**Int. J. Nonlinear Anal. Appl. 13 (2022) 1, 1167-1173**
**ISSN: 2008-6822 (electronic)**
http://dx.doi.org/10.22075/ijnaa.2022.5660

# Sparse minimum average variance estimation through signal extraction approach to multivariate regression

Abdulqader Ahmed[a,*], Saja Mohammad[a]

[a] *Department of Statistics, College of Administration and Economics, University of Baghdad, Baghdad, Iraq*

*(Communicated by Madjid Eshaghi Gordji)*

## Abstract

In this paper, a new sparse method called (MAVE-SiER) is proposed, to introduce MAVE-SiER, we combined the effective sufficient dimension reduction method MAVE with the sparse method Signal extraction approach to multivariate regression (SiER). MAVE-SiER has the benefit of expanding the Signal extraction method to multivariate regression (SiER) to nonlinear and multi-dimensional regression. MAVE-SiER also allows MAVE to deal with problems which the predictors are highly correlated. MAVE-SiER may estimate dimensions exhaustively while concurrently choosing useful variables. Simulation studies confirmed MAVE-SiER performance.

*Keywords:* High dimensional predictors, Dimension reduction, sparse, Minimum average variance estimation, Signal extraction approach to multivariate regression.

## 1. Introduction

High-dimensional data analysis has gained popularity in recent decades as a result of the huge data explosion. Many scientific areas, it has received a great deal of interest. In the literature, there are several model-based variable selection techniques. Finding efficient methods for processing high-dimensional data sets is one of the most challenging issues in current statistics.
In the context of multiple-linear regression, there is a massive of research on sparse, that is, for Y $=$ E(Y|X) $+ \epsilon = \beta^T$X $+ \epsilon$, where $\epsilon$ is i.i.d. N(0, $\sigma^2$).
To deal with the instability, new approaches such as Nonnegative Garrote [6], LASSO [11], SCAD

[*] Corresponding author
*Email addresses:* abdulqaderahmed79@gmail.com (Abdulqader Ahmed), zina.k.alabacy@uotechnology.edu.iq (Saja Mohammad)

*Received:* March 2021 *Accepted:* May 2021

[4], LARS [3] and Elastic Net [16], have been proposed. Through continuous penalty and automated variable selection, these approaches allow us to increase both model interpretability and prediction accuracy.
SIR [10], combined L1 penalty with a forward regression dimension reduction technique to come up with MAVE, MAVE (minimum average variance estimation, [14], MAVE is widely applied in other areas such as time series, economics and bioinformatics, and proposed sparse MAVE to select informative covariates. When compared to earlier works, Sparse MAVE is a model-free does not require any strong probabilistic assumptions about the predictors. [10] proposed penalised MAVE (P-MAVE) through combining bridge penalty with $l_1$-norms of the rows of a basis matrix.[15] combined MAVE with SCAD, adaptive Lasso and the MCP to produce SCAD-MAVE, ALMAVE and MCP-MAVE, respectively. [13] combined Lasso with the group-wise MAVE which suggested by [8].

In this paper ,we combine the dimension reduction method MAVE [14] with a Signal extraction approach to multivariate regression (SiER) [9] to propose a new variable selection method MAVE–SiER working under sufficient dimension reduction [1, 2] settings, MAVE - SiER has advantages over SiER because it extends SiER to multivariate response and nonlinear settings, has high efficient in dimension reduction and Computation, especially when the number of predictors is large.
The rest of the paper is organized as follows. We review of MAVE is provided in Section 2. SiER is discussed in Section 3 . Then in Section 4, we present the new approach MAVE – SiER. Section 5 contains the results from a simulation study. Finally, Section 6 ends of the paper with brief discussion.

