Supplementary Material for Orthogonalizing EM: A design-based least squares algorithm

Shifeng Xiong[1], Bin Dai[2], Jared Huling[3], and Peter Z. G. Qian[3]

[1] Academy of Mathematics and Systems Science

Chinese Academy of Sciences, Beijing 100190

[2] Tower Research Capital, 377 Broadway, New York, NY 10013

[3] Department of Statistics

University of Wisconsin-Madison, Madison, WI 53706

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

A APPENDIX

A.1 POSSESSING GROUPING COHERENCE

Data with fully aliased structures commonly appear in observational studies and designed experiments. For example, in survey data, household income is the sum of the income of all family members. If all income variables are present, there is perfect aliasing. For large scale problems it can be computationally infeasible to determine which columns of a design matrix are aliased. In such situations it is desirable to use a methodology that can accomodate aliasing gracefully. Unlike other grouping techniques that require a prespecified grouping structure, we show that OEM naturally groups columns that are fully aliased. The group lasso penalty requires knowledge of which columns must be grouped, whereas OEM groups aliased columns without group structure specification *a priori*. We provide an illustration of grouping coherence in Section B.3.2. In this section, we consider the convergence of the OEM algorithm when the regression matrix $\boldsymbol{X}$ in (6) is singular due to fully aliased columns. Let $\boldsymbol{X}$ be standardized as in (9) with columns $\mathbf{x}_1, \ldots, \mathbf{x}_p$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are fully aliased, i.e., $|\mathbf{x}_i| = |\mathbf{x}_j|$, then the objective function in (8) for the lasso is not strictly convex and has many minima (Zou and Hastie 2005).

If some columns of $\boldsymbol{X}$ are identical, it is desirable to have grouping coherence with the same regression coefficient. This is suggested by Zou and Hastie (2005) and others. Definition 1 makes this precise.

Definition 1. An estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ of $\boldsymbol{\beta}$ in (6) 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 $\hat{\beta}_i = -\hat{\beta}_j$.

Some penalties other than the lasso can produce estimators with grouping coherence (Zou and Hastie 2005; Bondell and Reich 2008; Tutz and Ulbricht 2009; Petry and Tutz 2012). But they often require more than one tuning parameters, which lead to more computational burden. *Instead of changing the penalty, OEM can give a lasso solution with this property.* This also holds for SCAD and MCP. Recall that $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}'\boldsymbol{X})^+\boldsymbol{X}'\mathbf{y}$, which can be obtained by OEM, has a stronger property than grouping coherence.

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 A.1. Suppose that $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}'\boldsymbol{\Delta} = d\boldsymbol{I}_p)$ 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. See Section C □

Remark 1. Lemma A.1 implies that, for $k = 1, 2, \ldots$, $\boldsymbol{\beta}^{(k)}$ has grouping coherence if $\boldsymbol{\beta}^{(0)}$ has grouping coherence. Thus, for this case, any limit point of $\{\boldsymbol{\beta}^{(k)}\}$ has grouping coherence.

When $\boldsymbol{X}$ in (6) has fully aliased columns, the objective function in (8) for the lasso has many minima and hence the condition in Theorem 4 does not hold. Theorem A.1 shows that, even with full aliasing, an OEM sequence (13) for the lasso converges to a point with grouping coherence.

Theorem A.1. Suppose that $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}'\boldsymbol{\Delta} = d\boldsymbol{I}_p)$ 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. See Section C □

A.1.1 Grouping Coherence Illustration

We illustrate grouping coherence of OEM in Section A.1 with a simulated data set of four predictors, where the variables X_1 and X_2 are generated from independent standard normal distributions. The degenerated design matrix is formulated by $X_3 = -X_1$ and $X_4 = -X_2$, where the predictors consist of two pairs of perfectly negative correlated random variables. The true relationship between the response and predictors is $y = -X_3 - 2X_4$.

Figure 4 displays the solution paths for the data using the lasso fitted by `R` packages `glmnet` and `oem` on the same set of tuning parameters λ. The package `lars` gives the same

