Advances and Applications in Statistics
Volume 34, Number 2, 2013, Pages 85-105
Published Online: June 2013
Available online at http://pphmj.com/journals/adas.htm
Published by Pushpa Publishing House, Allahabad, INDIA

SPARSE MAVE WITH ORACLE PENALTIES

Ali Alkenani[1,2,*] **and Keming Yu**[1]

[1]Mathematical Science Department
School of Information Systems, Computing and Mathematics
Brunel University
UB8 3PH, Uxbridge, Middlesex, U.K.
e-mail: mapgaja@brunel.ac.uk

[2]Statistics Department
College of Administration and Economics
Al-Qadisiyah University
Al Diwaniyah, Iraq

Abstract

Existing sufficient dimension reduction methods provide us with a way to find sufficient dimensions without the need to pre-specify a model or an error distribution. These methods replace the original variables with low-dimensional linear combinations of predictors without any loss of regression information. However, these methods suffer from the fact that each dimension reduction component is a linear combination of all the original predictors, so that it is difficult to interpret the resulting estimates.

In this article, we propose to combine the shrinkage ideas of the

2010 Mathematics Subject Classification: 62-XX.

Keywords and phrases: dimension reduction, variable selection, MAVE, Adaptive Lasso, SCAD, MCP.

*Corresponding author

Communicated by K. K. Azad

Received January 21, 2013

Adaptive Lasso, SCAD and MCP with a well-known sufficient dimension reduction method, the minimum average variance estimator MAVE, to produce sparse and accurate solutions. The performance of the proposed methods is verified by both simulation and real data analysis.

1. Introduction

In many statistical applications, the dimension p of $\mathbf{X}$ becomes large and therefore the statistical analysis becomes difficult. A usual approach to cope with this problem is to reduce the dimension of the explanatory part of the regression model without much loss of information on regression and without requiring a pre-specified parametric model. This has been obtained through the introduction of sufficient dimension reduction.

Sufficient dimension reduction (SDR) theory (Cook [1]) has been developed to reduce the predictors dimensions, while preserving full regression information and imposing few assumptions. Various methods have been proposed to estimate the SDR space. Some of these methods focus on the knowledge of the central subspace, which is denoted by $S_{Y|\mathbf{X}}$, to answer the question, how does the conditional distribution of $Y|\mathbf{X}$ change with the value assumed by $\mathbf{X}$?. This category includes ordinary least squares (OLS), graphical regression (Cook [2]), sliced average variance estimation (SAVE) (Cook and Weisberg [3]), and sliced inverse regression (SIR) (Li [4]).

In many situations, regression analysis is mostly concerned with deducing the conditional mean of the response given to the predictors, and less concerned with the other sides of the conditional distribution. Cook and Li [5] evolved dimension reduction methods that incorporate this consideration. The authors introduce the idea of the Central Mean Subspace (CMS), a natural inferential object for dimension reduction when the mean function is of interest. There are some dimension reduction methods included in this category, for example, principal Hessian direction (PHD) (Li [6]) and minimum average variance estimation (MAVE) (Xia et al. [7]) which are perhaps the most popular methods to estimate the CMS. However, all the

sufficient dimension methods suffer from the fact that each dimension reduction component is a linear combination of all the original predictors, so that it is difficult to interpret the resulting estimates.

The selection of predictors plays a decisive role in building a multiple regression model. The choice of an appropriate subset of predictors can help to improve prediction precision. Also, in practice, the interpretation of a smaller subset of predictors is easier than a large set of predictors. Variable selection by penalizing the least squares has attracted significant research interest. See for example: least absolute shrinkage and selection operator Lasso (Tibshirani [8]), smoothly clipped absolute deviation SCAD (Fan and Li [9]), Adaptive Lasso (Zou [10]) and the minimax concave penalty MCP (Zhang [11]).

Under the framework of sufficient dimension reduction, the work of Li et al. [12] has produced good results. For example, Ni et al. [13] suggested a shrinkage SIR; Li and Nachtsheim [14] proposed Sparse SIR; and Li [15] unified the inverse dimension reduction methods to have sparse sufficient dimension reduction. Zhou and He [16] suggested Constrained Canonical Correlation (*C*3), which uses CANCOR (Fung et al. [17]) with a l_1 norm constraint. Furthermore, a variable filtering and re-estimation procedure was added to promote sparsity and precision. However, Fung et al. [17] demonstrated that CANCOR is based on the SIR matrix; thus *C*3 may be thought of as an alternative approach to that of Li [15]. Thus, the major thrust of these methods concentrates on the conditional distribution of $\mathbf{X}|Y$ without assuming any particular model. However, they do need certain probabilistic assumptions on the predictors (**X**) such as the linearity condition, which restricts a more general use of these methods. Li and Yin [18] suggested a regularized SIR approach based on the least-squares formulation of SIR. The l_2 regularization is introduced, and an alternating least-squares algorithm is developed, to enable SIR to work with $n < p$ and highly correlated predictors. The l_1 regularization is further introduced to achieve a simultaneous reduction estimation and predictor selection. Wang

and Yin [19] proposed the Sparse MAVE (SMAVE) method which adds an l_1 penalty term to the MAVE loss function to generate a sparse estimate.

Fan and Li [9] studied a class of penalization methods including the Lasso. The authors stated a good penalty function should result in an estimator with three properties. These properties are unbiasedness, sparsity and continuity. They showed that the Lasso shrinkage produces biased estimates for the large coefficients and thus it could be suboptimal in terms of estimation risk. Fan and Li [9] conjectured that the oracle properties do not hold for the Lasso.

