Robust sparse MAVE through elastic net penalty

Naeem Abed Otaiwi [(1)] **Ali Alkenani** [(2)]

Stat.post15@qu.edu.iq ali.alkenani@qu.edu.iq

[1, 2] Dept. of Statistics, College of Administration and Economics

University of Al-Qadisiyah, Al Diwaniyah Iraq.

Abstract

The sparse MAVE-EN (SMAVE-EN) is a model-free variable selection method. The SMAVE-EN is a combination of Elastic Net (EN) and effective dimension reduction method *minimum average variance estimation* (MAVE).This approach is effective when the predictors are highly correlated under sufficient dimension reduction (SDR) settings. However, SMAVE-EN is not robust to outliers due to the use of least squares criterion which is sensitive to the presence of outlier in the data. In this article, we proposed a robust model-free variable selection method (RSMAVE-EN). This approach works under different error distributions settings. Thus, it gives robustness to existing outliers in the both dependent variable and independent variables. The effectiveness of the proposed approach is verified via both simulation studies and a real data analyses.

Key words: dimension reduction, MAVE, robust estimation, Elastic-Net

1. Introduction

Due to the explosion of big data in the last decades, high-dimensional data analysis has attracted significant research interest. However, due to the so-called "Curse of dimensionality " (Bellman,1961), it is complex to formulate validate parametric model for large number of covariates. The SDR (Li,1991; Cook,1998) methods provide an effective tool to deal with the mentioned problem in regression. The basic idea of SDR aims to replace the original high dimensional predictor vector with a suitable low-dimensional projection without much loss of the regression information. Let y is a response variable and $x = (x_1 ,x_2 , ….,x_p)^T$ is a $p \times 1$ predictor vector. The SDR explores a $p \times d$ matrix θ, such that $y \perp\!\!\!\perp x|x^T \theta$, where $\perp\!\!\!\perp$ indicates independence. The dimension reduction subspace (DRS) is the column space spanned by θ.The intersection of all DRS is known as the central subspace ($S_{y/x}$). The $S_{y/x}$ includes all the regression information of y/x (Yu and Zhu, 2013). Many approaches were introduced for finding ($S_{y/x}$) such as SIR (Li, 1991), SAVE (Cook and Weisberg,1991) and PHD (Li,1992) . Cook and Li (2002) proposed the concept of the central mean subspace ($S_{E(y/x)}$). For estimate ($S_{E(y/x)}$) , many methods of DR were introduced, for example the iterative Hessian transformation (Cook and Li, 2002) and , MAVE (Xia et al., 2002) . However, for SDR methods, the outcomes are stay linear combinations of all original predictors. Therefore, the SDR methods suffer from the difficulty in interpretation of the resulting estimates. Variable selection (v.s) methods aim to select the best subset of predictors among all possible subsets of predictors. Picking out the most important small subset of predictors makes the interpretation of the results easy, lower cost models and giving a good understanding of the dataset. Moreover, the selection of the important predictors

can improve the prediction accuracy of the model. In general the variable selection is divided into two types of methods, traditional and regularization methods .Examples for traditional methods are stepwise selection (Efroymson, 1960), AIC (Akaike, 1973) and BIC (Schwarz, 1978). When traditional methods are compared with regularization methods, traditional methods have several drawbacks, for example instability .To tackle the instability that affects traditional methods, a family of regularization methods are proposed to automatically select informative variables via continuous shrinkage. For example, Lasso (Tibshirani ,1996) , SCAD (Fan and Li , 2001) , Lars (Efron, Hastie, and Tibshirani, 2004), elastic net (Zou and Hastie , 2005) adaptive lasso (Zou , 2006) , adaptive elastic net (ADEN) (Zou and Zhang , 2009) and MCP (Zhang, 2010) among others . Under SDR framework settings, many procedures are proposed to combine the ideas of SDR methods and regularization methods by many researchers. For example, Li et al.(2005), Ni et al. (2005), Li and Nachtsheim (2006), Li (2007), Li and Yin(2008), Wang and Yin (2008), Wang et al. (2013) and Alkenani and Yu (2013). Recently, Alkenani and Rahman (2020) proposed SMAVE-EN method. The authors combine a popular SDR method, MAVE (Xia et al., 2002) and Elastic Net penalty to produce sparse and accurate estimates when the predictors are highly correlated under SDR settings. Moreover, variable selection and parameters estimation have been implemented simultaneously. However, SMAVE-EN is not robust to outliers due to the use of least squares criterion. In this article, we propose a robust SMAVE-EN (RSMAVE-EN), which can exhaustively estimate directions in the regression mean function also selects informative covariates simultaneously, whereas being robust to the existence of possible outliers. The effectiveness of our approach is verified through simulation studies and a real data analysis. The rest of the article is organized

