Robust sparse minimum average variance estimation through Adaptive elastic net

Sanaa Jabbar Tuama and Prof. Dr. Ali J. Kadhim Alkenani
E-mail:dmin.stat21.3@qu.edu.iq

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

Abstract

Regression analysis is a difficult method when there are many variables. In other words, as the number of variables increases, the model becomes more complex. This may lead to a dimensional problem. Some explanatory variables do not have a significant effect on the dependent variable, and some of these variables also have an internal correlation with each other, and this requires excluding such variables in order to increase the accuracy of the model. There are two ways to reduce the dimensions, namely the method of selecting variables (v.s) variable selection and variables extractions. Under the assumptions of the theory of SDR (Sufficient dimension reduction), the researchers worked on proposing methods to reduce the dimensions, including the integration of SDR methods with regularization methods (Regularization method) and the methods of regulation mean adding a penalty limit to control the complexity of the model as it greatly reduces the variance of the model, and among these methods SMAVE-AdEN (Alkenani and Rahman,2020) is a method for selecting a variable under the assumptions of SDR theory.

The SMAVE-AdEN method is a combination of Adaptive elastic net with MAVE (Minimum average variance estimator) method for estimating minimum average variance. This method is effective when the variables are highly correlated under SDR assumptions. But the SMAVE-AdEN method is not immune and it is a sensitive method that is affected when there are outliers in the data, owing to the least squares criteria that we employ. In this paper, we proposed a robust method (RSMAVE-AdEN), which can estimate parameters and select variables simultaneously, and is not affected by the presence of outliers in explanatory variables and response variables. The effectiveness of the proposed method was verified by a simulation study.

<u>Key words</u>: Adaptive Elastic Net, Robust estimation, MAVE, Dimension reduction.

1.Introduction

The study of regression when there are a large number of variables and a large sample size is a difficult and complex process, as it increases the complexity of the regression model, which prompted researchers to use the variable selection process because some explanatory variables are not essential in their impact on the variable. The dependent variable, or its effect is similar to the effects of other variables, and many of these variables have an internal connection with each other, which leads to the emergence of the problem of multicollinearity, and thus its effect is not significant, which calls for the exclusion of non-significant variables and the selection of significant variables To increase the accuracy of the model prediction.

This problem led the researchers to work on reducing the high dimensions of the data, as Cook proposed in (1998) the (Sufficient dimension reduction) method, this method is of high importance as one of the effective tools to address the issue of high-dimensional data analysis. Several dimension reduction SDR methods have been presented, one of which is the MAVE method (Xia et al., 2002). However, the results are linear combinations of all variables. Therefore, these methods suffer from the difficulty of interpreting the resulting estimates. Many methods have been proposed to combine SDR methods with regularization methods. These methods are able to deal with high-dimensional data, which are based on the principle of minimizing the sum of squares of error by adding a certain restriction to the parameters and reducing some coefficients and set others equal to zero,

It gives a sparse model that includes the least possible number of variables and is interpretable. For example, the researchers Alkenani and Rahman (2020) suggested the SMAVE-EN method, where the researchers combined the MAVE (Minimum Average Variance Estimator) method proposed by the researcher Xia and others in general. (2002) with the flexible EN network proposed by Zou and Hastie in (2005), and this method is characterized by the ability to deal with variables that are in highly correlated groups.

The researchers Alkenani and Rahman (2020) suggested the SMAVE-AdEN method, where the researchers combined the MAVE method (Minimum Average Variance Estimator) proposed by the researcher Xia et al. (2002) with the Adaptive Elastic Net (Adaptive Elastic Net) proposed by the researchers Zou and Helen (2009) to produce the SMAVE-AdEN method, and this method is characterized by giving accurate estimates when the variables are highly correlated. Moreover, the selection of the variable and the estimation of the parameters are done at the same time. Despite these good advantages of this method, it loses its efficiency if there are abnormal values in its data, and this is the problem.In this paper, we proposed a robust method (RSMAVE-AdEN), which can estimate parameters and select variables simultaneously, and is not affected by the presence of outliers in explanatory variables and response variables and this is the goal of the research. The effectiveness of the proposed method was verified by a simulation study.