## 2. Review of Minimum Average Variance Estimation (MAVE)

When applying regression models to high-dimensional data,[14] proposed the minimum average variance estimation (MAVE) approach to reduce dimension covariates for the conditional mean with fewer regularity conditions on the predictors.
Consider the regression model of a response Y $\in R^q$ on a vector $\mathbf{X} \in R^p$ can be written As

$$Y = g(B^T X) + \varepsilon \tag{2.1}$$

where g($\bullet$) is an unknown function, $B = (\beta_1, \dots, \beta_d)$ is a $p \times d$ orthogonal matrix ($B^T B = I_d$) with d < p and E( $\varepsilon$| $\mathbf{X}$) = 0 almost surely. [14] defined the d-dimensional subspace $B^T \mathbf{X}$ the effective dimension reduction (EDR) space. Given a random sample $\{(\mathbf{X_i},\ \mathbf{Y_i}),\ \ i = 1, \dots, n\}$, minimizes the objective function

$$argmin \left[ E|Y - E(Y|X^T B)| \right]^2 \tag{2.2}$$

It follows that

$$\left[ E|Y - E(Y|X^T B)| \right]^2 = E\{\sigma_B^2(B^T X)\} \tag{2.3}$$

over all $B \in$ R$^{p \times d}$ . It's equivalent to minimize the following problem

$$E\{\sigma_B^2(B^T X)\}, \ \ B^T B = I_d$$

A local linear expansion of at any For each j the following weighted sum of such linear approximation is minimized,

$$\sigma_B^2(B^T X) = argmin \left( \sum_{i=1}^{n} \left[ |Y - E(Y|X^T B)| \right]^2 \right) = argmin \left( \sum_{i=1}^{n} \left[ |Y_i - \{\alpha + (X_i - X_j)^T B b_j\}\right]^2 \right) \tag{2.4}$$

$w_{ij}$ is a function of the distance between $x_i$ and $x_j$.
The challenge of solving $B$ is same as to the following minimization:

$$argmin\left(\sum_{i=1}^{n}\left[|Y_i-\{\alpha+(X_i-X_j)^TBb_j\}\right]^2\right)$$

## 3. Signal extraction approach to multivariate regression

Propose by [10], for dimension reduction and regression in multiple response linear model with high-dimensional predictor variables . This approach considered the decomposition of the coefficient matrix B, let K denote the rank of B , each coefficient vector $\beta_j$ can be expressed as a linear combination of $w_1,\ldots,w_K$ . So we have the decomposition

$$\mathfrak{B}=AW^T=\alpha_{\mathbf{k}}\mathbf{w_1^T}+\ldots+\alpha_{\mathbf{k}}\mathbf{w_K^T} \tag{3.1}$$

Where A= [ $\alpha_1$ , . . . , $\alpha_K$] and W= [ $w_1$ , . . . , $w_K$] are p× K and q×K matrices, respectively. There are infinitely many choices of $w_1$ , . . . , $w_K$ hence the decomposition (3.1) is not unique, we will consider a different decomposition which leads to the best lower rank approximation to $\mathbf{X}\mathfrak{B}$ ,We call $\mathbf{X}\mathfrak{B}$ the signal matrix in the response matrix Y. Specifically, to find **A** and **W** we consider the singular value decomposition (SVD) of $\mathbf{X}\mathfrak{B}$,

$$\mathbf{X}\mathfrak{B}=\sigma_1\gamma_1\mu_1+\ldots+\sigma_K\gamma_K\mu_K \tag{3.2}$$

Where $\sigma_1 \geq \sigma_2 \geq \ . \ . \ . \geq \ \sigma_{\mathrm{K}} \ \geq 0$ are singular values of XB, $\gamma_k \in \mathbb{R}^n$and $u_k \in \mathbb{R}^q$, are the left singular vectors corresponding to $\sigma_k$. respectively, with $\|u_k\|_2=1$, $\|\gamma_k\|_2=1$ by the Eckart-Young Theorem $\sum_{j=1}^{k}\sigma_j\gamma_ju_j^T$ is the best rank k approximation to XB. We define the columns of **W** and **A**

$$\mathbf{W_k}=\frac{\sigma_k}{\sqrt{n}}u_k,\quad \alpha_k=\frac{n}{\sigma_k^2}\mathfrak{B}\mathbf{W_k},\qquad 1\leq k\leq K \tag{3.3}$$