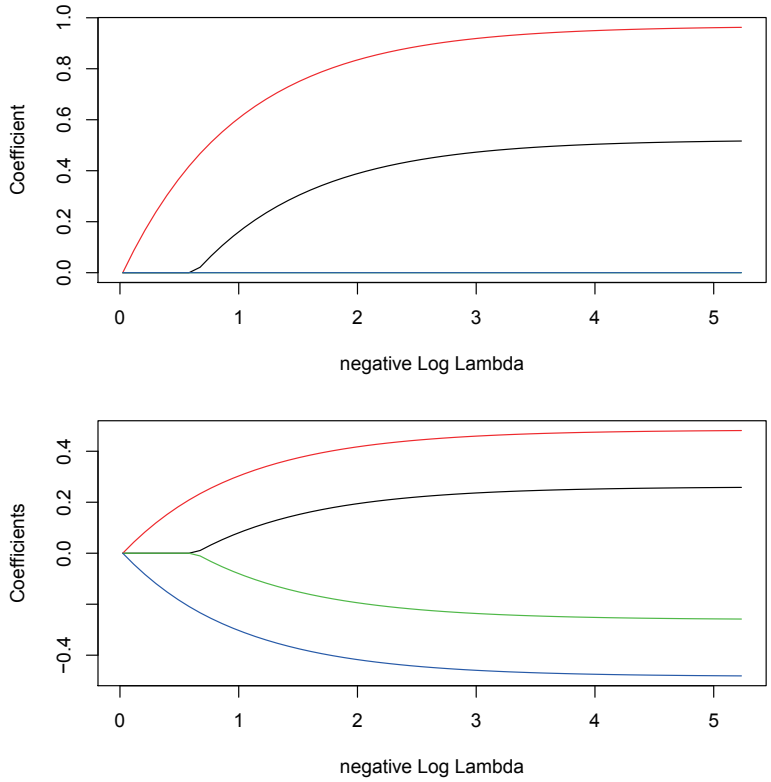

Figure 1: Solution paths of the lasso fitted by CD (the upper panel) and OEM (the lower panel).

solution path as `glmnet`. This figure reveals that OEM estimates the perfectly negative correlated pairs to have exactly the opposite signs but CD only has X_1 and X_2 in the model and fixes X_3 and X_4 to be zero for any λ. This difference is due to the fact that in every iteration, both CD and LARS will find the predictor with the largest improvement on the target function and if more than one coordinates can give better results, only the one with the smallest index will enter the model. OEM considers all the predictors in every iteration equally, so the ones with same contribution to the target will receive equal steps. The grouping coherence property of OEM also holds for non-convex penalties such as SCAD, with the solution paths shown in Figure 1 in the Supplementary Materials, where the same data are used as above for the lasso.

A.2 CONVERGENCE RATE OF OEM

We now derive the convergence rate of the OEM sequence in (7). 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 (7). We capture the convergence rate of the OEM sequence $\{\boldsymbol{\beta}^{(k)}\}$ through $\mathbf{M}$. Recall

$$\boldsymbol{S} = \mathrm{diag}(s_1, \ldots, s_p), \tag{1}$$

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

$$R = \limsup_{k\to\infty} \frac{\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^*\|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} = \limsup_{k\to\infty} \frac{\|\mathbf{M}(\boldsymbol{\beta}^{(k)}) - \mathbf{M}(\boldsymbol{\beta}^*)\|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|}, \tag{2}$$

the global rate of convergence for the OEM sequence. If there is no penalty in (8), i.e., computing the OLS estimator, the global rate of convergence R in (2) 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 $\partial M_i(\boldsymbol{\phi})/\partial\phi_j$ for $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$. For $\boldsymbol{S}$ in (1), $\boldsymbol{J}(\boldsymbol{\beta}^*) = \boldsymbol{S}^{-2}\boldsymbol{A}/d$ with $\boldsymbol{A} = \boldsymbol{\Delta}'\boldsymbol{\Delta}$. Recall that γ_p is the smallest eigenvalue of $\boldsymbol{S}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{S}^{-1}$. We have

$$R_0 = \frac{d - \gamma_p}{d}. \tag{3}$$

For (8), the penalty function $P(\boldsymbol{\beta}; \lambda)$ typically is not sufficiently smooth and R in (2) has no analytic form. Theorem A.2 gives an upper bound of R_{NET}, the value of R for the elastic-net penalty.

Theorem A.2. For $\boldsymbol{S}$ in (1), $R_{\mathrm{NET}} \leqslant R_0$.

Remark 2. Theorem A.2 indicates that, for the same $\boldsymbol{X}$ and $\mathbf{y}$ in (6), the OEM solution for the elastic-net numerically converges faster than its counterpart for the OLS. Since the lasso is a special case of the elastic-net, this theorem holds for the lasso as well.

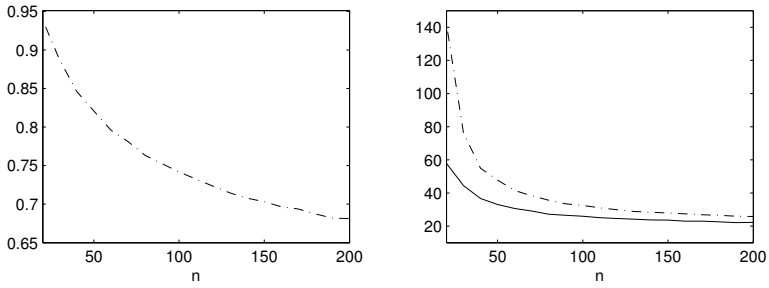

Figure 2: (Left) the average values of R_0 in (3) against increasing n for Example 1; (right) the average iteration numbers against increasing n for Example 1, where the dashed and solid lines denote the OLS estimator and the lasso, respectively.