2. Several dimension reduction (SDR)

A response variable's regression-type model yϵ R^1 on a Px1 predictor vector X and the error term ε, Suppose the following model:

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

where $f(x_1, x_2, \ldots, x_p) = E(y/x)$, $E(\varepsilon/x)=0$.

The aim of SDR for the mean function is to select a subset S of the predictor space

where y□ $E(y/x)/p_s x$, (2)

Thus,□ denotes independence, p(.)is an operator that performs projections.

Mean DRS are subspaces that satisfy condition (2) (Cook and Li ,2002).

If d=dim(s) ,$\theta = (\theta_1, \theta_2, .., \theta_d)$is a basis for S. The linear combinations can be used in place of the predictor X.

$\theta_1^T X, \theta_2^T X, \ldots, \theta_d^T X = f(\theta_x^T)$. (3)

The intersection of all subspaces (2) is referred to as the intersection. that is called the $d \leq p$ without loss of information on $E(y/x)$ that is ,

$f(x_1, x_2, \ldots, x_p)$ called the central mean subspace SE(y|x) (Cook and Li 2002). Many methods for estimation SE(y/x) have been proposed, with MAVE(Xia et al.,2002) being one of the most well-known. Methods used by them.

2.1 **Minimum average variance estimator** (MAVE)

In(2002) It has been suggested before Xia et al. the MAVE method and it means the least variance rate estimation method. This method is used on a wide range of regression models, and it is one of the SDR method to reduce dimensions, and this method has advantages, the most important of which is flexibility, its ability to combine with other methods, and its ability to choose variables. The estimation of the parameters simultaneously, as well as the availability of its algorithms and the ease of implementation, but the disadvantages of this method are that it does not give a single solution, but rather includes all the linear structures of all the original variables:

so $\theta^T \theta = I_d$,

The provided conditional variance $\theta^T x$ is

$\sigma_\theta^2(\theta^T x) = E[\{y - E(y/\theta^T x)\}^2 | \theta^T x]$ (4)

So,

$\min_\theta E[y - E(y|\theta^T x)]^2 = \min E\{\sigma_\theta^2(\theta^T x)\}$, (5) For any given X_0 , $\sigma_\theta^2(\theta^T x)$ can be approximated using local Linear smoothing as

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

$$\approx \sum_{i=1}^{n}[y_i - (a_{0+} b_0^T \theta^T (X_{i-}X_j))\}]^2 w_{i0} , \quad (6)$$

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 $\sum_{i=1}^{n} w_{i0}=1$ And you know like this

$$w_{ij} = k_h\left\{\hat{\theta}^T(X_i - X_j)\right\} / \sum_{i=1}^{n} k_h\left\{\hat{\theta}^T(X_i - X_j)\right\} , \quad (7)$$

$$\text{Min}_{\theta:\theta^T\theta=I_d}(\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - \left\{a_j + b_j^T\theta^T(x_i - x_j)\right\}\right]^2 w_{ij}, \quad (8)$$

2.2 Sparse minimum average variance estimator(SMAVE)

Although the MAVE method is an effective dimensionality reduction method, its outputs are still linear combinations of all variables, so it suffers from the difficulty of interpretation as other DR (dimensionality reduction) methods do. Therefore, several methods were proposed to combine V.S methods and SDR methods in one step. In 2008, researchers Wang and Yin proposed a (SMAVE) method that combines the Lasso method with the MAVE method. SMAVE has advantages over Lasso in that it extends multidimensional and nonlinear settings without assuming any particular form. SMAVE is defined by the following equation:

$$\min(\sum_{j=1}^{n}\sum_{i=1}^{n}[y_i\text{-}\left\{a_j + b_j^T\theta^T(x_i - x_j)\right\}]^2 W_{ij}+\lambda\sum_{k=1}^{p}\left|\theta_{m,k}\right| , \quad (9)$$