Therefore $\sum_{j=1}^{k}X\alpha_jw_j^T=\sum_{j=1}^{k}\sigma_j\gamma_ju_j^T$is the best rank k approximation to $\mathbf{X}\mathfrak{B}$ and $\alpha_k^TS\alpha_k$= 1, for any 1≤ $k$ ≤K.
In the following for any 1≤ $k$ ≤K, let

$$\mathfrak{B}_{\mathbf{k}}=AW^T=\alpha_{\mathbf{k}}\mathbf{w_1^T}+\ldots+\alpha_{\mathbf{k}}\mathbf{w_K^T} \tag{3.4}$$

The sum of the first **k** terms of our decomposition. Our decomposition of $\mathfrak{B}$ leads to the following model transformation

$$Y=X\mathfrak{B}+\varepsilon=TW^T+\varepsilon=t_1w_1^T+\ldots+t_Kw_K^T+\varepsilon, \tag{3.5}$$

Where

$$T=[t_1,\ldots,t_K],\ and\, t_j=X\alpha_{\mathbf{j}},\quad 1\leq j\leq K.$$

To estimate the decomposition, we first estimate $\alpha_{\mathbf{1},\ldots,}$ $\alpha_{\mathbf{k}}$ by then estimate $t_1,\ldots,t_K$ Finally, based on model (3.5) and the least squares method, we obtain the estimates of $w_1\ldots,w_K.\alpha_{\mathbf{k}}$ is the solution to

$$Max\alpha^TB\alpha,\ subject\ \alpha^TS\alpha=1,\alpha_l^TS\alpha=0,1\leq l\leq k-1 \tag{3.6}$$

The approximation error of the best rank k approximation to $X\mathfrak{B}$ is

$$\left\| X\mathfrak{B} - \sum_{i=1}^{k} t_i w_i^T \right\|_F^2 = \left\| X\mathfrak{B} - X(\sum_{i=1}^{k} \alpha_i w_i^T) \right\|_F^2 = \| X\mathfrak{B} - X\mathfrak{B}_k \|_F^2 = n \sum_{i=k+1}^{k} u_i(\Xi) \; for\; any\; 1 \le k \le K$$

As $\sum_{k=1}^{K} X\alpha_k w_k^T$ the SVD of $X\mathfrak{B}$, we have that X$\alpha_k$ is the kth left-singular vector of $X\mathfrak{B}$, and hence it is the kth eigenvector of the matrix $(X\mathfrak{B})(X\mathfrak{B})^T$ with the corresponding eigenvalue $\sigma_k^2$, So we have

$$(\mathbf{X}\alpha_\mathbf{k})^T(\mathbf{X}\mathfrak{B})(\mathbf{X}\mathfrak{B})^T(\mathbf{X}\alpha_\mathbf{k}) = \sigma_k^2(\mathbf{X}\alpha_\mathbf{k})^T(\mathbf{X}\alpha_\mathbf{k}) \tag{3.7}$$

which implies that

$$\alpha_\mathbf{k}^\mathbf{T}\mathbf{B}\alpha_\mathbf{k} = \frac{\sigma_\mathbf{k}^\mathbf{2}}{\mathbf{n}}\alpha_\mathbf{k}^\mathbf{T}\mathbf{S}\alpha_\mathbf{k} \tag{3.8}$$

Our choice of W and A makes the signal concentrated in the first few components as much as possible. The estimates $\alpha_\mathbf{k}$ can be obtained by solving the following generalized eigenvalue problem.

$$Max\alpha^T\widehat{B}\alpha, subject \;\; \alpha^T S\alpha = 1, \quad \widehat{\alpha}_l^T S\alpha = 0, 1 \le l \le k-1 \tag{3.9}$$

Where we estimate **B** by

$$\widehat{B} = \frac{1}{n^2}X^T(Y - 1_n\overline{y}^T)(Y - 1_n\overline{y}^T)^T X \tag{3.10}$$