Remark 3. From (3) and Theorem A.2, the convergence rate of the OEM algorithm depends on $\boldsymbol{S}$ and d. For a given $\boldsymbol{S}$, clearly $d = \gamma_1$ reaches the optimal rate. With such a d, it is desirable to use the $\boldsymbol{S}$ that maximizes γ_p/γ_1. Unfortunately, numerical experiments show that this optimization problem does not have a trivial solution like (5) or $\boldsymbol{S} = \boldsymbol{I}_p$. Therefore, we usually use $\boldsymbol{S} = \boldsymbol{I}_p$ in practice;

Remark 4. For $\boldsymbol{S} = \boldsymbol{I}_p$, the rate in (3) and Theorem A.2 is the fastest when $d = \gamma_1 = \gamma_p$, i.e., if $\boldsymbol{X}$ is orthogonal and standardized. This result suggests that OEM converges faster if $\boldsymbol{X}$ has controlled correlation like from a supersaturated design or a nearly orthogonal Latin hypercube design (Owen 1994).

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

$$\beta_j = (-1)^j \exp\big[-2(j-1)/20\big] \text{ for } j = 1, \ldots, p. \tag{4}$$

Values of $\mathbf{y}$ and $\boldsymbol{\beta}$ are generated by (6) and (4). 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 2 depicts the average values of R_0 in (3) 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 OLS estimator and the lasso, respectively. This

figure indicates that OEM requires *fewer* iterations as n becomes larger, which makes OEM particulary attractive for situations with big data. The OEM sequence for the lasso requires fewer iteration than its counterpart for the OLS, thus validating Theorem A.2.

B EXAMPLES

B.1 Orthogonalization Examples

Example 2. Suppose that $\boldsymbol{X}$ in (6) is orthogonal. Take $d = \gamma_1$ and

$$\boldsymbol{S} = \text{diag}\left[(\sum_{i=1}^{n} x_{i1}^2)^{1/2}, \ldots, (\sum_{i=1}^{n} x_{ip}^2)^{1/2}\right]. \tag{5}$$

Note that $\boldsymbol{S}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{S}^{-1}$ is an identity matrix. Consequently, $t = p$, and $\boldsymbol{\Delta}$ in (21) is empty, which indicates that active orthogonalization will not overshoot.

Example 3. Consider a two-level design in three factors

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

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 $\boldsymbol{S} = \boldsymbol{I}_3$ and $d = \gamma_1$, (21) 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 structure of $\boldsymbol{\Delta}$ is flexible in the sense that the interaction columns do not need to be a product of other two columns.

Example 4. 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 (5), (21) 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.

B.2 Iterative Formulas for Various Penalties

The regression model

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{6}$$

where $\boldsymbol{X} = (x_{ij})$ is an $n \times p$ regression matrix.

The third step of OEM is the M-step,

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}\in\Theta} Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(k)}), \tag{7}$$

Consider a penalized version of (6):

$$\min_{\boldsymbol{\beta}\in\Theta} \left[\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + P(\boldsymbol{\beta};\lambda)\right], \tag{8}$$

where $\boldsymbol{\beta} \in \Theta$, Θ is a subset of $\mathbb{R}^p$, P is a penalty function, and λ is the vector of tuning parameters. To apply the penalty P equally to all the variables, the regression matrix $\boldsymbol{X}$ is standardized so that

$$\sum_{i=1}^{n} x_{ij}^2 = 1, \text{ for } j = 1, \ldots, p. \tag{9}$$

$$\beta_j^{(k+1)} = \underset{\beta_j\in\Theta_j}{\arg\min} \left[d_j\beta_j^2 - 2u_j\beta_j + P_j(\beta_j;\lambda)\right], \text{ for } j = 1, \ldots, p, \tag{10}$$

with $\mathbf{u} = (u_1, \ldots, u_p)'$ defined in (23).

For the model in (6), denote the objective function in (8) by

$$l(\boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + P(\boldsymbol{\beta};\lambda), \tag{11}$$

which is defined on a subset Θ of $\mathbb{R}^p$

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

$$P_j(\beta_j;\lambda) = 2\lambda|\beta_j|, \tag{12}$$

and (10) becomes

$$\beta_j^{(k+1)} = \text{sign}(u_j)\left(\frac{|u_j| - \lambda}{d_j}\right)_+. \tag{13}$$

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 OLS estimator of β_j, and (10) 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 $\Theta_j = \mathbb{R}$,

$$P_j(\beta_j; \lambda) = 2\lambda_1 |\beta_j| + \lambda_2 \beta_j^2. \tag{14}$$

and (10) becomes

$$\beta_j^{(k+1)} = \text{sign}(u_j) \left(\frac{|u_j| - \lambda_1}{d_j + \lambda_2} \right)_+ . \tag{15}$$