In this paper, we consider sufficient dimension reduction and variable selection on the mean function $E(Y|\mathbf{X})$ only. We combine the dimension reduction method MAVE (Xia et al. [7]) with the SCAD (Fan and Li [9]), Adaptive Lasso (Zou [10]) and the MCP (Zhang [11]). Our proposed methods have advantages over SMAVE (Wang and Yin [19]) because all of these penalization methods have the oracle properties and have advantages over sparse inverse dimension reduction methods (Li [15]) in that it does not require any particular distribution on $\mathbf{X}$ and it can exhaustively estimate the dimensions in the conditional mean function.

The rest of the paper is organized as follows: In Section 2, a brief review of sufficient dimension reduction for the mean function and MAVE is given. SMAVE is reviewed in Section 3. Sparse MAVE with Adaptive Lasso penalty, SCAD and MCP penalties are introduced in Sections 4, 5 and 6, respectively. Simulation studies are conducted under different settings in Section 7. The applications of the methods using two sets of real data are reported in Section 8. Finally, the conclusions are summarized in Section 9.

2. Sufficient Dimension Reduction for the Mean Function and MAVE

For regression problems with a scalar response variable Y on a $p \times 1$ predictor vector $\mathbf{X}$, assume the following model:

$$Y = f(X_1, X_1, ..., X_p) + \varepsilon, \tag{1}$$

where $f(X_1, X_1, ..., X_p) = E(Y|\mathbf{X})$ and $E(\varepsilon|\mathbf{X}) = 0$. The aim of sufficient dimension reduction for the mean function is to find a subset S of the predictor space such that

$$Y \perp\!\!\!\perp E(Y|\mathbf{X})|P_s X, \tag{2}$$

where $\perp\!\!\!\perp$ indicates independence and $P_{(.)}$ stands for a projection operator with respect to the standard inner product. Subspaces satisfying condition (2) are called *mean dimension reduction subspaces* (Cook and Li [5]). Thus, if $d = \dim(S)$ and $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_d)$ is a basis for S, the predictors X can be replaced by linear combinations $\boldsymbol{\beta}_1^T\mathbf{X}, \boldsymbol{\beta}_2^T\mathbf{X}, ..., \boldsymbol{\beta}_d^T\mathbf{X}$, $d \le p$ without loss of information on the conditional mean function. That is, $f(X_1, X_1, ..., X_p)$ $= f(\boldsymbol{\beta}^T\mathbf{X})$. When the intersection of all subspaces satisfies condition (2), it is called the *central mean subspace* (CMS) (Cook and Li [5]) and denoted by $S_{E(Y|\mathbf{X})}$. $S_{E(Y|\mathbf{X})}$ is assumed existent throughout the paper. Several methods are available for estimating $S_{E(Y|\mathbf{X})}$, and one of the most well-known methods of them is MAVE (Xia et al. [7]). We will describe MAVE in details as follows:

Xia et al. [7] proposed MAVE such that the matrix $\mathbf{B}$ is the solution of

$$\min_{\mathbf{B}} \{E[Y - E(Y|\mathbf{B^T X})]^2\}, \tag{3}$$

where $\mathbf{B^T B} = \mathbf{I_d}$. The conditional variance given $\mathbf{B^T X}$ is

$$\boldsymbol{\sigma}^2_{\mathbf{B}}(\mathbf{B^T X}) = E[\{Y - E(Y|\mathbf{B^T X})\}^2 | \mathbf{B^T X}]. \tag{4}$$

Thus,

$$\min_{\mathbf{B}} E[Y - E(Y|\mathbf{B^T X})]^2 = \min_{\mathbf{B}} E\{\boldsymbol{\sigma}^2_{\mathbf{B}}(\mathbf{B^T X})\}. \tag{5}$$

For any given $\mathbf{X_0}$, $\boldsymbol{\sigma}^2_{\mathbf{B}}(\mathbf{B^T X_0})$ can be approximated using local linear smoothing as

$$\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{B}}(\mathbf{B^T X_0}) \approx \sum_{i=1}^{n} \{Y_i - E(Y_i \,|\, \mathbf{B^T X_i})\}^2 w_{i0}$$

$$\approx \sum_{i=1}^{n} [Y_i - \{a_0 + \mathbf{b}_0^T \mathbf{B^T}(\mathbf{X_i} - \mathbf{X_0})\}]^2 w_{i0},$$

where $a_0 + \mathbf{b}_0^T \mathbf{B^T}(\mathbf{X_i} - \mathbf{X_0})$ is the local linear expansion of $E(Y_i \,|\, \mathbf{B^T X_i})$ at $\mathbf{X_0}$, and $w_{i0} \geq 0$ are the kernel weights centred at $\mathbf{B^T X_0}$ with $\sum_{i=1}^{n} w_{i0} = 1$. So the problem of finding $\mathbf{B}$ is equivalent to that of solving the following optimization:

$$\sum_{j=1}^{n} \sum_{i=1}^{n} [Y_i - \{a_j + \mathbf{b}_j^T \mathbf{B^T}(\mathbf{X_i} - \mathbf{X_j})\}]^2 w_{ij}. \tag{6}$$

3. Sparse MAVE (SMAVE)

Wang and Yin [19] proposed the SMAVE. The authors add an l_1 penalty term to the MAVE loss function in (6) to produce a sparse estimate. The authors solve the following minimization problem:

$$\sum_{j=1}^{n} \sum_{i=1}^{n} [Y_i - \{a_j + \mathbf{b}_j^T \mathbf{B^T}(\mathbf{X_i} - \mathbf{X_j})\}]^2 w_{ij} + \lambda \sum_{k=1}^{p} | \boldsymbol{\beta}_{m,k} |, \tag{7}$$

$m = 1, ..., d$.