as follows, in section 2 we briefly review the SDR, MAVE and SMAVE . The robust extension of SMAVE-EN is detailed in section 3. We compared the proposed method with a number of existing methods through simulation in section 4. In section 5, the real data analysis. Finally, the conclusions are reported in section 6.

2. Brief review of SDR, MAVE and SMAVE

2.1. Sufficient dimension reduction (SDR)

The regression-type model of a response variable $y \in R^1$ on a $p \times 1$ predictor vector X and the error term ε , assume the following model:

$$y = f(x_1, x_2, \ldots, x_p) + \varepsilon, \qquad (1)$$

where $f(x_1, x_2, \ldots, x_p) = E(y|x)$ and $E(\varepsilon|x) = 0$. SDR for the mean function aims to find a subset S of the predictor space such that

$$y \perp\!\!\!\perp E(y|x)|p_s x \qquad (2)$$

where $\perp\!\!\!\perp$ denotes independence and p (.) is a projection operator . Subspaces satisfying condition (2) are called mean DRS (Cook and Li, 2002). If d = dim(S) and $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ is a basis for S, the predictor X can be replaced by the linear combinations $\theta_1^T X, \theta_2^T X, \ldots, \theta_d^T X, = f(\theta^T x)$. The intersection of all subspaces satisfying (2), that is called the $d \leq p$ without loss of information on $E(y|x)$ that is, $f(x_1, x_2, \ldots, x_p)$ central mean subspace $S_{E(y|x)}$ (Cook and Li,2002). Many methods have been proposed to estimate $S_{E(y|x)}$ and MAVE (Xia et al., 2002) is one of the well-known methods of them.

2.2. MAVE

Xia et al.(2002) introduced MAVE such that the matrix θ is the solution of

$$\min_{\theta} \{E[y- E(y|\theta^T x)]^2 \}, \quad (3)$$

where $\theta^T\theta = I_d$ .The conditional variance given $\theta^T x$ is

$$\sigma^2_\theta(\theta^T x) = E[\{y – E(y|\theta^T x)\}^2 | \theta^T x]. \quad (4)$$

Thus,

$$\min_{\theta} E[y- E(y|\theta^T x)]^2 = \min_{\theta} E \{ \sigma^2_\theta(\theta^T x)\} \quad (5)$$

For any given X_0 , $\sigma^2_\theta(\theta^T x_0)$ can be approximated using local linear smoothing as

$$\sigma^2_\theta(\theta^T x_0) \approx \sum_{i=1}^{n}\{y_i - E (y_i|\theta^T x_i) \}^2 w_{i0}$$

$$\approx \sum_{i=1}^{n}[y_i - \{ a_0 + b_0^T \theta^T (X_i - X_j)\}]^2 w_{i0},$$

where $a_0 + b_0^T \theta^T (x_i – x_0)$ is the local linear expansion of $E (y_i| \theta^T x_i)$ at x_0 , and $w_{i0} \geq 0$ are the kernel weights centered at $\theta^T x_0$ with $\sum_{i=1}^{n} w_{i0}= 1$, so the problem of finding θ is by solving the following:

$$\min_{\theta: \theta^T\theta = I_d} \left(\sum_{j=1}^{n}\sum_{i=1}^{n}[y_i - \{ a_j + b_j^T\theta^T (x_i - x_j)\}]^2 w_{ij} \right) \quad (6)$$

where $\theta^T\theta = I_d$ and w_{ij} are kernel weights defined as a function of the distance between x_i and x_j . the minimization of (2) resolves iteratively with respect to { (a_j , b_j), j= 1, …. , n }, and θ separately. MAVE is a very efficient method, since only two quadratic programming problems are included and both have explicit solutions.

2.3. Sparse MAVE

Although MAVE is an efficient dimension reduction method, its outputs are still linear combinations of all original predictors. Therefore, it suffers the same difficulty in interpretation as other DR methods do. Wang and Yin (2008) combine a variable selection method Lasso (Tibshirani, 1996) with MAVE (Xia et al., 2002) to propose sparse MAVE (SMAVE). The authors incorporate an L_1 penalty term into the MAVE loss function in (2). SMAVE has advantages over Lasso because it extends Lasso to multidimensional and nonlinear settings without assuming any particular model. The SMAVE minimizes:

$$\left(\sum_{j=1}^{n} \sum_{i=1}^{n} [\, y_i - \{ a_j + b_j^T \theta^T (x_i - x_j) \}]^2 w_{ij} \right) + \lambda \sum_{k=1}^{p} |\theta_{m,k}|, \quad (7)$$

where m = 1, …. ,d and d is known and it can be estimated by BIC , $\| . \|_1$ is the L_1 *norm* and λ is nonnegative regularization parameter which controls the amount of shrinkage. Alkenani and Yu (2013) incorporate the adaptive Lasso, SCAD and MCP penalties with the loss function of MAVE in (2) to propose ALMAVE, SCAD-MAVE and MCP-MAVE, respectively. Wang et al. (2013) suggested penalized MAVE (P-MAVE) by incorporating a bridge penalty to L_1 norm of the rows of a basis matrix.

2.4. SMAVE-EN

The previous methods employed penalties that fail to work with grouped variables situation. Alkenani and Rahman (2020) proposed (SMAVE-EN) method. The authors combined a popular SDR method MAVE (Xia et al., 2002) with (EN) penalty (Zou and Hastie, 2005) to produce sparse and accurate

estimates when the predictors are highly correlated under SDR settings. The SMAVE-EN minimizes:

$$\left(\sum_{j=1}^{n}\sum_{i=1}^{n}[\, y_i - \{ a_j + b_j^T \theta^T (x_i - x_j) \}]^2 \, w_{ij} \right) + \lambda_1 \|\theta_m\|_2^2 + \lambda_2 \|\theta_m\|_1 , \quad (8)$$

where $\|.\|_2^2$ is L_2 norm related with ridge penalty and $\|\,.\,\|_1$ is L_1 norm related with Lasso penalty. λ_1 and λ_2 are the tuning parameters which control the amount of shrinkage. Under the same conditions of EN and MAVE, can be shown that the SMAVE-EN estimator has the same consistency rate as the MAVE estimator furthermore it is also as efficient as MAVE asymptotically.

3. The proposed approach

3.1 Robust estimation

Although SMAVE-EN has advantages over the other methods employed penalties. However, SMAVE-EN is not robust to outliers and the violation of distribution assumption error, due to the use of least squares criterion. Cizek and Hardle (2006) introduced a comprehensive study of the sensitivity of MAVE to outlier values and proposed the robust enhancement to MAVE by replacing the local least squares with local L- or M-estimation. The robust MAVE estimates can be written by minimizing:

$$\sum_{j=1}^{n}\sum_{i=1}^{n} p[\, y_i - \{ a_j + b_j^T \theta^T (x_i - x_j) \}] \, w_{ij} \quad , \qquad (9)$$

where $p(.)$ represent the robust loss function. Under this setting, Wang and Yao (2013) proposed a robust sparse MAVE to select the informative covariates robustly. The authors added an L_1 penalty into the expression (9) as follows:

$$\left(\sum_{j=1}^{n} \sum_{i=1}^{n} p\left[y_i - \left\{ a_j + b_j^T \theta^T \left(x_i - x_j \right) \right\} \right] w_{ij} \right) + \sum_{k=1}^{d} \lambda_k \left| \theta_k \right|_1 , \quad (10)$$

