Int. J. Nonlinear Anal. Appl. 13 (2022) No. 1, 2921–2931
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.6024

Sparse dimension reduction with group identification

Ali Alkenani[a,*]

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

(Communicated by Madjid Eshaghi Gordji)

Abstract

Estimating the central mean subspace without requiring to designate a model is achieved via MAVE method. The original p predictors are replaced with d-linear combinations (LC) of predictors in MAVE, where $d < p$ without loss of any information about the regression. However, it is known that the interpretation of the estimated effective dimension reduction (EDR) direction is not easy due to each EDR direction is a LC of all the original predictors. The PACS method is an oracle procedure. In this article, a group variable selection method (SMAVE-PACS) is proposed. The sufficient dimension reduction (SDR) concepts and group variable selection are emerged through SMAVE-PACS. SMAVE-PACS produces sparse and accurate solutions with the ability of group identification. SMAVE-PACS extended PACS to multi-dimensional regression under SDR conditions. In addition, a method for estimating the structural dimension was proposed. The effectiveness of the SMAVE-PACS is checked through simulation and real data.

Keywords: Variable selection, Group identification, MAVE, Pairwise absolute clustering and sparsity.

1. Introduction

Regression analysis can be highly challenging because the dimension p of predictor vector $\mathbf{x}$ is large. A useful tool to deal with high dimensional data is to shrinkage the dimensions of $\mathbf{x}$ without the loss of the regression information and without assuming a specific model.
Cook (1998)[2] introduced the sufficient dimension reduction (SDR) theory to reduce the dimensions of $\mathbf{x}$ while saving the regression information. For regression problems, Let y is a response and $\mathbf{x}=(\mathrm{x}_1,\ldots,\mathrm{x}_p)^T$ is a $p\times 1$ predictor vector. The SDR searches for a $p\times d$ matrix $\mathbf{B}$, such that $y \perp\!\!\!\perp \mathbf{x}|\mathbf{x}^T\mathbf{B}$, where $\perp\!\!\!\perp$ indicates independence. The column spanned by $\mathbf{B}$ is known as the dimension

*Corresponding author
Email address: ali.alkenani@qu.edu.iq - http://orcid.org/0000-0001-5067-2321 (Ali Alkenani)

Received: May 2021 *Accepted:* October 2021

reduction subspace (DRS). The intersection of all DRS is called the central subspace, which is denoted by $S_{y|\mathbf{x}}$. The $S_{y|\mathbf{x}}$ contains all the regression information of y, given $\mathbf{x}$ (Yu and Zhu, 2013)[21]. Many methods were proposed for finding $S_{y|\mathbf{x}}$. For example, SIR (Li, 1991)[7], SAVE (Cook and Weisberg, 1991)[4] and PHD (Li, 1992)[8].

The idea of the central mean subspace, which is denoted by $S_{E(y|\mathbf{x})}$, was introduced in Cook and Li (2002)[3] for SDR when the mean function is of interest. In order to estimate $S_{E(y|\mathbf{x})}$, many methods were proposed, such as the iterative Hessian transformation (Cook and Li, 2002) and MAVE (Xia et al., 2002)[20].

The merits of MAVE method can be summarised as follow. Firstly, in order to obtain a faster consistency rate for the estimated parameters, the nonparametric link function estimator must undersmooth in the majority of the existing methods. In contrast, for MAVE method there is no need to undersmooth the estimator of the function to obtain a faster consistency rate. The dimension of the space can be estimated consistently via MAVE because it achieves a faster consistency rate for the estimators of parameter. Secondly, MAVE is applicable to a vast range of models and it is easy to implement with fewer restrictions on the probabilistic structure of x.

SDR methods were proved as efficient methods; however, the problem of these methods is that each DR component is a linear combination of all of the predictors, which may not be simple to explain the resulting estimates.

The selection of predictors is important in constructing the regression model. In addition, the prediction accuracy can be improved through the selection of appropriate subset of predictors. Moreover, in practice, it is simple to interpret the model with a small number of predictors. The regularisation methods were used significantly in the classical least squares problems. For example, Lasso (Tibshirani, 1996)[17], SCAD (Fan and Li, 2001)[5], Elastic Net (Zou and Hastie, 2005)[25], adaptive Lasso (Zou, 2006)[24] and MCP (Zhang, 2010[23]).

Under the framework of SDR, many researchers proposed to combine the ideas of regularisation with SDR. For example, Li et al. (2005)[9], Ni et al. (2005)[14], Li and Nachtsheim (2006)[10], Li (2007)[6], Li and Yin (2008)[11] and so on. Wang and Yin (2008)[18] proposed to combine Lasso with MAVE to obtain SMAVE. Wang et. al. (2013)[22] proposed the penalised MAVE (P-MAVE) method. A penalty of bridge was employed to penalise l_1-norms of the basis matrix. Alkenani and Yu (2013)[1] suggested to combine MAVE with SCAD, adaptive Lasso and the MCP penalties. Wang et. al. (2015)[19] proposed to merge Lasso with group-wise MAVE.