2.3 SMAVE-EN

The researchers Alkenani and Aljobori (2021) presented a study on the sensitivity of the SMAVE-EN method to outliers and proposed a robust enhancement to SMAVE-EN that can estimate trends in the mean regression function and identify covariates simultaneously, while it is impervious to the presence of possible outliers in each of dependent and independent variables. It is defined by the following equation: (10):

$$\sum_{j=1}^{n}\sum_{i=1}^{n}[y_i - \{a_j+b_j^T\theta^T(x_i - x_j)\}]^2 W + \lambda_1\|\theta_m\|_2^2 + \lambda_2\|\theta_m\|_1 , \quad (10)$$

2.4 SMAVE-AdEN

In the year (2020) the researchers (Alkenani and Rahman) proposed a new method (SMAVE-AdEN), the SDR method is integrated, Resulting from a combination of MAVE (Xia, 2002) with with the Adaptive Elastic Net method (Zou and Zhang, 2009), which is to combine the ridge regression method with the Lasso penalty function method to get an accurate scattered estimate. The SMAVE-AdEN method is defined by the following equation:

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

3. Robust Estimation

The most used methods for estimating the parameters of the statistical model are the maximum possibility (ML) Maximum liklihooh, the least squares (OLS) Ordinary Least Sequared and moments (M.OM) and others. In recent years, we have dealt with the case of anomalies in the data. In other words, when there are anomalies in the data, how are they dealt with? The answer is dealt with through robust estimation methods or robust estimation methods, where robust capabilities with high efficiency are obtained compared to the usual methods in the event that there are abnormal values in the data. It is also assumed that the robust method's capabilities are very close to the capabilities of the ordinary method when No outliers.

3.1 Robust SMAVE

The least squares criteria used by the SMAVE method make it sensitive to outliers In their 2006 investigation of the susceptibility of MAVE to outliers, researchers Cizek and Hardle proposed a significant improvement over SMAVE by the following equation Definition of the RMAVE method :

$$\sum_{j=1}^{n}\sum_{i=1}^{n} \rho[y_i - \{a_j + b_j^T\, \theta^T (X_i - X_j)\}]w_{ij} \ , \qquad (12)$$

Researchers Wang and Yao (2013) suggested the R SMAVE method and added a penalty term to equation 12), so that equation (13) is as follows:

$$\sum_{j=1}^{n}\sum_{i=1}^{n} \rho[y_i - \{a_j + b_j^T\, \theta^T (X_i - X_j)\}]w_{ij} + \sum_{K=1}^{d} \lambda_k\, |\theta_k| \quad , \qquad (13)$$

Alkenani (2021) proposed the (RSSIR) method to select a immune variable in the SIR method, using Tukeys Biweight, Criterion for Bioweight and Ball Covariance.

3.2 Robust SMAVE-EN

The researchers Alkenani and Aljobori (2021) presented a study on the sensitivity of the SMAVE-EN method to outliers and proposed a robust reinforcement for RSMAVE -EN, which can estimate trends in the mean regression function and identify covariates simultaneously, while it is immune to the presence of possible outliers in each of the dependent and independent variables. It is important to know this method by equation:

$$\sum_{j=1}^{n}\sum_{i=1}^{n} \rho\left[y_i - \left\{a_j + b_j^T \theta^T \left(x_i - x_j\right)\right\}\right] w_{ij} + \lambda_1 \|\theta_m\|_2^2 + \lambda_2 \|\theta_m\|_1, \qquad (14)$$

3.3 Robust SMAVE-AdEN

Although the SMAVE-AdEN method proposed by researchers (Alkenani and Rahman) in (2020) has good advantages for selecting variables and estimating parameters, easy to implement, and has good prediction accuracy compared to the current methods, it is not vulnerable to outliers, and we suggested a robust enhancement for SMAVE- AdEN by replacing the local least squares with an estimate of -L or -M. The RSMAVE-AdEN immune method can be defined by (15) as follows:

$\sum_{j=1}^{n}\sum_{i=1}^{n} p[y_i - \{a_j + b_j^T\theta^T(x_i - x_j)b_j^T\theta^T(x_i - x_j)w_{ij} + \lambda_1\|\theta_m\|_2^2 + w_k^*\lambda_2\|\theta_m\|_{1,}$ (15)

We utilize p (.) as a Tukey's biweight function to get an estimate in both x and y. As a result, the loss function is robust and resistant to outliers in both x and y when it has a redescending derivative [Rousseeuw and Yohai (1984)].This is a characteristic of Tukey's biweight's loss function[Tukey(1960)]. As a result, the proposed RSMAVE-AdEN is not sensitive to x and y outliers. By substituting Tukey's biweight function for the least squares loss function used in the minimizing in (8), the minimizing in (11) is a robust version of the minimizing in (8). And Tukey's function can be expressed in equation (16) as follows:

$$P_C(u)=\begin{cases}(\frac{c^2}{6})\left\{1-\left[1-(\frac{u}{c})^2\right]^3\right. , & if\ |u|\le c \\ \frac{c^2}{6}, if\ |u|>c \end{cases} , \quad (16)$$

RSMAVE-AdEN estimates can be obtained according to the following algorithm:

1-Let m=1,$\theta=\theta_0$, any arbitrary Px1 vector

2-To know θ,get(a_j,b_j),where j=1,.,…,n,from

$\text{Min}_{a_jb_j=1,\dots,n}(\sum_{j=1}^{n}\sum_{i=1}^{n} p[y_i - \{a_j + b_j^T\,\theta^T(X_i - X_j)\}]w_{ij}$, (17)

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

$\min_{\theta:\theta^T\theta=I_m}(\sum_{j=1}^{n}\sum_{i=1}^{n} P[y_i - \{\hat{a}_j + \hat{b}_j^T(\hat{\theta}_1,\hat{\theta}_2,\dots\dots,\hat{\theta}_{m-1},\hat{\theta}_{m,})^T$

$$(x_i - x_j)\}w_{ij} + \quad \lambda_1\|\theta_m\|_2^2 + w_k^*\lambda_2\|\theta_m\|.$$

4-Substitute the mth colom of θ by $\hat{\theta}_{mRSMAVE-AdEN}$ and use 2 and 3 unit convergence.

5- Upgrade θ by($\hat{\theta}_{1RSMAVE-AdEN}$, $\hat{\theta}_{2RSMAVE-AdEN}$,….., $\hat{\theta}_{mRSMAVE-AdEN}$, $\hat{\theta}_0$)

And group m to be m+1.

6-In case m<d,keep on with 2 to 5 until m=d.

The kernel weights defined by the following equation:

$W_{ij}= k_h\{\hat{\theta}^T(x_i - x_j)\}/k_h\{\hat{\theta}^T(x_i - x_j)\}$, (18)

4. Simulation Study

This section's goal is to evaluate our suggested RSMAVE-AdEN method's finite sample performance using simulation tests. We evaluate the SMAVE-AdEN (Alkenani and Rahman,2020)and the recommended approach (RSMAVE-AdEN), RSMAVE-EN Alkenani and Aljobori (2021). We compare the results to show how well the RSMAVE-AdEN method performs in terms of prediction accuracy and variable selection. The researcher wrote a code in R language to calculate the proposed method,The reported simulation results were based on 200 iterations of the data.

1. *Normal* (0, 1), the standard normal.
2. $t_3/\sqrt{3}$, *t*- distribution with(3) degree of freedom.
3- (1-α) N(0,1)+ α N(0,10^2).

4- (1-α) N(0,1)+ α U(-50,50)

With regard to the distributions in cases 4 and 3, (1-α)% of the data comes from the standard normal distribution and α% from other distributions.(Wang &Yao,2013)

Example: First example: $y = 1+2(\theta^T x + 3) \times \log(3|\theta^T x| + 1) + \varepsilon$, Let $d = 1$,n=100,200, $p = 40$.