Where p (.) is a robust loss function, $|\,.\,|_1$ is the L_1norm and $\{\lambda_k , k = 1, 2 ,... , d\}$ are the nonnegative regularization parameters.

Alkenani (2020) proposed robust variable selection in SIR using Tukey's biweight criterion and ball covariance (RSSIR).

3.2 Robust SMAVE-EN

In this paper, we extend the robust estimation to variable selection and proposed RSMAVE-EN, which can exhaustively estimate directions in the regression mean function also select informative covariates simultaneously, whereas being robust to the existence of possible outliers in both the dependent and independent variables. To select the formative covariates robustly, the (EN) penalty can be introduced into the expression (9). RSMAVE-EN is proposed can be obtained by minimizing the following:

$$\sum_{j=1}^{n} \sum_{i=1}^{n} p\left[y_i - \left\{ a_j + b_j^T \theta^T \left(x_i - x_j \right) \right\} \right] w_{ij} + \lambda_1 \|\theta_m\|_2^2 + \lambda_2 \|\theta_m\|_1 , \quad (11)$$

where $p(.)$ represents a robust loss function, $\|\,.\,\|_1$ is the L_1 norm, $\|.\|_2^2$ is the L_2 norm and λ_1 and $\lambda 2$ are the tuning parameters. We choose $p(.)$ as a Tukey's biweight function to obtain robust estimation in both x and y, when the loss function has a redescending derivative, then the loss function is robust and resistant to outliers in x and y (Rousseeuw and yohai, 1984). The loss function of Tukey's biweight has this property (Tukey, 1960). Therefore the suggested

RSMAVE-EN is no sensitive to outliers in x and y. The minimizing in (11) is an robust version of the minimizing in (8) by replacing the least squares loss function in (8) by robust loss function with Tukey's biweight function. The function of Tukey's biweight is:

$$p_c(u) = \begin{cases} \left(\frac{c^2}{6}\right)\left\{1-\left[1-\left(\frac{u}{c}\right)^2\right]^3\right\} & if\,|u| \le c \\ \frac{c^2}{6} & if\,|u| > c, \end{cases} \qquad (12)$$

where c is tuning level of the robustness. For obtain 95% asymptotic efficiency at the standard normal distribution, Value of c assumed 4.685.

The RSMAVE-EN algorithm is as follows:

For a given sample $\{(y_i, x_i),\ i = 1,2,\ldots..,\ n\}$,

1.initialize m=1 and set $\theta = \theta_0$,any arbitrary $p\times 1$ vector.

2. For given θ, solve (a_j, b_j), where $j= 1,2, \ldots., n$, from the following minimization problem:

$$\min_{a_j,\ b_j,\ j=1,2,\ldots,n} \left(\sum_{j=1}^{n}\sum_{i=1}^{n} p[\, y_i - \{ a_j + b_j^T \theta^T (x_i - x_j)\}]\, w_{ij} \right) \qquad (13)$$

3. For given $(\hat{a}_j, \hat{b}_j)$, j=1,2,….,n, solve $\theta_{\text{mRSMAVE-EN}}$ from :

$$\min_{\theta:\theta^T\theta = I_m} \left(\sum_{j=1}^{n}\sum_{i=1}^{n} p\left[y_i - \left\{\hat{a}_j + \hat{b}_j^T\left(\hat{\theta}_1, \hat{\theta}_2, \ldots., \hat{\theta}_{m-1}, \hat{\theta}_m\right)^T (x_i - x_j)\right\}\right] w_{ij} + \lambda_1 \|\theta_m\|_2^2 + \lambda_2 \|\theta_m\|_1 \right) \qquad (14)$$

4. Replace the mth column of θ by $\hat{\theta}_{\text{mRSMAVE-EN}}$ and iterate between step2and 3 until convergence is attained.