They suppose that d is known, then propose a modified BIC criterion to estimate d. The algorithm for SMAVE is as follows:

1. Initialize $m = 1$, and set $\mathbf{B} = \boldsymbol{\beta_0}$, any arbitrary $p \times 1$ vector.

2. For given $\mathbf{B}$, obtain $(a_j, \mathbf{b}_j)$, where $j = 1, ..., n$, by solving the following minimization problem:

$$\min_{a_j, \mathbf{b}_j, j=\mathbf{1}, \ldots, \mathbf{n}} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} [Y_i - \{a_j + \mathbf{b}_j^T \mathbf{B}^{\mathbf{T}} (\mathbf{X_i} - \mathbf{X_j})\}]^2 w_{ij} \right). \tag{8}$$

3. For given $(\hat{a}_j, \hat{\mathbf{b}}_j)$, $j = 1, \ldots, n$, $\boldsymbol{\beta}_{\boldsymbol{mLasso}}$ can be obtained by solving the following minimization problem:

$$\min_{\mathbf{B}: \mathbf{B}^T \mathbf{B} = \mathbf{I}_m} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} [Y_i - \{\hat{a}_j + \hat{\mathbf{b}}_{\mathbf{j}}^T (\hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}, \ldots, \hat{\boldsymbol{\beta}}_{\boldsymbol{m-1}}, \boldsymbol{\beta}_{\boldsymbol{m}})^T (\mathbf{X_i} - \mathbf{X_j})\}]^2 w_{ij} + \lambda \sum_{k=1}^{p} |\boldsymbol{\beta}_{m,k}| \right). \tag{9}$$

4. Replace the mth column of $\mathbf{B}$ by $\hat{\boldsymbol{\beta}}_{\boldsymbol{mLasso}}$, and repeat steps 2 and 3 until convergence.

5. Update $\mathbf{B}$ by $(\hat{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{Lasso}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}\boldsymbol{Lasso}}, \ldots, \hat{\boldsymbol{\beta}}_{\boldsymbol{mLasso}}, \boldsymbol{\beta}_{\mathbf{0}})$, and set m to be $m + 1$.

6. If $m < d$, then continue steps 2 to 5 until $m = d$.

Wang and Yin [19] adopted the refined multidimensional Gaussian Kernel proposed by Xia et al. [7] for MAVE,

$$w_{ij} = K_h\{\hat{\mathbf{B}}^{\mathbf{T}}(\mathbf{X_i} - \mathbf{X_j})\} \Big/ \sum_{i=1}^{n} K_h\{\hat{\mathbf{B}}^{\mathbf{T}}(\mathbf{X_i} - \mathbf{X_j})\}$$

and the optimal bandwidth in the sense of minimizing the mean integrated squared errors. Also, they used the Gaussian product kernel, and $h_{opt} = A(d) n^{-1/(4+d)}$, where $A(d) = \left\{ \frac{4}{(d+2)} \right\}^{1/(4+d)}$, and d is the dimension of the kernel function.

4. Sparse MAVE with Adaptive Lasso Penalty (ALMAVE)

Fan and Li [9] studied a class of penalization methods including the Lasso. They showed that the Lasso shrinkage produces biased estimates for the large coefficients, and thus it could be suboptimal in terms of estimation

risk. Fan and Li [9] conjectured that the oracle properties do not hold for the Lasso. The Adaptive Lasso can be viewed as a generalization of the Lasso penalty. Basically, the idea is to penalize the coefficients of different covariates at a different level by using adaptive weights. In the case of least squares regression, Zou [10] proposed the Adaptive Lasso in which adaptive weights are used to penalize different coefficients in the l_1 penalty. The author showed that the Adaptive Lasso benefits from the oracle properties that Lasso does not have. Zou [10] defined the Adaptive Lasso as follows:

$$\min_{\boldsymbol{\beta}}\left(\sum_{i=1}^{n}(Y_i-\boldsymbol{\beta}^T\mathbf{X_i})^2+\lambda_n\sum_{k=1}^{p}\widetilde{w}_k|\beta_k|\right), \tag{10}$$

where $\lambda > 0$ is the tuning parameter controlling the amount of penalty given. The weights are set to be $\widetilde{w}_k = 1/|\widetilde{\beta}_k|^{\delta}$, $k = 1, ..., p$, $\widetilde{\beta}$ is non-penalized regression estimates and $\delta > 0$.

Sparse MAVE with the Adaptive Lasso penalty has been proposed as follows:

$$\min_{\mathbf{B}}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}[Y_i-\{a_j+\mathbf{b_j}^T\mathbf{B}^T(\mathbf{X_i}-\mathbf{X_j})\}]^2w_{ij}+\lambda_n\sum_{k=1}^{p}\widetilde{w}_k|\boldsymbol{\beta}_{m,k}|\right) \tag{11}$$

for $m = 1, ..., d$.