The omission of insignificant predictors and combining the predictors of indistinguishable coefficients (ICs) are two important matters in the search for the correct model (Sharma et al., 2013)[15]. The above-mentioned regularisation penalties do well with removing unimportant predictors but fail in merging predictors with ICs. PACS (Sharma et al., 2013) can achieve both aims. Moreover, PACS was shown as oracle method (Sharma et al., 2013). In order to make the concept of " group identification" clear, we can cite the following sentences from Sharma et al. (2013) "if the coefficients of two predictors are truly equal in magnitude, we would combine these two columns of the design matrix by their sum and if a coefficient were truly zero, we would exclude the corresponding predictor".

In this article, MAVE-PACS method is proposed. The MAVE-PACS method has the ability to omit insignificant predictors and combine the predictors with ICs under the framework of the SDR. MAVE-PACS has advantages over the SMAVE, SPMAVE and P-MAVE methods. It benefits from the strength of PACS in deleting unimportant predictors and merging the predictors with ICs, which does not hold for the above mensioned penalties.

The rest is organised as follows. In Section 2, a short review of SDR and MAVE is mentioned. The MAVE-PACS method is proposed in Section 3. Simulations are carried out in Section 4. In Section 5, the methods are applied on real data. In Section 6, conclusions are given.

2. SDR and MAVE

In this section, SDR and MAVE are briefly reviewed. Assume the following model:

$$y = f\left(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_p\right) + \varepsilon, \tag{2.1}$$

where y is the response on a $p \times 1$ vector $\mathbf{x}$ of predictors and ε is the error., Also, $E\left(y|\mathbf{x}\right) = f\left(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_p\right)$ and $E\left(\varepsilon\ |\mathbf{x}\right) = 0$. SDR for mean function investigates a subspace S such that

$$y \ \perp\!\!\!\perp E\left(y|\mathbf{x}\right)|P_s\mathbf{x}, \tag{2.2}$$

where $P_{(.)}$ refers to an operator of projection. Cook and Li (2002) stated that the subspaces satisfying (2.2) are the mean DR subspaces. If $d = dim\left(S\right)$ and $\mathbf{B} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_d\right)$ is a basis for S, $\mathbf{x}$ will replace with ${\boldsymbol{\beta}_1}^T\mathbf{x}, {\boldsymbol{\beta}_2}^T\mathbf{x}, \ldots, {\boldsymbol{\beta}_d}^T\mathbf{x}$, $d \leq p$ without loss any information on $E\left(y|\mathbf{x}\right)$. The central mean subspace, which is denoted by $S_{E(y|\mathbf{x})}$, is the intersection of all subspaces satisfying (2.2) (Cook and Li, 2002). One of the most popular methods for estimating $S_{E(y|\mathbf{x})}$ is the MAVE.
The MAVE method was proposed in (Xia et al., 2002). The matrix $\mathbf{B}$ is obtained by

$$\min_{\mathbf{B}} \left\{\mathrm{E}\left[y - \mathrm{E}\left(y|\mathbf{B}^T\mathbf{x}\right)\right]^2\right\}, \tag{2.3}$$

where $\mathbf{B}^T\mathbf{B}=\mathbf{I}_d$ and

$$\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{B}}\left(\mathbf{B}^T\mathbf{x}\right) = \mathrm{E}\left[\left\{y - \mathrm{E}\left(y|\mathbf{B}^T\mathbf{x}\right)\right\}^2 |\mathbf{B}^T\mathbf{x}\right]. \tag{2.4}$$

Thus,

$$\min_{\mathbf{B}}\ \mathrm{E}\left[y - \mathrm{E}\left(y|\mathbf{B}^T\mathbf{x}\right)\right]^2 = \min_{\mathbf{B}}\ \mathrm{E}\left\{\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{B}}\left(\mathbf{B}^T\mathbf{x}\right)\right\}. \tag{2.5}$$

For any given $\mathbf{x}_0$, $\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{B}}\left(\mathbf{B}^T\mathbf{x}\right)$ can be locally approximated as follows:

$$\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{B}}\left(\mathbf{B}^T\mathbf{x}_0\right) \approx \sum_{i=1}^{n}\left\{y_i - \mathrm{E}\left(y_i|\mathbf{x}_i^T\mathbf{B}\right)\right\}^2\omega_{i0}$$

$$\approx \sum_{i=1}^{n}\left[y_i - \left\{a_0 + \boldsymbol{b}_0^T\mathbf{B}^T\left(\mathbf{x}_i - \mathbf{x}_0\right)\right\}\right]^2\omega_{i0},$$

where $\omega_{i0} \geq 0$ are the kernel weights with $\sum_{i=1}^{n}\omega_{i0} = 1$. So, $\mathbf{B}$ can be found by solving the following minimisation:

$$\min_{\mathbf{B}:\mathbf{B}^T\mathbf{B}=\mathbf{I}_m}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - \left\{a_j + \boldsymbol{b}_j^T\mathbf{B}^T\left(\mathbf{x}_i - \mathbf{x}_\mathrm{j}\right)\right\}\right]^2\omega_{ij}\right). \tag{2.6}$$

3. The SMAVE with PACS penalty (SMAVE-PACS)

The SMAVE method was proposed by Wang and Yin (2008)[18] through incorporating l_1 penalty into (2.6). It minimises:

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - \left\{a_j + \boldsymbol{b}_j^T\mathbf{B}^T\left(\mathbf{x}_i - \mathbf{x}_\mathrm{j}\right)\right\}\right]^2\omega_{ij} + \lambda\sum_{k=1}^{p}\left|\boldsymbol{\beta}_{m,k}\right|, \tag{3.1}$$

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

The preceding authors assumed that d is known and then they proposed to estimate it according to BIC. Alkenani and Yu (2013)[1] proposed another version of sparse MAVE by combining the adaptive Lasso, SCAD and MCP penalties with MAVE in (2.6).

