling Correlations in Latin Hypercube Samples," *Journal of the American Statistical Association*, 89, 1517–1522.

Owen, A. B. (2006), "A Robust Hybrid of Lasso and Ridge Regression," *Technical Report.*

Pang, F., Liu, M. Q. and Lin, D. K. J. (2009), "A Construction Method for Orthogonal Latin Hypercube Designs with Prime Power Levels," *Statistica Sinica*, 19, 1721–1728.

Phoa, F. K. H., Pan, Y-H. and Xu, H. (2009), "Analysis of Supersaturated Designs via The Dantzig Selector," *Journal of Statistical Planning and Inference*, 139, 2362–2372.

Rubin, D. B. and Szatrowski, T. H. (1982), "Finding Maximum Likelihood Estimates of Patterned Covariance Matrices by the EM Algorithm," *Biometrika*, 69, 657–660.

SAMSI (2012), Program on Statistical and Computational Methodology for Massive Datasets, 2012-2013.

Santner, T. J., Williams, B. J. and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer.

Schifano, E. D., Strawderman, R. and Wells, M. T. (2010), "Majorization-Minimization Algorithms for Nonsmoothly Penalized Objective Functions," *Electronic Journal of Statistics*, 23, 1258–1299.

Steinberg, D. M. and Lin, D. K. J. (2006), "A Construction Method for Orthogonal Lain Hypercube Designs," *Biometrika*, 93, 279–288.

Sun, F., Liu, M. and Lin, D. K. J. (2009), "Construction of Orthogonal Latin Hypercube Designs," *Biometrika*, 96, 971–974.

Tang, B. (1998), "Selecting Latin Hypercubes Using Correlation Criteria," *Statistica Sinica*, 8, 965–977.

Tang, B. and Wu, C. F. J. (1996), "Characterization of Minimum Aberration 2^{n-k} Designs in terms of Their Complementary Designs," *The Annals of Statistics*, 24, 2549–2559.

Tang, B. and Wu, C. F. J. (1997), "A Method for Constructing Supersaturated Designs and Its $E(s^2)$ Optimality," *Canadian Journal of Statistics*, 25, 191–201.

Tanner, M. A. and Wong, W. H. (1987). "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.

The Economist (2010), "Special Report on the Data Deluge, (February 27)," *The Economist*, 394, 3–18.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.

Tseng, P. (2001), "Convergence of a Block Coordnate Descent Method for Nondifferentialble Minimization," *Journal of Optimization Theory and Applications*, 109, 475–494.

Tseng, P. and Yun, S. (2009), "A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization," *Mathematical Programming B*, 117, 387–423.

Varadhan, R. and Roland, C. (2008), "Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm," *Scandinavian Journal of Statistics*, 35, 335–353.

Wahba, G. and Luo, Z. (1997), "Smoothing Spline ANOVA Fits for Very Large, Nearly Regular Data Sets, with Application to Historical Global Climate Data," *Annals of Numerical Mathematics*, 4579-4598. (Festschrift in Honor of Ted Rivlin, C. Micchelli, Ed.)

Wang, L., Chen, G. and Li, H. (2007), "Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data," *Bioinformatics*, 23, 1486–1494.

Watkins, D. S. (2002), *Fundamentals of Matrix Computations*, 2nd ed., New York: Wiley.

Wilkinson, J. H. (1965), *The Algebraic Eigenvalue Problem*, New York: Oxford University Press.

Witten, D. M. and Tibshirani, R. (2011), "Scout," R package.

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.

Wu, C. F. J. (1993), "Construction of Supersaturated Designs through Partially Aliased Interactions," *Biometrika*, 80, 661–669.

Wu, C. F. J. and Hamada M. (2009), *Experiments: Planning, Analysis, and Optimization*, 2nd ed., New York: Wiley.

Wu, T. and Lange, K. (2008), "Coordinate Descent Algorithm for Lasso Penalized Regression," *The Annals of Applied Statistics*, 2, 224–244.

Xiong, S. (2010), "Some Notes on the Nonnegative Garrote," *Technometrics*, 52, 349–361.

Xu, H. (2009), "Algorithmic Construction of Efficient Fractional Factorial Designs With Large Run Sizes," *Technometrics*, 51, 262–277.

Xu, H. and Cheng, C-S. (2008), "A Complementary Design Theory for Doubling," *The Annals of Statistics*, 36, 445–457.

Xu, H. and Wu, C. F. J. (2005), "Construction of Optimal Multi-Level Supersaturated Designs," *The Annals of Statistics*, 33, 2811–2836.

Ye, K. Q. (1998). "Orthogonal Column Latin Hypercubes and Their Applications in Computer Experiments," *Journal of the American Statistical Association*, 93, 1430–1439.

Yuan, M. and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society, Ser. B*, 68, 49–68.

Yuan, M., Joseph, V. R. and Lin, Y. (2007), "An Efficient Variable Selection Approach for Analyzing Designed Experiments," *Technometrics*, 49, 430–439.

Yuan, M., Joseph, V. R. and Zou, H. (2009), "Structured Variable Selection and Estimation," *Annals of Applied Statistics*, 3, 1738–1757.

Zangwill, W. I. (1969), *Nonlinear Programming: A Unified Approach*, Englewood Cliffs, New Jersey: Prentice Hall.

Zhang. C-H. (2010), "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942.

Zhao, P., Rocha, G. and Yu, B. (2009), "The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection," *The Annals of Statistics*, 37, 3468–3497.

Zhou, N. and Zhu, J. (2008), "Group Variable Selection via a Hierarchical Lasso and Its Oracle Property," *Technical Report*.

Zhu, Y. (2011), *Personal communication*.

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Ser. B*, 67, 301–320.

Zou, H. and Li, R. (2008), "One-step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, 36, 1509–1533.