5. SCAD (Fan and Li 2001), where $\Theta_j = \mathbb{R}$, $P_j(\beta_j; \lambda) = 2P_\lambda(|\beta_j|)$, and

$$P'_\lambda(\theta) = \lambda I(\theta \leqslant \lambda) + (a\lambda - \theta)_+ I(\theta > \lambda)/(a-1), \tag{16}$$

with $a > 2$, $\lambda \geqslant 0$, and $\theta > 0$. Here, I is the indicator function. If $\boldsymbol{X}$ in (6) is standardized as in (9) with $d_j \geqslant 1$ for all j, (10) becomes

$$\beta_j^{(k+1)} = \begin{cases} \text{sign}(u_j)\big(|u_j| - \lambda\big)_+ / d_j, & \text{when } |u_j| \leqslant (d_j+1)\lambda, \\ \text{sign}(u_j)\big[(a-1)|u_j| - a\lambda\big] / \big[(a-1)d_j - 1\big], & \text{when } (d_j+1)\lambda < |u_j| \leqslant a\lambda d_j, \\ u_j/d_j, & \text{when } |u_j| > a\lambda d_j. \end{cases} \tag{17}$$

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

$$P'_\lambda(\theta) = (\lambda - \theta/a) I(\theta \leqslant a\lambda) \tag{18}$$

with $a > 1$ and $\theta > 0$. If $\boldsymbol{X}$ in (6) is standardized as in (9) with $d_j \geqslant 1$ for all j, (10)

becomes

$$\beta_j^{(k+1)} = \begin{cases} \text{sign}(u_j)a\big(|u_j| - \lambda\big)_+/(ad_j - 1), & \text{when } |u_j| \leqslant a\lambda d_j, \\ u_j/d_j, & \text{when } |u_j| > a\lambda d_j. \end{cases} \tag{19}$$

7. 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 (10) becomes

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

B.3 FIGURES

B.3.1 Illustration of the geometry of OEM

For illustration of the implications of Lemma 1, Figure 3 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 $\mathbb{R}^3$.

B.3.2 Grouping Coherence

We illustrate grouping coherence of OEM in Section A.1 with a simulated data set of four predictors, where the variables X_1 and X_2 are generated from independent standard normal distributions. The degenerated design matrix is formulated by $X_3 = -X_1$ and $X_4 = -X_2$, where the predictors consist of two pairs of perfectly negative correlated random variables. The true relationship between the response and predictors is $y = -X_3 - 2X_4$.

Figure 4 displays the solution paths for the data using the lasso fitted by `R` packages `glmnet` and `oem` on the same set of tuning parameters λ. The package `lars` gives the same solution path as `glmnet`. This figure reveals that OEM estimates the perfectly negative correlated pairs to have exactly the opposite signs but CD only has X_1 and X_2 in the model and fixes X_3 and X_4 to be zero for any λ. This difference is due to the fact that in every iteration, both CD and LARS will find the predictor with the largest improvement on the target function and if more than one coordinates can give better results, only the one with

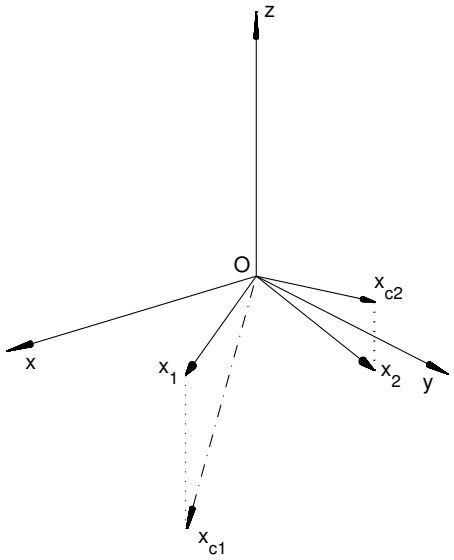

Figure 3: Expand two two-dimensional vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ to two three-dimensional vectors $\mathbf{x}_{c1}$ and $\mathbf{x}_{c2}$ with $\mathbf{x}_{c1}'\mathbf{x}_{c2} = 0$.

the smallest index will enter the model. OEM considers all the predictors in every iteration equally, so the ones with same contribution to the target will receive equal steps. The grouping coherence property of OEM also holds for non-convex penalties such as SCAD, with the solution paths shown in Figure 1 in the Supplementary Materials, where the same data are used as above for the lasso.