The algorithm of Sparse MAVE with the Adaptive Lasso penalty is similar to the algorithm in Section 3, except in step 3, for given $(\hat{a}_j, \hat{\mathbf{b}}_j)$, $j = 1, ..., n$, $\boldsymbol{\beta}_{\boldsymbol{mALasso}}$ can be obtained by solving the following minimization problem:

$$\min_{\mathbf{B}:\mathbf{B}^T\mathbf{B}=\mathbf{I}_m}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}[Y_i-\{\hat{a}_j+\hat{\mathbf{b}}_\mathbf{j}^T(\hat{\boldsymbol{\beta}}_\mathbf{1},\hat{\boldsymbol{\beta}}_\mathbf{2},...,\hat{\boldsymbol{\beta}}_{\boldsymbol{m-1}},\boldsymbol{\beta}_{\boldsymbol{m}})^T(\mathbf{X_i}-\mathbf{X_j})\}]^2w_{ij}+\lambda_n\sum_{k=1}^{p}\widetilde{w}_k|\boldsymbol{\beta}_{m,k}|\right) \tag{12}$$

and then we follow the same steps in the algorithm in Section 3.

5. Sparse MAVE with SCAD Penalty (SCADMAVE)

Fan and Li [9] demonstrated the oracle properties for the SCAD in the variable selection aspect. The SCAD penalty Fan and Li [9] defined on $[0, \infty)$ is given by

$$p_{SCAD\,\lambda, c}(\theta) = \begin{cases} \lambda\theta & \text{if } \theta \le \lambda, \\ \dfrac{c\lambda\theta - 0.5(\theta^2 + \lambda^2)}{c-1} & \text{if } \lambda < \theta \le c\lambda, \\ \dfrac{\lambda^2(c^2-1)}{2(c-1)} & \text{if } \theta > c\lambda, \end{cases} \tag{13}$$

and its first derivative is given by

$$p'_{SCAD\,\lambda}(\theta) = \begin{cases} \lambda & \text{if } \theta \le \lambda, \\ \dfrac{c\lambda - \theta}{c-1} & \text{if } \lambda < \theta \le c\lambda, \\ 0 & \text{if } \theta > c\lambda, \end{cases} \tag{14}$$

where $c > 2$ and $\lambda \ge 0$ are tuning parameters.

The SCAD penalized regression solves the following minimization problem:

$$\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}^T \mathbf{X_i})^2 + n \sum_{k=1}^{p} p_{SCAD\,\lambda, c}(\beta_k) \right). \tag{15}$$

Sparse MAVE with SCAD penalty has been proposed as follows:

$$\min_{\mathbf{B}} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} [Y_i - \{a_j + \mathbf{b_j}^T \mathbf{B}^T (\mathbf{X_i} - \mathbf{X_j})\}]^2 w_{ij} + n \sum_{k=1}^{p} p_{SCAD\,\lambda, c}(\beta_k) \right). \tag{16}$$

The algorithm of Sparse MAVE with SCAD penalty is similar to the algorithm in Section 3, except in step 3, for given $(\hat{a}_j, \hat{\mathbf{b}}_j)$, $j = 1, ..., n$,

$\boldsymbol{\beta}_{mSCAD}$ can be obtained by solving the following minimization problem:

$$\min_{\mathbf{B}:\mathbf{B}^T\mathbf{B}=\mathbf{I}_m}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}[Y_i-\{\hat{a}_j+\hat{\mathbf{b}}_{\mathbf{j}}^T(\hat{\boldsymbol{\beta}}_{\mathbf{1}},\hat{\boldsymbol{\beta}}_{\mathbf{2}},...,\hat{\boldsymbol{\beta}}_{\boldsymbol{m-1}},\boldsymbol{\beta}_{\boldsymbol{m}})^T(\mathbf{X_i}-\mathbf{X_j})\}]^2 w_{ij}\right.$$

$$\left.+\,n\sum_{k=1}^{p}p_{SCAD\lambda,c}(\beta_k)\right) \tag{17}$$

and then we follow the same steps in the algorithm in Section 3.

6. Sparse MAVE with MCP Penalty (MCPMAVE)

Zhang [11] proposed a minimax concave penalty MCP. The MCP provides the convexity of the penalized loss in sparse regions to the greatest extent given certain thresholds for variable selection and unbiasedness.

The MCP Zhang [11] defined on $[0,\infty)$ is given by

$$p_{MCP\lambda,c}(\theta)=\begin{cases}\lambda\theta-\dfrac{\theta^2}{2c} & \text{if } \theta\le c\lambda,\\[2ex] \dfrac{1}{2}c\lambda^2 & \text{if } \theta>c\lambda,\end{cases} \tag{18}$$

and its first derivative is given by

$$p'_{MCP\lambda,c}(\theta)=\begin{cases}\lambda-\dfrac{\theta}{c} & \text{if } \theta\le c\lambda,\\[2ex] 0 & \text{if } \theta>c\lambda,\end{cases} \tag{19}$$

where $c>1$ and $\lambda\ge 0$ are tuning parameters. The rationale behind the penalty can be understood by considering its derivative: MCP begins by applying the same rate of penalization as the Lasso, but continuously relaxes that penalization until, when $\theta>c\lambda$, the rate of penalization drops to 0.

The MCP penalized regression solves the following minimization problem:

$$\min_{\boldsymbol{\beta}}\left(\sum_{i=1}^{n}(Y_i-\boldsymbol{\beta}^T\mathbf{X_i})^2+n\sum_{k=1}^{p}p_{MCP\,\lambda,c}(\beta_k)\right). \tag{20}$$

Sparse MAVE with MCP penalty has been proposed as follows:

$$\min_{\mathbf{B}}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}[Y_i-\{a_j+\mathbf{b_j}^T\mathbf{B}^T(\mathbf{X_i}-\mathbf{X_j})\}]^2w_{ij}+n\sum_{k=1}^{p}p_{MCP\,\lambda,c}(\boldsymbol{\beta}_k)\right). \tag{21}$$