Consider $\theta = (\underbrace{0,\dots,0}_{10}, \underbrace{2,\dots,2}_{10}, \underbrace{0,\dots,0}_{10}, \underbrace{2,\dots,2}_{10},)^{T}$,

Example: Second example: $y = 1+2(\theta^T x + 3) \times \log(3|\theta^T x| + 1) + \varepsilon$, Let $d = 1$,n=100,200, $p = 40$.

Consider $\theta = (, \underbrace{3,\dots,3}_{15}, \underbrace{0,\dots,0}_{25})^{T}$,

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

$x_i = z_1 + \varepsilon$, $i = 1, \dots, 5$, $x_i = z_2 + \varepsilon$, $i = 6, \dots, 10$,

$x_i = z_3 + \varepsilon$, $i = 11, \dots, 15$, x_i, $i = 16, \dots, 40$.

When, I = 1, …., 15. In this model, there are three groups and five predictors in each group. Additionally, we set the coefficients of 25 predictors.

Table 1: Results of the first example, when(Sample volume) n = 100, p = 40 and contamination percentage is 5% for two distributions. 3 and 4.

Distribution	Method	Ave.0's	MSE	$\|Corr (\theta^T x, \hat{\theta}^T x)\|$
1	RSMAVE-AdEN	14	0.073717	0.957768
	RSMAVE-EN	14	0.074843	0.956479
	SMAVE-AdEN	13	0.065121	0.969947
2	RSMAVE-AdEN	15	0.098880	0.975045
	RSMAVE-EN	14	0.231806	0.952287
	SMAVE-AdEN	11	0.382312	0.867518
3	RSMAVE-AdEN	15	1.954264	0.783236
	RSMAVE-EN	14	1.545765	0.786991
	SMAVE-AdEN	11	2.275063	0.683299
4	RSMAVE-AdEN	15	0.656871	0.972751
	RSMAVE-EN	15	0.862388	0.791489
	SMAVE-AdEN	13	2.888324	0.653346

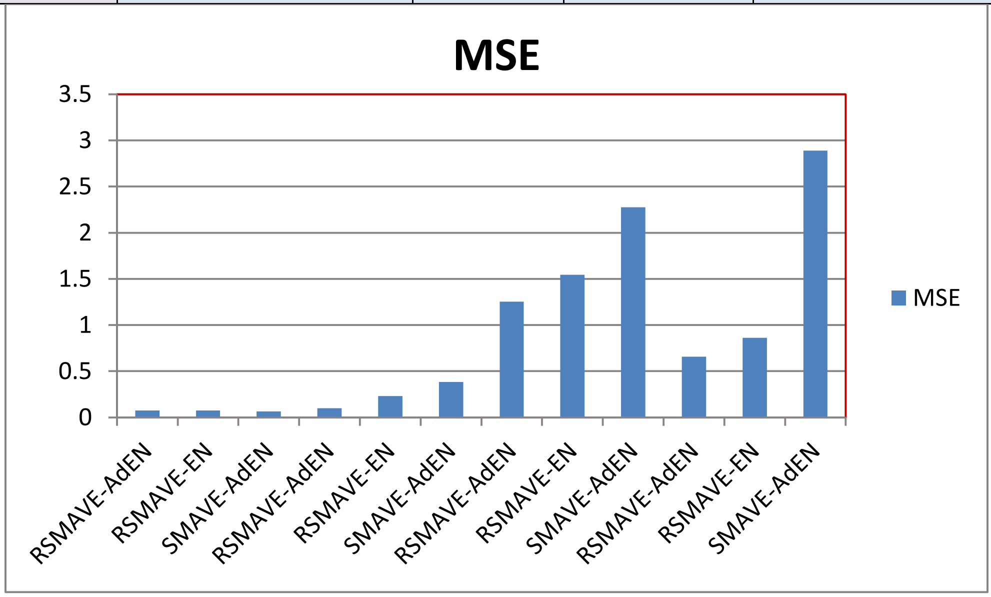

Figure (1) at n= 100 and a pollution rate of 5%

Table 2: Results of the first example, when (Sample volume) n = 100, p = 40 and contamination percentage is 10% for two distributions. 3 and 4..