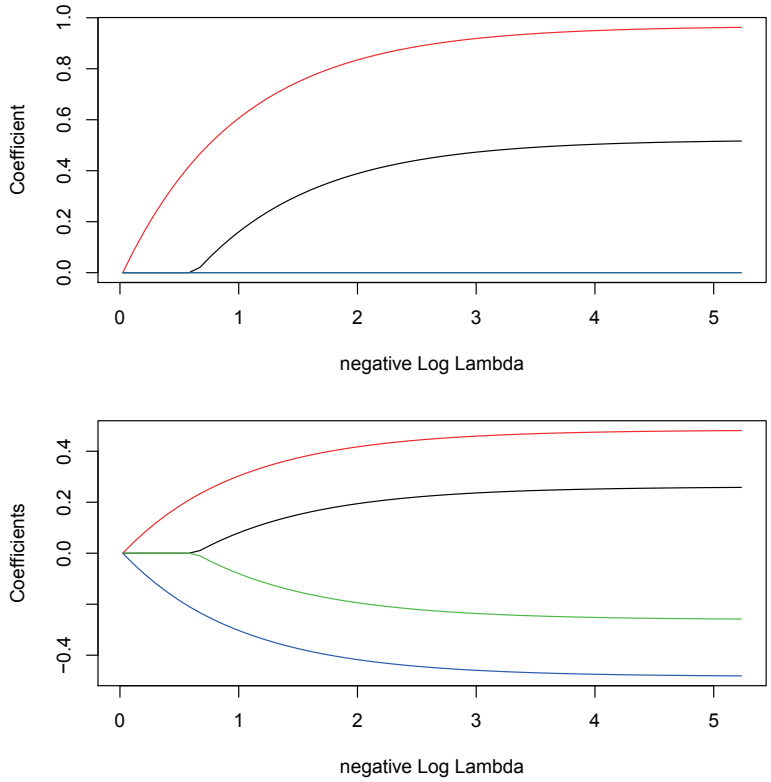

Figure 4: Solution paths of the lasso fitted by CD (the upper panel) and OEM (the lower panel).

B.3.3 Data Analysis Solution Paths

C PROOFS

C.1 Proof of Lemma 1

Define

$$\boldsymbol{B} = \text{diag}(d - \gamma_{t+1}, \ldots, d - \gamma_p) \tag{20}$$

and

$$\boldsymbol{\Delta} = \boldsymbol{B}^{1/2}\boldsymbol{V}_1\boldsymbol{S}, \tag{21}$$

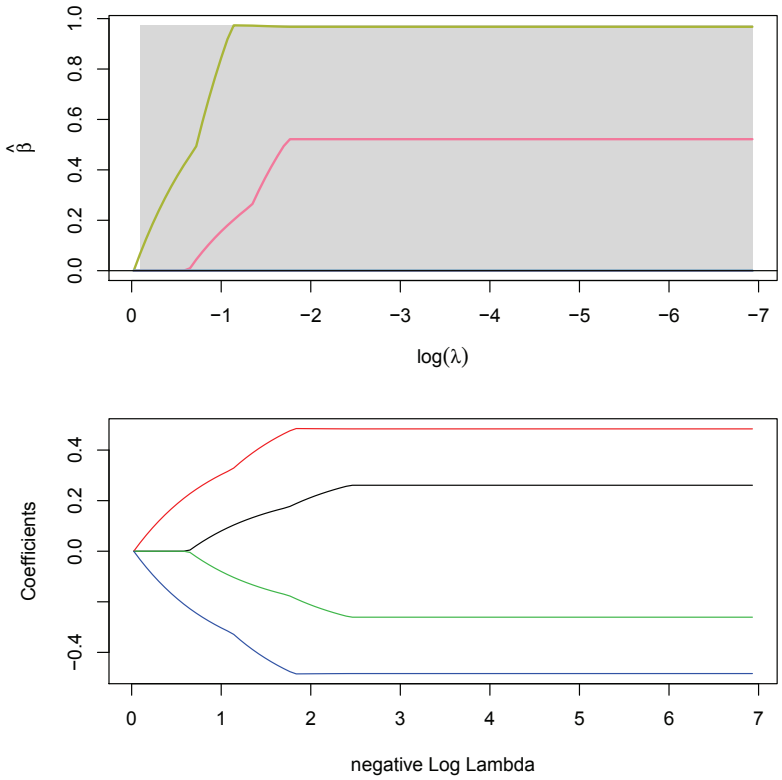

Figure 5: Solution paths of SCAD fitted by CD (from package `ncvreg`) in the upper panel and OEM for the lower panel

where $\boldsymbol{V}_1$ is the submatrix of $\boldsymbol{V}$ consisting of the last $p-t$ rows. Put $\boldsymbol{X}$ and $\boldsymbol{\Delta}$ row by row together to form a complete matrix $\boldsymbol{X}_{\mathrm{c}}$

Proof. From (20) and (21),

$$\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}} = \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}'\boldsymbol{\Delta} = \boldsymbol{S}(\boldsymbol{V}'\boldsymbol{\Gamma}\boldsymbol{V} + \boldsymbol{V}_1'\boldsymbol{B}\boldsymbol{V}_1)\boldsymbol{S}.$$

For the $p \times p$ identity matrix $\boldsymbol{I}_p$,

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

It then follows that $\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}} = \boldsymbol{S}[\boldsymbol{V}'\boldsymbol{\Gamma}\boldsymbol{V} + \boldsymbol{V}'(d\boldsymbol{I}_p - \boldsymbol{\Gamma})\boldsymbol{V}]\boldsymbol{S} = d\boldsymbol{S}^2$, which completes the