The algorithm of Sparse MAVE with the MCP penalty is similar to the algorithm in Section 3, except in step 3, for given $(\hat{a}_j, \hat{\mathbf{b}}_j)$, $j = 1, ..., n$, $\boldsymbol{\beta}_{\boldsymbol{mSCAD}}$ can be obtained by solving the following minimization problem:

$$\min_{\mathbf{B}:\mathbf{B}^T\mathbf{B}=\mathbf{I}_m}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}[Y_i-\{\hat{a}_j+\hat{\mathbf{b}}_\mathbf{j}^T(\hat{\boldsymbol{\beta}}_\mathbf{1},\hat{\boldsymbol{\beta}}_\mathbf{2},...,\hat{\boldsymbol{\beta}}_{\boldsymbol{m-1}},\boldsymbol{\beta}_{\boldsymbol{m}})^T(\mathbf{X_i}-\mathbf{X_j})\}]^2w_{ij}\right.$$

$$\left.+\,n\sum_{k=1}^{p}p_{MCP\,\lambda,c}(\beta_k)\right) \tag{22}$$

and then we follow the same steps in the algorithm in Section 3.

The R codes for the proposed methods are available from the authors.

7. A Simulation Study

Many simulations have been carried out in order to check the feasibility of the proposed methods and some typical examples are reported below:

Example 1. $R = 200$ data-sets were generated with size $n = 200$ observations from the model $y=\dfrac{(\beta_1^T\mathbf{X})}{\{0.5+(\beta_2^T\mathbf{X}+1.5)^2\}}+0.2\,\varepsilon$, where $\mathbf{X} = (x_1, ..., x_{10})^T$, x_i and ε are independent and are identically distributed standard normal random variables, $\boldsymbol{\beta}_1 = (1, 0, ..., 0)^T$ and $\boldsymbol{\beta}_2 = (0, 1, 0, ..., 0)^T$. $S_{E(Y|\mathbf{X})} = span(\mathbf{B_2})$.

Example 2. $R = 200$ data-sets were generated with size $n = 60$ and 120 observations from the linear model $y = \beta^T \mathbf{X} + 0.5\,\varepsilon$, where $\mathbf{X} = (x_1, ..., x_{24})^T$, x_i and ε are independent and are identically distributed standard normal random variables, $\boldsymbol{\beta} = (1, 1, 1, 0, ..., 0)^T$. $S_{E(Y|\mathbf{X})} = span(\mathbf{B_1})$. To assess the impact of correlated predictors on the performance of our proposed methods, we also adopt Tibshirani's [8] correlated predictors setting to generate $\mathbf{X}$ from a multivariate normal with $cov(x_i, x_j) = 0.5^{|i-j|}$ for this model.

Example 3. $R = 200$ data-sets were generated with size $n = 200$ observations from the model $y = \text{sign}(\beta_1^T \mathbf{X}) \log(|\beta_2^T \mathbf{X} + \mathbf{5}|) + 0.2\,\varepsilon$, where $\mathbf{X} = (x_1, ..., x_{20})^T$, x_i and ε are independent and are identically distributed standard normal random variables. There are three different forms for β_1 and β_2 as follows:

(1) $\boldsymbol{\beta}_1 = (1, 1, 1, 1, 0, ..., 0)^T$ and $\boldsymbol{\beta}_2 = (0, ..., 0, 1, 1, 1, 1)^T$.

(2) $\boldsymbol{\beta}_1 = (1, 1, 0.1, 0.1, 0, ..., 0)^T$ and $\boldsymbol{\beta}_2 = (0, ..., 0, 0.1, 0.1, 1, 1)^T$.

(3) $\boldsymbol{\beta}_1 = (1, ..., 1, 0, ..., 0)^T$ and $\boldsymbol{\beta}_2 = (0, ..., 0, 1, ..., 1)^T$, where there are 10 coordinates equal to one in each direction.

$S_{E(Y|\mathbf{X})} = span(\mathbf{B_2})$.

Example 4. $R = 200$ data-sets were generated with size $n = 60$ and 120 observations from the linear model $y = \beta^T \mathbf{X} + 0.5\,\varepsilon$, where $\mathbf{X} = (x_1, ..., x_{24})^T$, x_i and ε are standard normal random variables, $\boldsymbol{\beta} = (1, ..., 1)^T$. $S_{E(Y|\mathbf{X})} = span(\mathbf{B_1})$. This model (Zhou and He [16, model 4]) has no sparseness in the predictors.

After we write the first term in equations (12), (17) and (22) in least squares form, we use the functions (adalasso) from Package 'parcor' (Kraemer and Schaefer [20]), ncvreg (, penalty=c("SCAD")), and ncvreg (, penalty=c("MCP")) from Package 'ncvreg' in R (Breheny and Huang [21]) to do the computations in equations (12), (17) and (22), respectively.

To evaluate the estimation accuracy, we report the mean and standard deviation of the absolute correlation $|r_i|$ between the estimated predictor $\hat{\beta}_j^T\mathbf{X}$ and the true one $\beta_j^T\mathbf{X}$ and the mean and standard deviation of the mean squared error, $\hat{E}(\hat{\beta}_j^T\mathbf{X} - \beta_j^T\mathbf{X})^2$.

According to the mean and standard deviation of the absolute correlation $|r_i|$ between the estimated predictor $\hat{\beta}_j^T\mathbf{X}$ and the true one $\beta_j^T\mathbf{X}$, and the mean and standard deviation of the mean squared error, $\hat{E}(\hat{\beta}_j^T\mathbf{X} - \beta_j^T\mathbf{X})^2$. From Tables 1, 2 and 3, it can be seen that the proposed methods (ALMAVE and SCADMAVE) have a better performance than the other methods for all cases under consideration except in Example 3, case (1) where the proposed methods (ALMAVE and MCPMAVE) were the best two methods amongst of all the methods. Also, we can see from Table 4 that the proposed methods (SCADMAVE and MCPMAVE) have a better performance than the other methods. In general, this indicates that the proposed methods give precise estimates and these methods are more significantly efficient than the SMAVE method.