Choosing the true model needs the omission of unimportant predictors and merging the predictors of ICs. SMAVE and SPMAVE employed penalties that failed to combine the predictors with ICs. The PACS penalty able to achieve the two aims.

In this study, sparse MAVE with PACS penalty (SMAVE-PACS) is proposed to minimise

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i-\left\{a_j+\boldsymbol{b}_j^T\mathbf{B}^T\left(\mathbf{x}_i-\mathbf{x}_j\right)\right\}\right]^2\omega_{ij} + \lambda\left\{\sum_{j=1}^{p}\omega_j\left|\boldsymbol{\beta}_{m,j}\right|+\sum_{1\le j<k\le p}\omega_{jk(-)}\left|\boldsymbol{\beta}_{m,k}-\boldsymbol{\beta}_{m,j}\right|+\sum_{1\le j<k\le p}\omega_{jk(+)}\left|\boldsymbol{\beta}_{m,k}+\boldsymbol{\beta}_{m,j}\right|\right\}, \tag{3.2}$$

The penalty in (3.2) consists of $\lambda\left\{\sum_{j=1}^{p}\omega_j\left|\boldsymbol{\beta}_{m,j}\right|\right\}$ that encourages sparseness, $\lambda\left\{\sum_{1\le j<k\le p}\omega_{jk(-)}\left|\boldsymbol{\beta}_{m,k}-\boldsymbol{\beta}_{m,j}\right|\right\}$ that enables the coefficients with same sign to set as equal and $\lambda\left\{\sum_{1\le j<k\le p}\omega_{jk(+)}\left|\boldsymbol{\beta}_{m,k}+\boldsymbol{\beta}_{m,j}\right|\right\}$ that enables the coefficients with different sign to set in magnitude as equal.

The choice of suitable adaptive weights plays a crucial role for PACS to be an oracle procedure. As a result, Sharma et al. (2013) proposed adaptive PACS that incorporates into the weights correlations as follows:

$$\omega_j=\left|\widetilde{\beta}_j\right|^{-1},\omega_{jk(-)}=\left(1-r_{jk}\right)^{-1}\left|\widetilde{\beta}_k-\widetilde{\beta}_j\right|^{-1}\text{and}\quad\omega_{jk(+)}=\left(1+r_{jk}\right)^{-1}\left|\widetilde{\beta}_k+\widetilde{\beta}_j\right|^{-1}\text{for}\quad 1\le j<k\le p, \tag{3.3}$$

where $\widetilde{\beta}$ is a $\sqrt{n}$ consistent estimator of β and r_{jk} is Pearson's correlation.

3.1. SMAVE-PACS

The algorithm of SMAVE-PACS is as below:

1. Let $m = 1$, and $\mathbf{B}=\boldsymbol{\beta}_0$, any arbitrary $p \times 1$ vector.
2. when $\mathbf{B}$ is known, get $(a_j, \boldsymbol{b}_j)$ where $j = 1, \ldots, n$, by solving

$$\min_{a_j,\boldsymbol{b}_j,j=1,\ldots,n}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i-\left\{a_j+\boldsymbol{b}_j^T\mathbf{B}^T\left(\mathbf{x}_i-\mathbf{x}_j\right)\right\}\right]^2\omega_{ij}\right). \tag{3.4}$$

3. when $\left(\hat{a}_j,\widehat{\boldsymbol{b}}_j\right)$ are given, $j = 1, \ldots, n$, solve $\boldsymbol{\beta}_{m\text{SMAVE–PACS}}$ from the following:

$$\min_{\mathbf{B}:\mathbf{B}^T\mathbf{B}=\mathbf{I}_m}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i-\left\{\hat{a}_j+\widehat{\boldsymbol{b}}_j^{\boldsymbol{T}}\left(\widehat{\boldsymbol{\beta}}_1,\widehat{\boldsymbol{\beta}}_2,\ldots,\widehat{\boldsymbol{\beta}}_{m-1},\boldsymbol{\beta}_m\right)^T\left(\mathbf{x}_i-\mathbf{x}_j\right)\right\}\right]^2\omega_{ij} + \lambda\left\{\begin{array}{c}\sum_{j=1}^{p}\omega_j\left|\boldsymbol{\beta}_{m,j}\right|+\sum_{1\le j<k\le p}\omega_{jk(-)}\left|\boldsymbol{\beta}_{m,k}-\boldsymbol{\beta}_{m,j}\right|\\+\sum_{1\le j<k\le p}\omega_{jk(+)}\left|\boldsymbol{\beta}_{m,k}+\boldsymbol{\beta}_{m,j}\right|\end{array}\right\}\right) \tag{3.5}$$

4. Replace the mth column of $\mathbf{B}$ by $\widehat{\boldsymbol{\beta}}_{m\text{SMAVE–PACS}}$ and repeat till convergence the steps 2 and 3.
5. Update $\mathbf{B}$ by $\left(\widehat{\boldsymbol{\beta}}_{1\text{SMAVE–PACS}},\widehat{\boldsymbol{\beta}}_{2\text{SMAVE–PACS}},\ldots,\widehat{\boldsymbol{\beta}}_{m\text{SMAVE–PACS}},\boldsymbol{\beta}_0\right)$, and set m to be $m + 1$.
6. If $m < d$, continue steps 2 to 5 until $m = d$.