Where $\overline{y}$ is the sample mean of $y_1, \ldots, y_n$, and $1_n$ is an n-dimensional vector with all elements equal to one. In the classic setting of small p and large n, the estimates $\widehat{\alpha}_\mathbf{1}, \ldots, \widehat{\alpha}_\mathbf{k}$ can be sequentially obtained by solving (3.10)

only a small number of the coefficient vectors $\beta_1$,..., $\beta_p$, are nonzero vectors. Since these vectors are the row vectors of $\mathfrak{B}$, this assumption is the row-wise sparsity of $\mathfrak{B}$. implies that $\alpha_\mathbf{k}$ is a sparse vector and the number of its nonzero coordinates is less than or equal to the number of nonzero vectors among $\beta_1$,..., $\beta_p$. Motivated by the sparsity of $\alpha_\mathbf{k}$, we propose the following penalized optimization problem whose solution is the sparse estimate$\widehat{\alpha}_\mathbf{k}$ of $\alpha_\mathbf{k}$

$$Max\frac{\alpha^T\widehat{B}\alpha}{\alpha^T S\alpha \; + \tau \left\|\alpha\right\|_\lambda^2} \quad subject \; \alpha^T S\alpha = 0, \qquad 1 \le l \le k-1 \tag{3.11}$$

Where $\|\alpha\|_\lambda^2$ = (1- $\lambda$) $\|\alpha\|_2^2$ +$\lambda$ $\|\alpha\|_1^2$ , is a mixture of the squared $l_2$ and squared $l_1$ norms. and both $\tau \ge 0$ and $0 < \lambda < 1$ are tuning parameters. In the penalty $\tau \|\alpha\|_\lambda^2$ the $l_2$ term is used to overcome the singularity problem of **S** and the $l_1$ term encourages the sparsity of $\widehat{\alpha}_\mathbf{k}$.

In (3.11) scale-invariant, that is, if we replace $\alpha$ by $\mathbf{t}\alpha$, where **t** is any nonzero number, the value of the objective function is unchanged, Due to the scale-invariant property, (3.11) is equivalent to

$$Max\alpha^T\widehat{B}\alpha, \;\; subject \;\; \alpha^T S\alpha + \tau \left\|\alpha\right\|_\lambda^2 \le 1, \qquad \widehat{\alpha}_l^T S\alpha = 0, \quad 1 \le l \le k-1 \tag{3.12}$$

## 4. MAVE – SiER

Despite the fact that MAVE is a promising dimension reduction approach, the reduced variables are still linear combinations of all the original predictors. As a result, it faces the same interpretive challenges as most dimension reduction approaches. We use SiER in the following section to optimize (2.4) for MAVE since MAVE can be built easily as an iterative “ordinary least squares” technique, as shown in Section 2, from which we may estimate and choose relevant variables at the same time. This approach is known as MAVE – SiER .

### *4.1. Algorithm for MAVE – SiER*

1. Initialize m = 1, and set $\mathbf{B} = B_0$, any arbitrary p× 1 vector.

2. For given $\mathbf{B}$, solve $(a_j,\ \mathbf{b_j})$ where j = 1, . . . , n, from the following quadratic minimization problem:

$$margin \sum_{j=1}^{n} \sum_{i=1}^{n} [Y_i - \{\ \alpha + (X_i - X_j)^T Bb_j]^2 w_{ij}$$

3. For a given $(\widehat{a}_j,\ \widehat{\mathbf{b_j}})$ solve $\mathbf{B}_{\text{MSiER}}$ from the following minimization problem:

$$argmin\ [\sum_{j=1}^{n} \sum_{i=1}^{n} \{\ Y_i - \{\ \alpha + (X_i - X_j)^T Bb_j]^2 w_{ij}\} + \sum_{i=1}^{p} P_{\lambda}^{\text{SiER}}\, |\mathbf{B_m}| \tag{4.1}$$

4. Replace the $m-th$ column of $\mathbf{B}$ by $\mathbf{B_{MSiER}}$ and Iterate steps 2 and 3 to convergence.

5. Update $\mathbf{B}$ by $(\mathbf{B_1}_{\text{MSiER}}\ \mathbf{B_2}_{\text{MSiER}}, \ldots,\ \mathbf{B_3}_{\text{MSiER}}, \beta_0)$, and set $m$ to be $m+1$.