It can be observed that in all of the examples, the proposed methods produce a lower mean squared error and bigger absolute correlation $|r_i|$ than the SMAVE method. The variations in the ALMAVE, SCADMAVE and MCPMAVE estimates are approximately similar in the majority of cases and less than the variations in the estimate of the SMAVE method.

8. Real Data Examples

8.1. Air pollution data

In this subsection, we illustrate our methods through an analysis of air pollution data. The data consist of $n = 500$ observations that originate in a study where air pollution on a road is related to traffic volume and meteorological variables.

The data-set is available at the website http://lib.stat.cmu.edu/datasets/NO2.dat. The response variable Y is hourly values of the logarithm of the concentration of Nitrogen dioxide (particles), measured at Alnabru in Oslo, Norway, between October 2001 and August 2003. The $p = 7$ predictor X variables are the logarithm of the number of cars per hour (x_1), temperature 2m above ground $(x_2$, degree C), wind speed $(x_3$, m/s), the temperature difference between 25 and 2m above ground $(x_4$, degree C), wind direction $(x_5$, degrees between 0 and 360), hour of day (x_6) and day number from October 1st, 2001 (x_7).

Table 5 reports the values of the adjusted R-squared for the model fit based on air pollution data for all the studied methods. All of these methods find nonlinear structure, which can be approximated by a cubic fit, and the adjusted R-squared is a little bit larger than SMAVE (Wang and Yin [19]) for the ALMAVE method (adjusted R-squared=0.94), and it is similar to the SMAVE for the other methods (adjusted R-squared=0.93).

Table 6 presents the prediction error of the models which are selected by the studied methods based on air pollution data for the cubic fit. It is clear that all of the proposed methods have less prediction error than the SMAVE method. This means, these methods have better performance than the SMAVE method.

Figure 1 presents a plot and explains the estimated $\hat{\beta}$'s which are estimated by studied methods based on air pollution data. It can be seen from this figure that the estimated coefficients for the SMAVE, SCADMAVE and

MCPMAVE methods were approximately similar maybe because they have the same value for the adjusted R-squared.

8.2. Body fat data

Percentage of body fat is an important measure of health, which can be accurately estimated by underwater weighing techniques. These techniques often require special equipment and are sometimes not easily achieved, thus fitting percentage body fat to simple body measurements is a convenient way to predict body fat. Johnson [22] introduced a data-set in which percentage body fat and 13 simple body measurements (such as weight, height and abdomen circumference) are recorded for 252 men. The data-set is available at the package ('mfp') in R. The response variable Y is the percent body fat (%). The $k = 13$ predictor variables x are the age (years) x_1, the weight (pounds) x_2, the height (inches) x_3, the neck circumference (cm) x_4, the chest circumference (cm) x_5, the abdomen circumference (cm) x_6, the hip circumference (cm) x_7, the thigh circumference (cm) x_8, the knee circumference (cm) x_9, the ankle circumference (cm) x_{10}, the extended biceps circumference x_{11}, the forearm circumference (cm) x_{12} and the wrist circumference (cm) x_{13}.

Table 7 reports the values of the adjusted R-squared for the model fit based on body fat data for all the studied methods. All of these methods find the nonlinear structure better than the linear, and the adjusted R-squared is same for the all methods and for the all fitted models.

Table 8 presents the prediction error of the models which are selected by the studied methods based on body fat data for the cubic fit. It is clear that all of the proposed methods have better performance than the SMAVE method. In general, the results are similar to those which are based on the air pollution data in Table 6.

Figure 2 presents a plot and explains the estimated $\hat{\beta}$'s which are estimated by studied methods based on body fat data. It can be seen from this

figure that there are no big differences among the estimated coefficients for all of the methods.

9. Conclusions

In this study, Sparse MAVE with Adaptive Lasso, SCAD and MCP penalties methods have been proposed. The proposed methods have been theoretically investigated and numerically compared with Sparse MAVE (Wang and Yin [19]). In order to assess the numerical performance, a simulation study was conducted based on the models in Examples 1, 2, 3 and 4 as described in Section 7. From the simulation study and the real data examples, it can be concluded that the proposed methods perform well in comparison to Sparse MAVE (Wang and Yin [19]) and thus the authors believe that the proposed methods are useful practically.

Acknowledgements

The authors wish to thank the Editor, an Associate Editor and anonymous referees. Also, we thank Prof. Xiangrong Yin and Dr. Qin Wang for sending us the code for the SMAVE method in Wang and Yin [19].

References

[1] R. Cook, Regression Graphics: Ideas for Studying the Regression through Graphics, Wiley, New York, 1998.

[2] R. D. Cook, On the interpretation of regression plots, J. Amer. Statist. Assoc. 89 (1994), 177-189.

[3] R. D. Cook and S. Weisberg, Discussion of Li (1991), J. Amer. Statist. Assoc. 86 (1991), 328-332.

[4] K. Li, Sliced inverse regression for dimension reduction (with discussion), J. Amer. Statist. Assoc. 86 (1991), 316-342.

[5] R. D. Cook and B. Li, Dimension reduction for the conditional mean in regression, Ann. Stat. 30 (2002), 455-474.