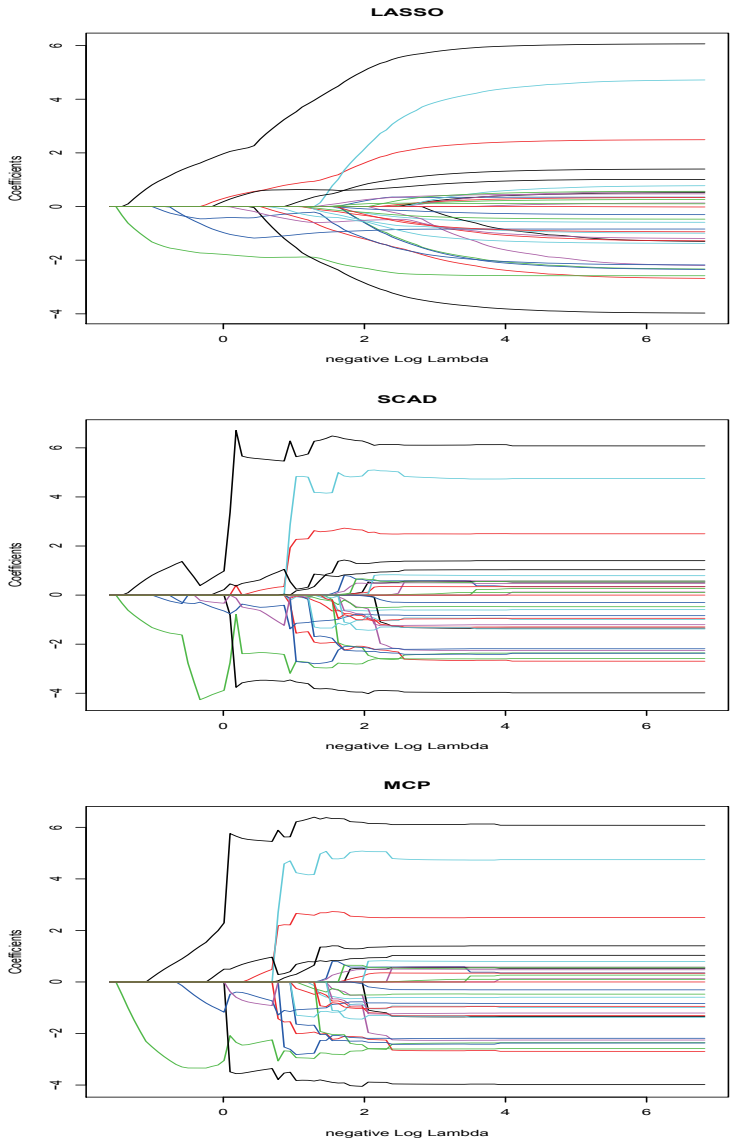

Figure 6: Solution paths for LASSO, SCAD and MCP for US census bureau data.

proof. □

C.2 Proof of Lemma 2

Proof. Let

$$\mathbf{u} = (u_1, \ldots, u_p)' = \boldsymbol{X}'\mathbf{y} + \boldsymbol{A}\boldsymbol{\beta}^{(k)}, \tag{22}$$

We have $\boldsymbol{\Delta}'\boldsymbol{\Delta} = d\boldsymbol{S}^2 - \boldsymbol{X}'\boldsymbol{X} = \boldsymbol{S}(d\boldsymbol{I}_p - \boldsymbol{S}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{S}^{-1})\boldsymbol{S}$. By the eigenvalue decomposition of $\boldsymbol{S}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{S}^{-1}$, the rank of the right side is $p - t$. We complete the proof by noting that the ranks of $\boldsymbol{\Delta}'\boldsymbol{\Delta}$ and $\boldsymbol{\Delta}$ are identical. □

C.3 Proof of Theorem 1

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

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

As $k \to \infty$,

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

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. □

C.4 Proof of Theorem 4

Proof. It suffices to show that $S = \{\boldsymbol{\beta}^*\}$. For $\boldsymbol{\phi} \in \Theta$ with $\boldsymbol{\phi} \neq \boldsymbol{\beta}^*$ and $t > 0$,

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

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

C.5 Proof of Theorem 5

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). □

C.6 Proof of Theorem A.2

Proof. Let $\mathbf{x}_j$ denote the jth column of $n \times p$ matrix $\boldsymbol{X}$ in (6) 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 (15),

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

where

$$f_j(u) = \text{sign}(u)\left(\frac{|u| - \lambda_1}{ds_j^2 + \lambda_2}\right)_+.$$

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

$$\begin{aligned}\frac{|M_j(\boldsymbol{\beta}^{(k)}) - M_j(\boldsymbol{\beta}^*)|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} &= \frac{|f_j(\mathbf{x}_j'\mathbf{y} + \mathbf{a}_j'\boldsymbol{\beta}^{(k)}) - f_j(\mathbf{x}_j'\mathbf{y} + \mathbf{a}_j'\boldsymbol{\beta}^*)|}{|(\mathbf{x}_j'\mathbf{y} + \mathbf{a}_j'\boldsymbol{\beta}^{(k)}) - (\mathbf{x}_j'\mathbf{y} + \mathbf{a}_j'\boldsymbol{\beta}^*)|} \\ &\quad \cdot \frac{|(\mathbf{x}_j'\mathbf{y} + \mathbf{a}_j'\boldsymbol{\beta}^{(k)}) - (\mathbf{x}_j'\mathbf{y} + \mathbf{a}_j'\boldsymbol{\beta}^*)|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} \\ &\leqslant \frac{1}{ds_j^2} \cdot \frac{|\mathbf{a}_j'(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*)|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|}.\end{aligned}$$