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

## 5. Simulation study

The aim of this section, we compare the performance of the proposed MAVE -SiER method with three related methods on simulated data.
The first method elastic net [16] proposed a technique which it can select groups of correlated variables and overcomes the difficulty of $p > n$. The elastic net is based on a combination of the ridge (L2) and the lasso (L1) penalties. The elastic net is defined in two stages. Assuming that the response is centered and the predictors are standardized. The third method is the SPLS [9] which identifies sparse latent components by maximizing the covariance between them and the responses with sparsity penalty imposed. The last method is the SiER, [10] it considers the decomposition of the coefficient matrix that leads to the best approximation to the signal part in the response given any rank, and estimates the decomposition by solving a penalized generalized eigenvalue problem .

The data were generated from the model $Y = XB + \epsilon$ , Sample size n was chosen as 50, 100 and 200, and we drawn 500 data replicates in each case. We set $\beta_{j1} = \frac{1}{\sqrt{15}}$ , for j = 1, . . . , 15 , $\beta_{j2} = \frac{0.5}{\sqrt{30}}$ , for j = 16 , . . . , 45 , $\beta_{j3} = \frac{0.25}{\sqrt{60}}$ , for j = 46 , . . . , 105 ,$\beta_{\text{jk}} = 0$ for others.For each i = 1, . . . , n, the first 150 predictors $(X_{i1}\ , \ldots,\ X_{i150})^T \sim \mathcal{N}_{150}$ (0,Σ ), where the 150×150 matrix Σ has the (j, k)th element $\Sigma_{\text{jk}} = \rho^{|j-k|}$ for $1 \le j, k \le 150$, and the other predictors are independent normal variables , $X_{\text{ij}} \sim \mathcal{N}$ (0,0.1$^2$ ) for $1 \le j \le$ p . The noise vector $\varepsilon$ case form $\mathcal{N}(0, \sigma^2 R)$

$$R = \begin{bmatrix} 1 & r & r^2 \\ r & 1 & r \\ r^2 & r & 1 \end{bmatrix}$$

Table 1: Summary of mean square error at Various Sample Size of 50 replicates for first case

| *p* | *n* | *q* | *p* | *r* | *MAVE– SiER* | *SiER* | *SPLS* | *Elastic net* |
|---|---|---|---|---|---|---|---|---|
| *500* | *100* | *3* | *0.3* | *0.2* | *2.150* | *1.299* | *2.077* | *2.217* |
| | | | | *0.9* | *2.082* | *1.249* | *1.698* | *1.806* |
| | | | *0.7* | *0.2* | *1.240* | *0.383* | *0.392* | *0.453* |
| | | | | *0.9* | *1.314* | *0.390* | *0.403* | *0.483* |
| | *500* | *3* | *0.3* | *0.2* | *2.474* | *2.014* | *2.257* | *2.622* |
| | | | | *0.9* | *2.623* | *1.973* | *2.205* | *2.427* |
| | | | *0.7* | *0.2* | *1.474* | *0.619* | *0.749* | *0.968* |
| | | | | *0.9* | *1.669* | *0.441* | *0.593* | *0.835* |
| | *1000* | *3* | *0.3* | *0.2* | *2.435* | *1.933* | *2.117* | *2.367* |
| | | | | *0.9* | *2.286* | *1.571* | *1.793* | *2.304* |
| | | | *0.7* | *0.2* | *1.511* | *0.724* | *0.734* | *0.884* |
| | | | | *0.9* | *1.526* | *0.424* | *0.465* | *0.473* |

In terms of sparsity and prediction precision, the SiER clearly outperforms the other methods (see Table 1). Among the competitors, MAVE-SiER is the worst. The mean square error of the SiER is lower than that of all methods.