We employed Gaussian kernel which was suggested by Xia et al. (2002) in computing the weights:

$$\omega_{ij} = K_h\left\{\widehat{\mathbf{B}}^T(\mathbf{x}_i - \mathbf{x}_j)\right\} / \sum_{i=1}^{n} K_h\left\{\widehat{\mathbf{B}}^T(\mathbf{x}_i - \mathbf{x}_j)\right\},$$

where K is Gaussian product kernel function and h is the bandwidth of the weights ω_{ij}.
The minimisation method of (3.5) contains of two parts. The first is the loss function of MAVE. The second is PACS penalty, which consists of $\lambda\left\{\sum_{j=1}^{p}\omega_j|\beta_{m,j}|\right\}$, $\lambda\left\{\sum_{1\le j<k\le p}\omega_{jk(-)}|\beta_{m,k}-\beta_{m,j}|\right\}$ and $\lambda\left\{\sum_{1\le j<k\le p}\omega_{jk(+)}|\beta_{m,k}+\beta_{m,j}|\right\}$.
The minimisation in (3.5) can be obtained using PACS algorithms. The $\boldsymbol{\beta}_{m\text{SMAVE-PACS}}$ is the PACS estimator for the regression of $\mathbf{Y}_{\boldsymbol{s}}$ on the data matrix $\mathbf{V}_{\boldsymbol{s}}$, where $\mathbf{Y}_{\boldsymbol{s}}$ and $\mathbf{V}_{\boldsymbol{s}}$ are as described in step 2 of the "OLS formulation" of MAVE in Section 2.2 of Wang and Yin (2008).
In summary, SMAVE-PACS is a procedure of two-steps: first, use MAVE to get the dimension d, $\mathbf{Y}_{\boldsymbol{s}}$ and $\mathbf{V}_{\boldsymbol{s}}$; secondly, compute $\boldsymbol{\beta}_{m\text{SMAVE-PACS}}$ via the PACS approach.
The predictors are standardised in the simulation studies and real data analysis. The Gaussian product kernel and $h_{opt} = A(d)\,n^{-1/(4+d)}$, which are reported in (Silverman, 1986)[16], are used, where $?(d) = \left\{\frac{4}{(d+2)}\right\}^{1/(4+d)}$.
SMAVE-PACS combines PACS into "OLS formulation" of MAVE. Under the settings of MAVE and PACS, the algorithm of SMAVE-PACS converges to the global minimum. Our simulation experiments indicated that the proposed algorithm converges within five to ten iterations. The PACS algorithm is an efficient with order of computation is similar to a single OLS fit (Sharma et al., 2013)[15]. The penalty of SMAVE-PACS appears in the "OLS formulation" of MAVE. Consequently, SMAVE-PACS is as efficient as MAVE. Under some conditions, the PACS is $\sqrt{n}$-consistent (Sharma et al., 2013), while the consistency rate for the estimator of MAVE is $O\left(h_{opt}^3\log(n)\right)$(Xia et al., 2002)[20]. Because of that the estimator of MAVE has a lower consistency rate than the consistency rate of PACS, the consistency rate for SMAVE-PACS is controlled by that of MAVE. Under the conditions of Sharma et al. (2013) and Xia et al. (2002), it can be shown that the estimator of SMAVE-PACS has consistency rate similar to that of MAVE estimator.
As a simple illustration, an example was implemented to show the necessity of SMAVE-PACS. Let $y = 5\cos(\mathbf{x}^T\boldsymbol{\beta}) + \exp\left(-(\mathbf{x}^T\boldsymbol{\beta})^2\right) + \varepsilon$,
where $\beta = (2,2,2,1,1,0,0,0,0,0)^T$ and $X \in \mathbb{R}^{10}$. Also, x_i and ε are i.i.d from $N(0,1)$, with $S_{E(y|\mathbf{x})} = span(\mathbf{B}_1)$. The first three predictors and the second two are correlated with correlation is 0.7 and the values of coefficients are equal in magnitude. The rest are uncorrelated. A single but representative simulated data set with size $n = 120$ was generated, and the direction estimates were:
MAVE = (1.273,1.571, 1.353,0.673, 0.554,0.048, 0.181,0.042, 0.114,0.022)
SMAVE-PACS = (1.858,1.858, 1.858,0.879, 0.879,0, 0.034,0, 0.042,0)
SMAVE = (1.359,1.619, 1.389,0.789, 0.503,0, 0.121,0.056, 0.089,0)
ALMAVE = (1.649,1.759, 1.689,0.827, 0.764,0, 0,0, 0.065,0)
SCAD-MAVE = (1.479,1.659, 1.449,0.799, 0.680,0, 0.091,0, 0.069,0)
MCP-MAVE = (1.455,1.640, 1.429,0.788, 0.669,0, 0.097,0, 0.071,0)
Note that none of MAVE, SMAVE, ALMAVE, SCAD-MAVE and MCP-MAVE perform grouping and the overall estimation accuracy is poor. An important gain in selection accuracy and grouping accuracy has achieved by SMAVE-PACS.