5. Update θ by $(\hat{\theta}_{1\text{RSMAVE-EN}}, \hat{\theta}_{2\text{RSMAVE-EN}}, \ldots, \hat{\theta}_{\text{mRSMAVE-EN}}, \hat{\theta}_0)$, and set m to m+1

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

where w_{ij} are the kernel weights :

$$w_{ij} = k_h\{\hat{\theta}^T(X_i - X_j)\} / \sum_{i=1}^{n} k_h\{\hat{\theta}^T(X_i - X_j)\}$$

k_h represent the refined multidimensional Gaussian kernel,

$h_{opt} = A(d)\, n^{-1/(4+d)}$ is the optimal bandwidth , where $A(d) = \left\{\frac{4}{(d+2)}\right\}^{1/(4+d)}$

and d is the dimension of the kernel function. See (Xia et al., 2002) for the more details.

3.3 Tuning parameter selection

Some information criterion, for example Akaike's information criterion (AIC) (Akaike, 1973), Bayesian information criterion (BIC) (Schwars,1978) and the residual information criterion (RIC) (Shi and Tsai, 2002) are often used for selecting λ according to the following formulas, respectively :

$$AIC = n \log(RSS / n) + 2\, p(\lambda) \qquad (15)$$

$$BIC = n \log(RSS / n) + \log(n)\, p(\lambda) \qquad (16)$$

$$RIC = \{n - p(\lambda)\}\log(RSS/ n - p(\lambda)) + p(\lambda)\{\log(n)-1\} + \{4/ (n - p(\lambda))\}, \qquad (17)$$

where $p(\lambda)$ denotes the number of non-zero coefficients and *RSS* is the residual sum of squares of the Lasso fit, that defined as:

$$RSS=\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - \left\{\hat{a}_j + \hat{b}_j^T(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{m-1}, \hat{\theta}_m)^T(x_i - x_j)\right\}\right]^2 w_{ij}, \qquad (18)$$

Shi and Tsai (2002) showed that using RIC for selection λ gives better performance, and it is a consistent criterion. In this paper, we employed a robust version of RIC, which is proposed by Alkenani (2020) as follows:

$$RRIC = \{n - p(\lambda)\}\log(RRSS/n - p(\lambda)) + p(\lambda)\{\log(n) - 1\} + \{4/(n - p(\lambda))\}, \quad (19)$$

where $$RRSS = \sum_{j=1}^{n}\sum_{i=1}^{n} p\left[y_i - \left\{a_j + b_j^T \theta^T (x_i - x_j)\right\}\right] w_{ij}, \quad (20)$$

4. Simulation study

The purpose of this section is to assess the finite sample performance of our proposed RSMAVE-EN method through simulation studies. We compare the suggested method (RSMAVE-EN) with SMAVE-EN (Alkenani and Rahman, 2020) and RSMAVE (Wang and Yao, 2013). We conduct a comparison to show the behavior of RSMAVE-EN method in terms of the prediction accuracy and variable selection of our proposed method. The reported simulation outcomes were based on 100 data replications. Also, we consider the distributions of x and ε for each of the following three examples are as follow:

1. $N(0, 1)$, the standard normal.
2. $t_3/\sqrt{3}$, t- distribution with 3 degree of freedom.
3. $0.95\,N(0, 1) + 0.05\,N(0, 10^2)$.
4. $0.95\,N(0, 1) + 0.05\,U(-50, 50)$, the standard normal were contaminated with 5% uniform distribution.