Thus,

$$\frac{\|\mathbf{M}(\boldsymbol{\beta}^{(k)}) - \mathbf{M}(\boldsymbol{\beta}^*)\|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} \leqslant \frac{1}{d} \cdot \frac{\|\boldsymbol{S}^{-2}\boldsymbol{A}(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*)\|}{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*\|} \leqslant \frac{d - \gamma_p}{d}.$$

This completes the proof. □

C.7 Proof of Lemma A.1

$$\mathbf{u} = (u_1, \ldots, u_p)' = \boldsymbol{X}'\mathbf{y} + \boldsymbol{A}\boldsymbol{\beta}^{(k)}, \tag{23}$$

Proof. For $\mathbf{u}$ in (13), $\boldsymbol{\alpha}'\mathbf{u} = \boldsymbol{\alpha}'\boldsymbol{X}'\mathbf{y} + \boldsymbol{\alpha}'(d\boldsymbol{I}_p - \boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta}^{(k)} = 0$ for any $\boldsymbol{\alpha} \in \mathcal{E}$ with $\boldsymbol{X}\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\alpha}'\boldsymbol{\beta}^{(k)} = 0$. Then by (13), (17) and (19), 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. □

C.8 Proof of Theorem A.1

Proof. Partition columns of $\boldsymbol{X}$ in (6) 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 3, 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 $\{\boldsymbol{\beta}_2^{(k)}\}$ is an OEM sequence for solving

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \tilde{\boldsymbol{X}}\boldsymbol{\theta}\|^2 + 2\sum_{j=1}^{p-J} |\theta_j|, \tag{24}$$

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 (24) is strictly convex, by Theorem 4, $\{\boldsymbol{\beta}_2^{(k)}\}$ converges to a limit with grouping coherence. This completes the proof. □

D NUMERICAL RESULTS

D.1 Simulation Speed for Moore-Penrose Generalized Inverse-based Least Squares

p	n	OEM	SVD
	10	0.0482	0.1153
	50	0.4203	0.4176
$50,000$	200	1.9159	5.2053
	1000	8.4626	47.7653
	$5,000$	71.8477	440.6741

Table 1: Average runtime (second) comparison between OEM and the SVD least squares method for $p > n$

D.2 Simulation Speed for OEM and CD for SCAD Penalty

p	n	OEM			CD		
		$\rho = 0$	$\rho = 0.2$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.8$
	100	0.8074	0.7742	0.8038	0.1209	0.1278	0.1073
200	200	1.4081	1.3519	1.5147	0.4489	0.2995	0.4458
	400	0.0770	0.0730	0.1361	0.5716	0.6296	0.7261
	250	6.4954	6.7119	6.6319	1.0902	1.5966	1.2908
500	500	9.5813	9.6927	9.8703	2.6176	4.5323	3.5150
	1000	0.5645	0.5636	0.9473	5.0932	5.1822	6.0234
	500	25.9380	26.2019	26.3059	7.4023	8.1918	8.3594
1000	1000	43.3394	44.0728	45.5136	17.4205	22.6260	15.9565
	2000	2.6293	2.5586	4.5881	23.9226	21.3800	34.5852
	100	8.1952	8.4076	8.2836	0.3108	0.4289	0.3863
1200	150	10.9885	11.1938	11.4019	0.5889	0.7644	0.6864
	240	16.2885	16.4598	16.4883	1.7953	2.1551	1.8841

Table 2: Average runtime (seconds) comparison between OEM and CD for SCAD for large p

E ALGORITHM

E.1 OEM: Least Squares

Algorithm 1 OEM algorithm for least-squares

Input: Regression matrix $\mathbf{X}$, response vector $\mathbf{y}$, threshold ϵ

Initialize estimate as $\boldsymbol{\beta}^{(0)}$
Compute $d \leftarrow \gamma_1$ by the Lanczos algorithm
Set $\boldsymbol{A} \leftarrow d\boldsymbol{I}_p - \boldsymbol{X}'\boldsymbol{X}$
$k \leftarrow 0$
while $(\beta_j^{(k)} - \beta_j^{(k-1)})/\beta_j^{(k-1)} > \epsilon$ for $j = 1, \ldots, p$ **do**
  $k \leftarrow k + 1$
  $\mathbf{u}^{(k)} \leftarrow \boldsymbol{X}'\mathbf{y} + \boldsymbol{A}\boldsymbol{\beta}^{(k-1)}$
  $\boldsymbol{\beta}^{(k)} \leftarrow \mathbf{u}^{(k)}/d$
end while