3.2. Estimation of the dimension d

Xia et al. (2002)[20] extended the cross-validation method of Yao and Tong (1994) to propose method for determining the dimension d. The authors estimated d by $\hat{d}_1 = \{arg\min_{0\le k\le p} CV_k\}$, where

$$CV_k = n^{-1}\sum_{i=1}^{n}\left(y_i - \frac{\sum_{j\neq i} y_j K_h\left\{\widehat{\mathbf{B}}^T(\mathbf{x}_l - \mathbf{x}_i)\right\}}{\sum_{l\neq i} K_h\left\{\widehat{\mathbf{B}}^T(\mathbf{x}_j - \mathbf{x}_i)\right\}}\right)^2, \tag{3.6}$$

where k is the estimate of the dimension.
Wang and Yin (2008) modified the BIC criterion to estimate d. The authors estimated d by $\hat{d}_2 = min\{\aleph : \aleph = arg\min_{0\le k\le p}(BIC_k)\}$, where

$$BIC_k = log\left(\frac{RSS_k}{n}\right) + \frac{\log(n)\ k}{nh^k} \tag{3.7}$$

where RSS_k is as follows:

$$RSS_k = \sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - \left\{\hat{a}_j + \widehat{\boldsymbol{b}}_{\boldsymbol{j}}^{\boldsymbol{T}}\left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \ldots, \boldsymbol{\beta}_k\right)^T(\mathbf{x}_i - \mathbf{x}_j)\right\}\right]^2 \omega_{ij}, \tag{3.8}$$

In this section, we combined (3.6) and (3.7) in one formula to estimate d as follows:

$$\hat{d} = min\left\{\hat{d}_1, \hat{d}_2\right\} \tag{3.9}$$

4. Simulation study

In this section, the behaviour of the SMAVE-PACS is demonstrated through the use of many simulation examples. The methods were assessed in terms of model error (ME) which is $(\widehat{\beta} - \beta)' V(\widehat{\beta} - \beta)$, where V is the covariance matrix of X. We report the median and standard error of ME. In addition, we computed and reported the SA which is the percentage of correct models identified, the GA which is the percentage of correct groups identified and the percentage of selection and grouping accuracy together (SGA). None of the ALMAVE, SCAD-MAVE, MCP-MAVE, SMAVE and sparse SIR perform grouping. The optimal λ in the PACS estimation can be selected by using the tenfold Cross-validation. The sample sizes were 60 and 120 and the simulation was replicated 200 times.

4.1. The estimation of directions and variable selection

Example 4.1. *$\mathcal{R} = 200$ datasets with size $n = 60$ and 120 were generated from $y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$, x_i and ε are independent and are identically distributed (i.i.d) from an $N(0,1)$ and $\beta = (2,2,2,0,0,0,0,0)^T$, $X \in \mathbb{R}^8$ with $d = 1$. The first three predictors have correlation equal to 0.7 and the values of their coefficients in magnitude were equal. The rest are uncorrelated. The model is $y = 2\mathrm{x}_1 + 2\mathrm{x}_2 + 2\mathrm{x}_3 + \varepsilon$.*

Example 4.2. *$\mathcal{R} = 200$ datasets with size $n = 60$ and 120 were generated from $y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$, x_i and ε are i.i.d from an $N(0,1)$ and $\beta = (0.5,1,2,0,0,0,0,0)^T$, $X \in \mathbb{R}^8$ with $d = 1$. The first three predictors have correlation equal to 0.7 and the values of their coefficients are different. The rest are uncorrelated. The model is $y = 0.5\mathrm{x}_1 + 1\mathrm{x}_2 + 2\mathrm{x}_3 + 0.5\ \varepsilon$.*

Example 4.3. $\mathcal{R} = 200$ *datasets with size* $n = 60$ *and* 120 *were generated from* $y = exp(\mathbf{x}^T\boldsymbol{\beta}) + \varepsilon$, *where* $\beta = (1,1,1,0.5,1,2,0,0,0,0)^T$, $X \in \mathbb{R}^{10}$ *with* $d = 1$, x_i *and* ε *are i.i.d from an* $N(0,1)$. *The first three predictors have correlation equal to 0.3 and the values of their coefficients are equal. The pairwise correlation of the second three is 0.7 with different magnitudes for the coefficients. The rest are uncorrelated. The model is* y=$\exp(\mathrm{x}_1+\mathrm{x}_2+\mathrm{x}_3+0.5\mathrm{x}_4+\mathrm{x}_5+2\mathrm{x}_6) + \varepsilon$.