Example 1: Let $d = 2$ and $p = 8$. The data are generate from the following regression model: $$Y = \frac{\theta_1^T X}{0.5 + (1.5 + \theta_2^T X)^2} + \sigma\varepsilon,$$

Where $\theta_1 = (3, 1.5, 2, 0, 0, 0, 0, 0)^T$, $\theta_2 = (0, 0, 0, 0, 0, 3, 1.5, 2)^T$, $X \in R^8$ and $\sigma = 3$. Consider θ_1, the first 3 predictors were highly correlated with pairwise correlation $r = 0.7$, whereas the last five were uncorrelated. θ_2, the first 5

predictors were uncorrelated, whereas the rest predictors were correlated with pair wise correlation r = 0.7 .

Example 2: consider the model: $y = 1+2(\theta^T x +3)\times \log(3|\theta^T x|+1) + \varepsilon$,

Let $d = 1$, $p = 40$. Consider $\theta = (\underbrace{0,\dots,0}_{10}, \underbrace{2,\dots,2}_{10}, \underbrace{0,\dots,0}_{10}, \underbrace{2,\dots,2}_{10})^T$,

where corr $(i, j) = 0.5$ for all i and j .

Example 3 :We adopt the same model as the previous example 2, where $d = 1$, $p = 40$ and consider $\theta = (\underbrace{3,\dots,3}_{15}, \underbrace{0,\dots,0}_{25})^T$, the predictors X are :

$x_i = z_1 + \varepsilon$, $i = 1,\dots, 5$,

$x_i = z_2 + \varepsilon$, $i = 6,\dots,10$,

$x_i = z_3 + \varepsilon$, $i = 11,\dots,15$,

xi , $i = 16,\dots,40$.

When $i = 1, \dots, 15$. We have three groups in this model, within each group there are five predictors. Also, we have 25 predictors and set the coefficients of there to zero.

Table1. Results of estimation accuracy comparisons for example1, based on the average number of zero coefficients (Ave.0's), mean squared error (MSE) and absolute correlation for($\theta_1^T x, \theta_1^{\hat{}T} x$) and ($\theta_2^T x, \theta_2^{\hat{}T} x$).

Dist.	method	Ave.0's	MSE	| corr($\theta_1^T x, \theta_1^{\hat{}T} x$) |	| corr($\theta_2^T x, \theta_2^{\hat{}T} x$)|
1	SMAVE- EN	8	1.642	0.848	0.389
	RSMAVE	8	1.631	0.828	0.579
	RSMAVE-EN	8	1.617	0.852	0.618
2	SMAVE- EN	7	1.727	0.814	0.190
	RSMAVE	7	1.714	0.796	0.534
	RSMAVE-EN	8	1.685	0.813	0.595
3	SMAVE- EN	6.5	1.784	0.477	0.242
	RSMAVE	7	1.643	0.811	0.533
	RSMAVE-EN	7.6	1.640	0.852	0.609
4	SMAVE- EN	6.5	1.775	0.746	0.288
	RSMAVE	7	1.723	0.761	0.362
	RSMAVE-EN	8	1.707	0.769	0.377

Table 2.results for example 2 , based on the average number of zero coefficients (Ave.0's), mean squared error (MSE) and the absolute of correlation between $(\theta^T x,\ \hat{\theta}^T x)$

dist.	method	Ave.0's	MSE	$\lvert Corr(\theta^T x,\ \hat{\theta}^T x)\rvert$
1	SMAVE-EN	12	1.653	0.994
	RSMAVE	12.5	1.608	0.996
	RSMAVE-EN	12.5	1.593	0.998
2	SMAVE-EN	10	1.650	0.978
	RSMAVE	11	1.591	0.999
	RSMAVE-EN	13	1.589	0.999
3	SMAVE-EN	11.5	1.691	0.977
	RSMAVE	11	1.677	0.987
	RSMAVE-EN	12.5	1.652	0.994
4	SMAVE-EN	11.5	1.701	0.980
	RSMAVE	11	1.640	0.996
	RSMAVE-EN	12.5	1.635	0.997