[6] K. C. Li, On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma, J. Amer. Statist. Assoc. 87 (1992), 1025-1039.

[7] Y. Xia, H. Tong, W. Li and L. Zhu, An adaptive estimation of dimension reduction space, J. Royal Stat. Soc. Ser. B 64 (2002), 363-410.

[8] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Royal Stat. Soc. Ser. B 58 (1996), 267-288.

[9] J. Fan and R. Z. Li, Variable selection via non-concave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001), 1348-1360.

[10] H. Zou, The adaptive Lasso and its oracle properties, J. Amer. Statist. Assoc. 101(476) (2006), 1418-1429.

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

[12] L. Li, R. D. Cook and C. J. Nachtsheim, Model-free variable selection, J. Royal Stat. Soc. Ser. B 67 (2005), 285-299.

[13] L. Ni, R. D. Cook and C. L. Tsai, A note on shrinkage sliced inverse regression, Biometrika 92 (2005), 242-247.

[14] L. Li and C. J. Nachtsheim, Sparse sliced inverse regression, Technometrics 48 (2006), 503-510.

[15] L. Li, Sparse sufficient dimension reduction, Biometrika 94 (2007), 603-613.

[16] J. Zhou and X. M. He, Dimension reduction based on constrained canonical correlation and variable filtering, Ann. Statist. 36 (2008), 1649-1668.

[17] W. K. Fung, X. He, L. Liu and P. Shi, Dimension reduction based on canonical correlation, Statist. Sinica 12 (2002), 1093-1113.

[18] L. Li and X. Yin, Sliced inverse regression with regularizations, Biometrics 64 (2008), 124-131.

[19] Q. Wang and X. Yin, A nonlinear multi-dimensional variable selection method for high dimensional data: sparse MAVE, Comput. Statist. Data Anal. 52 (2008), 4512-4520.

[20] N. Kraemer and J. Schaefer, Package “parcor”, http://cran.r-project.org/web/packages/parcor/parcor.pdf

[21] P. Breheny and J. Huang, Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, Ann. Appl. Stat. 5 (2011), 232-253.

[22] R. W. Johnson, Fitting percentage of body fat to simple body measurements, J. Stat. Edu. 4 (1996), 236-237.

Table 1. Simulation results for the studied methods based on the model in Example 1

	$\hat{\beta}_1$				$\hat{\beta}_2$			
	$\lvert r_1 \rvert$ Mean	$\lvert r_1 \rvert$ SD	MSE Mean	MSE SD	$\lvert r_2 \rvert$ Mean	$\lvert r_2 \rvert$ SD	MSE Mean	MSE SD
SMAVE	0.9516	0.0497	0.0005	0.0005	0.8090	0.1266	0.0060	0.0080
ALMAVE	**0.9791**	**0.0422**	**0.0003**	**0.0001**	**0.9840**	**0.0338**	**0.0003**	**0.0006**
SCADMAVE	**0.9619**	**0.0425**	**0.0004**	**0.0004**	**0.8511**	**0.1103**	**0.0028**	**0.0065**
MCPMAVE	0.9590	0.0479	0.0004	0.0005	0.8263	0.1146	0.0058	0.0075

Table 2. Simulation results for the studied methods based on the model in Example 2

	Independent predictors				Correlated predictors			
	$\lvert r \rvert$ Mean	$\lvert r \rvert$ SD	MSE Mean	MSE SD	$\lvert r \rvert$ Mean	$\lvert r \rvert$ SD	MSE Mean	MSE SD
		$\boldsymbol{n} = \mathbf{60}$				$\boldsymbol{n} = \mathbf{60}$		
SMAVE	0.9796	0.0075	0.0155	0.0400	0.9479	0.1085	0.0119	0.0088
ALMAVE	**0.9918**	**0.0074**	**0.0147**	**0.0360**	**0.9866**	**0.0348**	**0.0112**	**0.0074**
SCADMAVE	**0.9919**	**0.0074**	**0.0149**	**0.0363**	**0.9866**	**0.0349**	**0.0111**	**0.0074**
MCPMAVE	0.9920	0.0100	0.0157	0.0380	0.9865	0.0350	0.0111	0.0088
		$\boldsymbol{n} = \mathbf{120}$				$\boldsymbol{n} = \mathbf{120}$		
SMAVE	0.98985	0.00465	0.0050	0.0100	0.9934	0.0022	0.0062	0.0087
ALMAVE	**0.99689**	**0.00305**	**0.0043**	**0.0065**	**0.9988**	**0.0007**	**0.0057**	**0.0081**
SCADMAVE	**0.99561**	**0.00464**	**0.0044**	**0.0065**	**0.9982**	**0.0015**	**0.0060**	**0.0082**
MCPMAVE	0.99559	0.00489	0.0045	0.0066	0.9978	0.0019	0.0061	0.0085

Table 3. Simulation results for the studied methods based on the model in Example 3