Example 4.4. $\mathcal{R} = 200$ *datasets with size* $n = 60$ *and* 120 *were generated from* $y = 5\cos(\mathbf{x}^T\boldsymbol{\beta}) + \exp\left(-(\mathbf{x}^T\boldsymbol{\beta})^2\right) + \varepsilon$, *where* $\beta = (2,2,2,1,1,0,0,0,0,0)^T$, $X \in \mathbb{R}^{10}$. x_i *and* ε *are i.i.d from* $N(0,1)$, *with with* $d = 1$. *The first 3 and the second 2 predictors have correlation equal to 0.7 and the values of their coefficients in magnitude are equal. The rest are uncorrelated.*

Example 4.5. $\mathcal{R} = 200$ *datasets with size* $n = 60$ *and* 120 *were generated from the model* $y=\frac{\mathbf{x}^T\boldsymbol{\beta}_1}{0.5+(1.5+\mathbf{x}^T\boldsymbol{\beta}_2)}+\varepsilon$, *where* x_i *and* ε *are i.i.d from an* $N(0,1)$. *Also,* $\boldsymbol{\beta}_1 = (2,2,2,0,0,0,0,0)^T$ *and* $\boldsymbol{\beta}_2 = (0,0,0,0,0,2,2,2)^T$. $X \in \mathbb{R}^8$ *with* $d = 2$. *For* $\boldsymbol{\beta}_1$, *the first 3 predictors have correlation equal to 0.7 and the values of their coefficients are equal in magnitude. The rest are uncorrelated. For* $\boldsymbol{\beta}_2$, *the first 5 predictors are uncorrelated, while last 3 predictors have correlation equal to 0.7 and the values of their coefficients are equal in magnitude.*

Table 1: Results of Example 4.1

n	Criterion	Sparse SIR	SMAVE	ALMAVE	SCAD-MAVE	MCP-MAVE	SMAVE-PACS
60	ME (s.e)	0.1170 (0.0083)	0.1062 (0.0079)	0.0697 (0.0107)	0.0727 (0.0079)	0.0811 (0.0077)	0.0447 (0.0108)
	SA	61	61	79	79	76	64
	GA	0	0	0	0	0	82
	SGA	0	0	0	0	0	59
120	ME (s.e)	0.0531 (0.0046)	0.0423 (0.0055)	0.0288 (0.0033)	0.0302 (0.0034)	0.0331 (0.0042)	0.0040 (0.0014)
	SA	69	70	91	92	86	84
	GA	0	0	0	0	0	92
	SGA	0	0	0	0	0	79

From Table, the ME of SMAVE-PACS is the lowest for all sample sizes. Although the ALMAVE, SCAD-MAVE and MCP-MAVE have the highest SA, it is obvious that they do not perform grouping. It is clear that SMAVE-PACS identifies the groups of predictors, as seen in GA and SGA.

Table 2: Results of Example 4.2

n	Criterion	Sparse SIR	SMAVE	ALMAVE	SCAD-MAVE	MCP-MAVE	SMAVE-PACS
60	ME (s.e)	0.1677 (0.0079)	0.1087 (0.0085)	0.0828 (0.0144))	0.0901 (0.0117)	0.1105 (0.0122)	0.1341 (0.0080)
	SA	59	60	71	62	60	55
	NG	100	100	100	100	100	98
	SNG	58	59	70	60	59	48
120	ME (s.e)	0.0743 (0.0070)	0.0409 (0.0051)	0.0314 (0.0052)	0.0302 (0.0020)	0.0383 (0.0048)	0.0531 (0.0028)
	SA	68	70	87	84	72	67
	NG	100	100	100	100	100	100
	SNG	67	69	86	84	69	69

In Table 2, we reported NG which is refer to no groups found and the percentage of selection and no-grouping (SNG) instead of GA and SGA, respectively. In terms of prediction and selection, it can be seen that the SMAVE-PACS does not work as well, while ALMAVE, SCAD-MAVE and MCP-MAVE do the best. The methods do well in not identifying the group. Thus, the SMAVE-PACS is not an advisable when the correlations are high but the important coefficients not in a group.

Table 3: Results of Example 4.3

n	Criterion	Sparse SIR	SMAVE	ALMAVE	SCAD-MAVE	MCP-MAVE	SMAVE-PACS
60	ME (s.e)	0.1832 (0.0123)	0.1680 (0.0178)	0.1393 (0.0143)	0.1443 (0.0114)	0.1550 (0.0093)	0.1389 (0.0196)
	SA	47	47	79	75	70	59
	GA	0	0	0	0	0	59
	SGA	0	0	0	0	0	39
120	ME (s.e)	0.0778 (0.0060)	0.0692 (0.0051)	0.0430 (0.0031)	0.0441 (0.0030)	0.0456 (0.0035)	0.0350 (0.0056)
	SA	42	43	86	86	83	75
	GA	0	0	0	0	0	81
	SGA	0	0	0	0	0	61

Table 3 displays that the best SA are for ALMAVE, SCAD-MAVE and MCP-MAVE; however, the SMAVE-PACS do better according to ME. In this setting, it is clear that the SMAVE-PACS method identifies the significant group with high GA and SGA.