Table 3. results for example 3, based on the average number of zero coefficients (Ave.0's), mean squared error (MSE) and the absolute of correlation between $(\theta^T x,\ \hat{\theta}^T x)$

dist.	method	Ave.0's	MSE	$\lvert Corr(\theta^T x,\ \hat{\theta}^T x)\rvert$
1	SMAVE-EN	14.5	2.936	0.823
	RSMAVE	11.5	2.830	0.971
	RSMAVE-EN	14.5	2.823	0.990
2	SMAVE-EN	14	2.923	0.836
	RSMAVE	12	2.841	0.968
	RSMAVE-EN	14.5	2.838	0.973
3	SMAVE-EN	13.5	3.020	0.807
	RSMAVE	11.5	2.894	0.939
	RSMAVE-EN	14.5	2.892	0.940
4	SMAVE-EN	13.33	3.030	0.733
	RSMAVE	12.5	2.855	0.974
	RSMAVE-EN	14.5	2.840	0.975

From outcomes of tables 1, 2 and 3 for the previous three examples, the comparison demonstrated that, the three reported methods yielded similar results in case of standard normal distribution, in both variable selection and estimation accuracy. Whereas in case of other three distributions of x and error, we can note that SMAVE-EN method was sensitive about contamination but other methods RSMAVE and RSMAVE-EN were not affected because they have the robustness. Also, the performance of RSMAVE-EN outperformed RSMAVE method in terms of variable selection and estimation accuracy.

Depending on the above observations it is clear that under various settings, the proposed RSMAVE-EN has a very good performance in terms of variable selection and estimation accuracy.

5. Boston housing data

Harrison and Rubinfeld (1978) collected the data. This data set consist of n = 506 and p = 14, where y is the median value of owner-occupied homes in $1000's (medv). X includes 13 predictors on the 506 census tracts in Boston city. The predictors are : x_1 is (rate of crime), x_2 is (proportion of residential land zoned), x_3 is (proportion of non-retail business acres), x_4 is (the Charles river (= 1 if tract bounds river; 0 otherwise)), x_5 is (concentration of nitric oxides), x_6 is (average of rooms), x_7 is (proportion of owner-occupied units), x_8 is (weighted mean of distances), x_9 (index of accessibility),x_{10} is (rate of property tax), x_{11} (pupil – teacher ratio), x_{12} is (proportion of black population) and x_{13} is (lower status). The data set is available and public from R package. The predictors and y are standardized separately for ease of explanation. To verify the performance of the proposed RSMAVE-EN, we re-analyzed the data set by adding some outliers in x and y. four cases were considered in this analyzed, a percentage of 5%, 10% and 15% contaminated observations. Table 4 explains that, to evaluate the estimation accuracy for proposed method, we conducted a comparison based on the mean squared error (MSE), residual square error (RSE) and R-squared. Also, we reported the number of selected variables by SMAVE-EN, RSMAVE and RSMAVE-EN.

Table 4. Results of the comparison of estimation accuracy and variable selection for SMAVE-EN, RSMAVE and RSMAVE-EN.

Outliers	Method	Number of selected variables	MSE	RSE	R^2
No outlier	SMAVE-EN	11	0.236	0.489	0.764
	RSMAVE	11	0.238	0.491	0.761
	RSMAVE-EN	11	0.235	0.488	0.764
5%	SMAVE-EN	13	0.289	0.541	0.710
	RSMAVE	12	0.242	0.495	0.757
	RSMAVE-EN	11	0.241	0.494	0.758
10%	SMAVE-EN	13	0.303	0.554	0.696
	RSMAVE	13	0.260	0.512	0.741
	RSMAVE-EN	11	0.259	0.511	0.741
15%	SMAVE-EN	13	0.318	0.567	0.678
	RSMAVE	12	0.271	0.522	0.730
	RSMAVE-EN	11	0.269	0.520	0.731

From the results of table (4) it is obvious that the implement of SMAVE-EN, RSMAVE and RSMAVE-EN are very similar for the data set without contamination. Whereas after adding outliers to original data we can note that for all cases of contamination, a percentage of 5%, 10% and 15% the SMAVE-EN was sensitive to outliers and it is clearly affected in both estimation accuracy and variable selection. On the other hand, the results showed that the RSMAVE-EN has a slight superiority over its competitor RSMAVE. Thus, the outcomes of the comparison prove that the performance of proposed RSMAVE-EN was very consistent results even with all cases.