Table 2: Summary of mean square error at Various Sample Size of 50 replicates for second case

| *p* | *n* | *q* | *p* | *r* | *MAVE– SiER* | *SiER* | *SPLS* | *Elastic net* |
|---|---|---|---|---|---|---|---|---|
| *500* | *100* | *3* | *0.3* | *0.2* | *3.092* | *3.526* | *4.980* | *3.396* |
| | | | | *0.9* | *3.012* | *3.617* | *4.670* | *3.268* |
| | | | *0.7* | *0.2* | *3.181* | *3.551* | *4.435* | *3.445* |
| | | | | *0.9* | *3.029* | *3.936* | *4.981* | *3.362* |
| | *500* | *3* | *0.3* | *0.2* | *3.069* | *3.356* | *4.253* | *2.993* |
| | | | | *0.9* | *3.245* | *3.703* | *4.824* | *2.993* |
| | | | *0.7* | *0.2* | *3.093* | *3.790* | *4.618* | *3.162* |
| | | | | *0.9* | *2.993* | *3.674* | *4.643* | *3.293* |
| | *1000* | *3* | *0.3* | *0.2* | *2.823* | *3.682* | *3.898* | *3.082* |
| | | | | *0.9* | *3.129* | *3.698* | *3.912* | *3.512* |
| | | | *0.7* | *0.2* | *2.453* | *3.181* | *3.842* | *2.730* |
| | | | | *0.9* | *2.240* | *2.995* | *3.278* | *2.650* |

While in the second case, from (Table 2), The data were generated from the model, $Y = B^T X / \left(0.5 + (B^T X + 1.5)^2\right) + 0.5\epsilon$, the MAVE-SiER clearly outperforms the other methods. Among the competitors, SPLS is the worst. elastic net performance is comparable to that of MAVE-SiER and better than the performance of the other methods. Furthermore, the mean square error of the MAVE-SiER is lower than that of all methods.

## 6. Discussion

In this paper,we combine the strength of a MAVE and SiER and proposed new approach we called MAVE - SiER. MAVE can estimate $S_{E(y|x)}$ while SiER sparse estimation and dimension reduction. The MAVE-SiER enable SiER to work with nonlinear regression. From the results, it is obvious that MAVE-SiER gives accurate prediction and encourages variable selection under sufficient dimension reduction settings.

## References

[1] H.S.K. Chun, *Sparse partial least squares regression for simulatenous dimension reduction and variable selection*, J. R. Stat. Soc. Ser. B Stat. Meth. (2010) 3–25.

[2] R. Cook, *Regression Graphics: Ideas for Studying the Regression Through Graphics*, New York, Wily, 1998.

[3] B.H. Efron, *Least angle regression*, Ann. Stat. (2004) 407—499.

[4] J.A. Fan, *Variable selection via non-concave penalized likelihood and its oracle properties*, J. Amer. Stat. Assoc. (2001) 1348–1360.

[5] H and Z. Hastie, *Regularization and variable selection via the elastic net*, J. Royal Stat Soci (2005) 1418–142.

[6] B. Leo, *Better Subset Regression Using the Nonnegative Garrote*, Technomet. (1995) 373–384.

[7] K. Li, *Sliced inverse regression for dimension reduction (with discussion)*, J. Amer. Stat. Assoc. (1991) 316–342.

[8] L.L.-X. Li, *Groupwise dimension reduction*, J. Amer. Stat. Assoc. (2010) 1188–1201.

[9] R. Luo X. Qi, *Signal extraction approach for sparse multivariate response regression*, J. Multivar. Anal. (2017) 83–97.

[10] Q. Wang and X. Yin, *A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE*, Comput. Stat. Data Anal. 52(9) (2008) 4512—4520.

[11] R. Tibshirani, *Regression shrinkage and selection via the Lasso* , J. Royal Stat. Soc. (1996) 267–288.

[12] T.X. Wang, *Penalized minimum average variance estimation*, Stat. Sinica (2013) 543–569.

[13] T.X. Wang, *Variable selection and estimation for semiparametric multiple-index models*, Bernoulli 21(2015) 242–275.

[14] Y.T. Xia, *An adaptive estimation of dimension reduction space*, J. Royal Stat. Soc. (2002) 363–410.

[15] Yu and K. Alkenani, *Sparse MAVE with oracle penalties*, Adv. Appl. Stat. (2013) 85—105.

[16] H. A. Zou, *Regularization and variable selection via the elastic net*, J. Royal Stat. Soc. (2005) 301—320.