Table 4: Results of Example 4.4

n	Criterion	Sparse (SIR)	SMAVE	ALMAVE	SCAD-MAVE	MCP-MAVE	SMAVE-PACS
60	ME (s.e)	0.1933 (0.0126)	0.1750 (0.0170)	0.1695 (0.0128))	0.1698 (0.0161)	0.1705 (0.0105))	0.1700 (0.0149)
	SA	52	52	71	69	60	59
	GA	0	0	0	0	0	50
	SGA	0	0	0	0	0	39
120	ME (s.e)	0.0802 (0.0063)	0.0630 (0.0056)	0.0496 (0.0037)	0.0527 (0.0050)	0.0559 (0.0042)	0.0555e (0.0076)
	SA	57	58	83	81	76	74
	GA	0	0	0	0	0	68
	SGA	0	0	0	0	0	50

Table 4 demonstrates that the ALMAVE, SCAD-MAVE and MCP-MAVE have the best SA. In terms of ME, it is clear that the ALMAVE and SCAD-MAVE have better results than the SMAVE-PACS method. According to GA and SGA, it is clear that SMAVE-PACS does well in identifying the groups.

Table 5: Results of Example 4.5

		$\boldsymbol{\beta}_1$						$\boldsymbol{\beta}_2$					
n	Criterion	Sparse (SIR)	SMAVE	ALMAVE	SCAD-MAVE	MCP-MAVE	SMAVE-PACS	Sparse (SIR)	SMAVE	ALMAVE	SCAD-MAVE	MCP-MAVE	SMAVE-PACS
60	ME (s.e)	0.1664 (0.0107)	0.1518 (0.0127)	0.1308 (0.0120)	0.1325 (0.0122)	0.1370 (0.0093)	0.1186 (0.0119)	0.1680 (0.0107)	0.1551e (0.0127)	0.1318 (0.0120)	0.1343 (0.0122)	0.1392 (0.0093)	0.1213 (0.0131)
	SA	56	56	75	74	68	62	55	55	74	74	67	61
	GA	0	0	0	0	0	66	0	0	0	0	0	66
	SGA	0	0	0	0	0	59	0	0	0	0	0	50
120	ME (s.e)	0.0770 (0.0057)	0.0629 (0.0058)	0.0482 (0.0037)	0.0517 (0.0044)	0.0547 (0.0044)	0.0400 (0.0038)	0.0816e (0.0063)	0.0674 (0.0063)	0.0520 (0.0041)	0.0553 (0.0042)	0.0583 (0.0042)	0.0440 (0.0041)
	SA	63	64	87	87	81	79	62	64	86	86	81	79
	GA	0	0	0	0	0	80	0	0	0	0	0	79
	SGA	0	0	0	0	0	64	0	0	0	0	0	64

In Table 5, the dimension d = 2, it can be seen that the performance of SMAVE-PACS is not affected by d. In general, the performance of SMAVE-PACS was stable and did not change with the changing of d from 1 to 2. In terms of SA, the ALMAVE, SCAD-MAVE and MCP-MAVE have the largest SA values, respectively. In terms of ME, the results of ALMAVE and SCAD-MAVE are still better than the results of SMAVE-PACS. According to GA and SGA the performance of SMAVE-PACS is the best.

4.2. Estimation of the dimension d

The proposed method in (3.9) for estimating d was evaluated in this section. Data were generated according to the settings of example 5. The d value is 2. The results were reported for the sample sizes n = 100 and 200. We used 200 datasets for each case. The frequency of estimated d out of 200 datasets was summarised in Table 6. The results of the proposed method were compared with the results from CV in (3.6) and BIC in (3.7). It is clear that proposed formula in (3.9) produced highly consistent estimation for n = 100 and 200. It slightly outperforms the CV in (3.6) and BIC in (3.7).

Table 6: Frequency of $\hat{d}$ out of 200 datasets

	CV				BIC				$Min(CV,BIC)$			
n	$d=1$	$d=2$	$d=3$	$d\geq 4$	$d=1$	$d=2$	$d=3$	$d\geq 4$	$d=1$	$d=2$	$d=3$	$d\geq 4$
100	8	**157**	31	4	7	**158**	30	5	7	**160**	32	1
200	1	**180**	17	2	2	**183**	12	3	3	**184**	13	0

5. Analysis of real data

The performance of SMAVE-PACS with a number of existing selection methods is illustrated through real data. The NCAA data (Mangold et al., 2003) were analysed. The ALMAVE, SCAD-MAVE, MCP-MAVE (Alkenani and Yu, 2013), SMAVE (Wang and Yin, 2008), sparse SIR (Ni et al., 2005) and SMAVE-PACS methods were applied to the data. The response variable is centred and the predictors were standardised.

The data set was randomly split into a training and testing set, with 20% of the data used for testing. For stable comparisons, the data sets were split 100 times. SMAVE-PACS was applied to find $\hat{d}$. We find $\hat{d}$ = 1. The mean squared prediction error (MSPE) and the effective model size were reported. The effects of sociodemographic indicators and the sports programs were studied through the NCAA sport data. The data are available at the website:

(http://www4.stat.ncsu.edu/~boos/var.select/ncaa.html).