6. Conclusion

We have proposed RSMAVE-EN method. It is a robust approach to variable selection and dimension reduction simultaneously. The outcomes of numerical studies for both simulations and real data analysis have shown that the proposed RSMAVE-EN has a good behavior in a variable selection and estimation accuracy even with the outliers exist in predictors x and response variable y. Our simulation studies demonstrated for various distributions of error and predictors x that the proposed RSMAVE-EN outperformed the competitors RSMAVE and SMAVE-EN. In addition, the results of real data analysis demonstrated that the suggested method has good and very consistent results even with all contamination cases which considered through comparison with other methods RSMAVE and SMAVE-EN. The proposed method can be extended to other SDR approaches.

References

Alkenani, A. (2020). Robust variable selection in sliced inverse regression using Tukey biweight criterion and ball covariance. Journal of Physics Conference Series, 1664 012034.

Alkenani, A. and Rahman, E. (2021). Regularized MAVE through the elastic net with correlated predictors, Journal of Physics Conference Series, 1897 012018.

Alkenani, A. and Yu, K. (2013). Sparse MAVE with oracle penalties. Advances and Applications in Statistics 34, 85–105.

Cook, R. (1998). Regression graphics: ideas for studying the regression through graphics. New York, Wily.

Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. The Annals of Statistics 30, 455–474.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). Journal of the American Statistical Association 86, 328–332.

Efron, B. et al. (2004). Least angle regression. The Annals of Statistics 32, 407–499.

Fan, J. and Li, R. Z. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Li, L. (2007). Sparse sufficient dimension reduction. Biometrika 94, 603–613.

Li, K. (1991). Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association 86, 316–342.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. Journal of the American Statistical Association 87, 1025–1039.

Li, L. et al. (2005). Model-free variable selection. Journal of the Royal Statistical Society Series B, 67, 285–299.

Li, L. et al. (2010). Groupwise dimension reduction. Journal of the American Statistical Association. 105, 1188–1201.

Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. Technometrics 48, 503–510.

Li, L. and Yin, X. (2008). Sliced Inverse Regression with regularizations. Biometrics 64, 124–131.

Ni, L. et al. (2005). A note on shrinkage sliced inverse regression. Biometrika 92, 242–247.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of s-estimators. In Robust and Nonlinear Time Series Analysis, pages 256-272.

Stamey, T. et al. (1989) Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii: radical prostatectomy treated patients. J. Urol., 16, 1076–1083.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. Contributions to Probability and statistics, 2:448-485.

Wang, Q. and Yin, X. (2008). A Nonlinear Multi-Dimensional Variable Selection Method for High Dimensional Data: Sparse MAVE. Computational Statistics and Data Analysis 52, 4512–4520.

Wang, Q. and Yao, W. (2013). Robust Variable Selection Through MAVE. Computational Statistics and Data Analysis 63,42-49.

Wang, T. et al. (2013). Penalized minimum average variance estimation. Statist. Sinica 23 543–569.

Wang, T. et al. (2015). Variable selection and estimation for semi parametric multiple-index models. Bernoulli 21 (1), 242–275.10

Xia, Y.et al. (2002). An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society Series B 64, 363–410.

Yu, Z. and Zhu, L. (2013). Dimension reduction and predictor selection in semi parametric models. Biometrika, 100, 641-654.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. 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, Series B 67, 301–320.

Zou, H. (2006). The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–142