	$\hat{\beta}_1$				$\hat{\beta}_2$			
	$\|r\|$ Mean	$\|r\|$ SD	MSE Mean	MSE SD	$\|r\|$ Mean	$\|r\|$ SD	MSE Mean	MSE SD
				Case (1)				
SMAVE	0.9760	0.0071	0.0062	0.0087	0.8603	0.1045	0.0095	0.0096
ALMAVE	**0.9938**	**0.0029**	**0.0061**	**0.0080**	**0.9641**	**0.0882**	**0.0045**	**0.0028**
SCADMAVE	0.9924	0.0062	0.0062	0.0082	0.8674	0.1075	0.0088	0.0076
MCPMAVE	**0.9934**	**0.0043**	**0.0062**	**0.0080**	**0.8729**	**0.0889**	**0.0086**	**0.0074**
				Case (2)				
SMAVE	0.9798	0.0078	0.0009	0.0012	0.6903	0.2194	0.0036	0.0066
ALMAVE	**0.9954**	**0.0024**	**0.0007**	**0.0008**	**0.8049**	**0.1202**	**0.0015**	**0.0032**
SCADMAVE	**0.9920**	**0.0055**	**0.0008**	**0.0010**	**0.7348**	**0.1213**	**0.0016**	**0.0038**
MCPMAVE	0.9918	0.0059	0.0009	0.0011	0.6972	0.2359	0.0039	0.0069
				Case (3)				
SMAVE	0.9159	0.0408	0.0192	0.0165	0.9313	0.0398	0.0239	0.0355
ALMAVE	**0.9409**	**0.0360**	**0.0179**	**0.0145**	**0.9545**	**0.0309**	**0.0231**	**0.0354**
SCADMAVE	**0.9189**	**0.0360**	**0.0190**	**0.0163**	**0.9352**	**0.0309**	**0.0238**	**0.0355**
MCPMAVE	0.9158	0.0365	0.0192	0.0163	0.9332	0.0337	0.0240	0.0356

Table 4. Simulation results for the studied methods based on the model in Example 4

	$\|r\|$ Mean	$\|r\|$ SD	MSE Mean	MSE SD
		$\boldsymbol{n} = \mathbf{60}$		
SMAVE	0.9959	0.0101	0.0564	0.0645
ALMAVE	0.9466	0.1895	0.0587	0.0653
SCADMAVE	**0.9958**	**0.0100**	**0.0563**	**0.0645**
MCPMAVE	**0.9958**	**0.0100**	**0.0563**	**0.0645**
		$\boldsymbol{n} = \mathbf{120}$		
SMAVE	0.9976	0.0008	0.0619	0.0557
ALMAVE	0.9976	0.0009	0.0619	0.0559
SCADMAVE	**0.9975**	**0.0008**	**0.0619**	**0.0557**
MCPMAVE	**0.9975**	**0.0008**	**0.0619**	**0.0557**

Table 5. The values of the adjusted R-squared for the model fit based on air pollution data

		SMAVE	ALMAVE	SCADMAVE	MCPMAVE
Model fit	Linear	0.76	0.93	0.76	0.76
	Quadratic	0.90	0.94	0.90	0.90
	Cubic	0.93	0.94	0.93	0.93
	Quartic	0.93	0.94	0.93	0.93

Table 6. Prediction error of the models which are selected by the studied methods based on air pollution data

Method	Prediction error
SMAVE	0.7768
ALMAVE	**0.6692**
SCADMAVE	**0.7740**
MCPMAVE	**0.7741**

Table 7. The values of the adjusted R-squared for the model fit based on body fat data

		SMAVE	ALMAVE	SCADMAVE	MCPMAVE
Model fit	Linear	0.92	0.92	0.92	0.92
	Quadratic	0.95	0.95	0.95	0.95
	Cubic	0.96	0.96	0.96	0.96
	Quartic	0.96	0.96	0.96	0.96

Table 8. Prediction error of the models selected by the studied methods based on body fat data

Method	Prediction error
SMAVE	24.4095
ALMAVE	**22.6263**
SCADMAVE	**23.5089**
MCPMAVE	**23.0635**

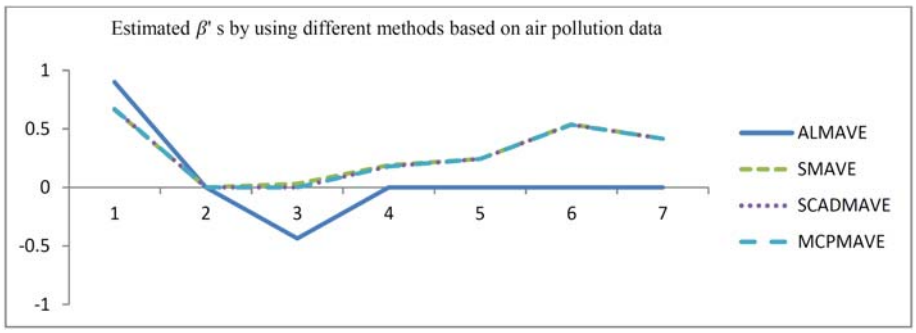

Figure 1. Plot and explanation of the estimated $\hat{\beta}$ which is estimated by studied methods based on air pollution data.

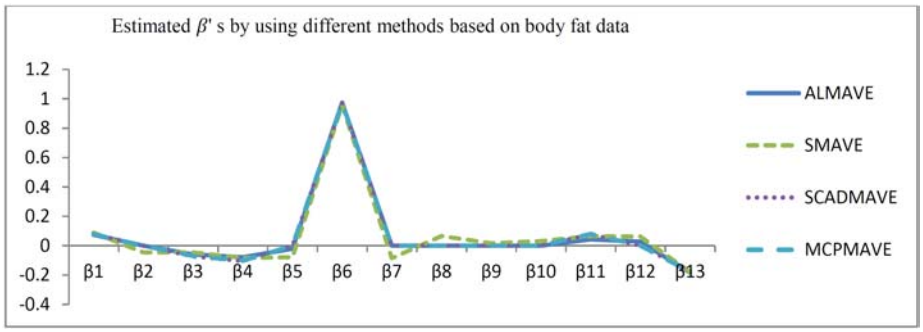

Figure 2. Plot and explanation of the estimated $\hat{\beta}$ which is estimated by studied methods based on body fat data.