The data size is n = 94 and p = 19 predictors. The dependent variable is the average of a six-year graduation. The predictors are students in top 10% HS (x_1), COMPOSITE (x_2), living on campus (x_3), undergraduates (x_4), enrolment/1000 (x_5), courses taught (x_6), basketball ranking (x_7), tuition/1000 (x_8), room and board/1000 (x_9), avg. BB attendance (x_{10}), salary of prof. (x_{11}), student/faculty (x_{12}), white (x_{13}), salary of assist. Prof. (x_{14}), city popul.(x_{15}), PhD (x_{16}), accept. rate (x_{17}), receiving loans (x_{18}) and out of state (x_{19}).

Table 7: Results of NCAA sports data

Criteria	MAVE	Sparse (SIR)	SMAVE	ALMAVE	SCAD-MAVE	MCP-MAVE	SMAVE-PACS
Model Size	19	12	12	10	10	11	9
MSPE	0.65	0.63	0.62	0.55	0.58	0.58	0.53

Table 7 makes it obvious that SMAVE-PACS method is significantly better than the considered approaches in terms of MSPE. The effective model size is 9 for SMAVE-PACS.

6. Conclusion

In this study, SMAVE-PACS is developed to incorporate PACS into MAVE. Since MAVE can efficiently estimate $S_{E(y|\mathbf{x})}$ while PACS does consistent group identification and variable selection, SMAVE-PACS can simultaneously achieve the two aims. PACS is extended to multi-dimensional regression without needing any specific model through SMAVE-PACS. SMAVE-PACS is proved to have an effective computational algorithm. According to the results of our simulations, the proposed criterion for estimating the dimensionality d significantly improves MAVE and SMAVE in correctly estimating the dimensionality over using CV (Xia et al., 2002) and modified BIC (Wang and Yin, 2008), respectively. This work shows that SMAVE-PACS can yield promising predictive precision, as well as identify related groups.

References

[1] A. Alkenani and K. Yu, *Sparse MAVE with oracle penalties*, Adv. Appl. Stat., 34(2013) 85-105.
[2] R. Cook, *Regression graphics: ideas for studying the regression through graphics*, New York, Wily, 1998.
[3] R. D. Cook and B. Li *Dimension reduction for the conditional mean in regression*, Ann. Stat., 30(2002) 455-474.
[4] R. D. Cook, and S. Weisberg, *Discussion of Li* , J. Am. Stat Assoc., 86 (1991) 328-332.
[5] J. Fan and R. Z.Li, *Variable selection via non-concave penalized likelihood and its oracle properties*, J. Am. Stat. Assoc., 96 (2001) 1348-1360.
[6] L. Li, *Sparse sufficient dimension reduction*, Biometrika, 94 (2007) 603-613.
[7] K. Li, *Sliced inverse regression for dimension reduction (with discussion)*, J. Am. Stat. Assoc., 86(1991) 316-342.
[8] K. C.Li, *On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma*, J. Am. Stat. Assoc., 87(1992) 1025-1039.
[9] L. Li, R. D.Cook and C. J.Nachtsheim, *Model-free variable selection*, J. R. Stat. Soc. Ser. B, 67(2005) 285-299.
[10] L.Li and C. J. Nachtsheim, *Sparse sliced inverse regression*, Technometrics , 48(2006) 503-510.
[11] L. Li and X. Yin, *Sliced Inverse Regression with regularizations*, Biometrics, 64(2008) 124-131.
[12] W. D. Mangold, L. Bean, and D. Adams, *The impact of intercollegiate athletics on graduation rates among major NCAA Division I universities: Implications for college persistence theory and practice*, J. Higher Educ, 74(5)(2003) 540-562.
[13] G. C. McDonald and R. C. Schwing, *Instabilities of regression estimates relating air pollution to mortality*, Technometrics, 15(3) (1973) 463-481.
[14] L. Ni, R. D. Cook and C. L. Tsai, *A note on shrinkage sliced inverse regression* ,Biometrika , 92(2005) 242-247.
[15] D. B. Sharma, H. D. Bondell and H. H.Zhang, *Consistent group identification and variable selection in regression with correlated predictors*, J. Comput. Graphical Stat., 22(2)(2013) 319-340.
[16] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.
[17] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Royal Stat. Soc. Ser. B, 58 (1996) 267-288.
[18] Q. Wang and X. Yin, *A Nonlinear Multi-Dimensional Variable Selection Method for High Dimensional Data: Sparse MAVE*, Comput. Stat. Data Anal. , 52(2008) 4512-4520.
[19] T. Wang, P. Xu and L. Zhu, *Variable selection and estimation for semiparametric multiple-index models*, Bernoulli, 21(1) (2015) 242-275.
[20] 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.
[21] Z. Yu and L. Zhu, *Dimension reduction and predictor selection in semiparametric models*, Biometrika, 100 (2013) 641-654.
[22] T. Wang, P. Xu and L. Zhu, *Penalized minimum average variance estimation*, Statist. Sinica , 23(2013) 543-569.
[23] C. H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, Annal. Stat., 38 (2010) 894-942.
[24] H. Zou, *The adaptive Lasso and its oracle properties*. J. Am. Stat. Assoc., 101(2006) 1418-142.
[25] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. Royal Stat. Soc., Ser. B ,67(2005) 301-320.