E.2 OEM: Lasso

Algorithm 2 OEM algorithm for lasso least-squares

Input: Regression matrix $\mathbf{X}$, response vector $\mathbf{y}$, threshold ϵ

Initialize estimate as $\boldsymbol{\beta}^{(0)}$
Compute $d \leftarrow \gamma_1$ by the Lanczos algorithm
Set $\boldsymbol{A} \leftarrow d\boldsymbol{I}_p - \boldsymbol{X}'\boldsymbol{X}$
$k \leftarrow 0$
while $(\beta_j^{(k)} - \beta_j^{(k-1)})/\beta_j^{(k-1)} > \epsilon$ for $\{j : \beta_j^{(k-1)} \neq 0\}$ and $\beta_j^{(k)} \neq 0$ for $\{j : \beta_j^{(k-1)} = 0\}$ **do**
  $k \leftarrow k + 1$
  $\mathbf{u}^{(k)} \leftarrow \boldsymbol{X}'\mathbf{y} + \boldsymbol{A}\boldsymbol{\beta}^{(k-1)}$
  $\beta_j^{(k)} \leftarrow \text{sign}(u_j)\left(\frac{|u_j| - \lambda}{d}\right)_+$ for $j = 1, \ldots, p$
end while

F DESCRIPTIONS AND DISCUSSION

F.1 Lanczos Algorithm Description

The Lanczos algorithm generates an orthonormal basis of the Krylov subspaces

$$\mathcal{K}^j(\boldsymbol{X}'\boldsymbol{X}, b) = \text{span}\{b, \boldsymbol{X}'\boldsymbol{X}b, (\boldsymbol{X}'\boldsymbol{X})^2 b, \ldots, (\boldsymbol{X}'\boldsymbol{X})^{j-1} b\}$$

in order to find an efficient approximation of the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$. Specifically, it constructs tridiagonal matrices $\boldsymbol{T_j}$, called Lanczos matrices, and Lanczos vectors $\boldsymbol{V_j}$. The eigenvalues of $\boldsymbol{T_j}$ are the Ritz values $\vartheta_i^{(j)}$, which are essentially the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ restricted to $\mathcal{K}^j(\boldsymbol{X}'\boldsymbol{X}, \boldsymbol{V_j})$. $\vartheta_i^{(j)}$ approximate γ_i as j increases. These can be computed efficiently by the tridiagonal QR algorithm. In practice only a few iterations are required to achieve discretization error as noted in Hanke (1997). This method is attractive because it converges quickly and requires very few arithmetic operations per iteration. Each iteration only requires one matrix-vector multiplication as stated in (Kuczynski and Wozniakowski 1992).

F.2 Data Description

The 36 covariates in the dataset in Section 7.2 include

1. Economic variables like income per capita, household income, poverty.
2. Population distribution like percentages of different races, education levels.
3. Crime rates like violent crimes and property thefts.
4. Miscellaneous variables like Republic, Democratic, death and birth rates.

These variables are in percentage of population of the individual counties.

F.3 Data Example Selected Variables

The selected significant variables include

- Percentage of Household income above $750,000$ dollars, which has large positive effect on the percentage of population change.
- Social security program beneficiaries. The larger the number of beneficiaries in the program, the higher the population change.
- Both the percentages of retired people and under 18 years old have negative effects since they are major sources of migrants leaving the county.
- Birth and death rate with positive and negative effects, respectively.

F.4 Discussion

The algorithm can be sped 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 (7) 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 (6) into G groups and in each iteration, it minimizes the objective function l in (11) 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

$$\boldsymbol{\beta}_g^{(k+1)} = \arg\min_{\boldsymbol{\beta}_g} l(\boldsymbol{\beta}_1^{(k+1)}, \ldots, \boldsymbol{\beta}_{g-1}^{(k+1)}, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{g+1}^{(k)}, \ldots, \boldsymbol{\beta}_G^{(k)}) \text{ for } g = 1, \ldots, G \qquad (25)$$

by OEM until convergence. Note that (25) has a much lower dimension than the iteration in (7). 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.

REFERENCES

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

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.

Hanke, M. (1997). "Superlinear Convergence Rates For The Lanczos Method Applied To Elliptic Operators ," *Numer. Math.*, 77, 487-499.

Kuczynski, J. and Wozniakowski H. (1992), "Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start," *SIAM J. Matrix Anal. Appl.*, 13(4), 1094-1122.

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

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

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

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.

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

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.

Bondell, H. D. and Reich, B. J. (2008), "Simultaneous Regression Shrinkage, Variable Selection and Clustering of Predictors With Oscar," *Biometrics* 64, 115–123.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977),

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

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

Petry, S. and Tutz, G. (2012), "Shrinkage and variable selection by polytopes," *Journal of Statistical Planning and Inference*, 9, 48–64.

Tutz, G. and Ulbricht, J. (2009), "Penalized Regression With Correlation-Based Penalty," *Statistics and Computing*, 19, 239–253 .

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