Distribution	Method	Ave.0's	MSE	$\lvert Corr (\theta^T x, \hat{\theta}^T x) \rvert$
1	RSMAVE-AdEN	16	0.103626	0.957768
	RSMAVE-EN	15	0.103893	0.956479
	SMAVE-AdEN	12	0.101244	0.966947
2	RSMAVE-AdEN	17	0.188978	0.925891
	RSMAVE-EN	15	0.287284	0.894435
	SMAVE-AdEN	11	0.474869	0.817289
3	RSMAVE-AdEN	16	2.236655	0.674032
	RSMAVE-EN	15	2.667088	0.542085
	SMAVE-AdEN	11	4.825417	0.400728
4	RSMAVE-AdEN	17	0.827871	0.961538
	RSMAVE-EN	15	1.437581	0.940663
	SMAVE-AdEN	12	3.993977	0.850105

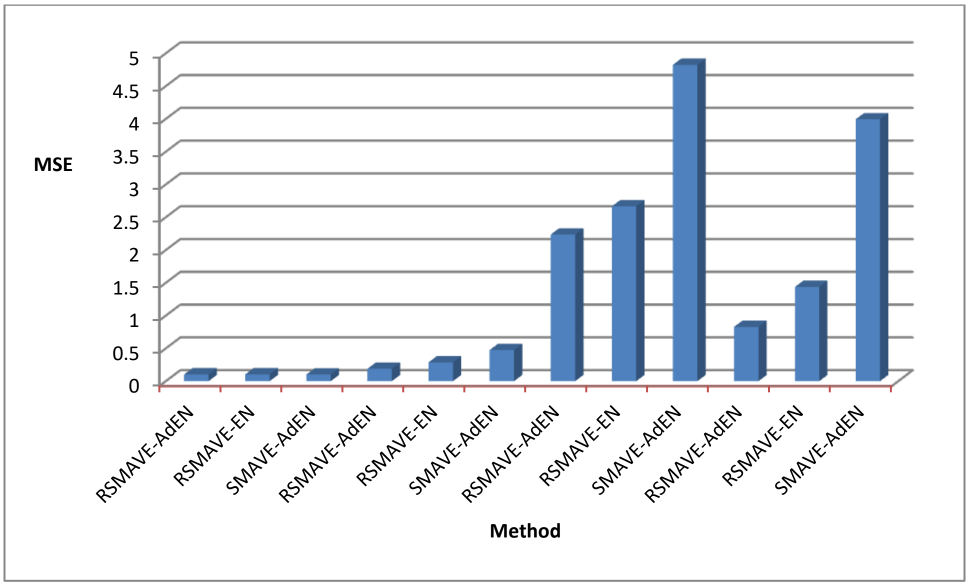

Figure (2) at n= 100 and a pollution rate of 10%

Table 3:

Results of the first example, when(Sample volume)n = 200, p = 40 and contamination percentage is 5% for two distributions. 3 and 4.

Distribution	Method	Ave.0's	MSE	$\lvert Corr (\theta^T x, \hat{\theta}^T x) \rvert$
1	RSMAVE-AdEN	17	0.030207	0.989215
	RSMAVE-EN	15	0.031847	0.977498
	SMAVE-AdEN	11	0.026728	0.994402
2	RSMAVE-AdEN	16	0.042231	0.983752
	RSMAVE-EN	14	0.156892	0.975321
	SMAVE-AdEN	11	0.228871	0.927181
3	RSMAVE-AdEN	17	0.307871	0.949726
	RSMAVE-EN	15	0.664467	0.925868
	SMAVE-AdEN	11	0.953373	0.923392
4	RSMAVE-AdEN	17	0.048539	0.973941
	RSMAVE-EN	15	0.193108	0.963446
	SMAVE-AdEN	11	0.500422	0.952597

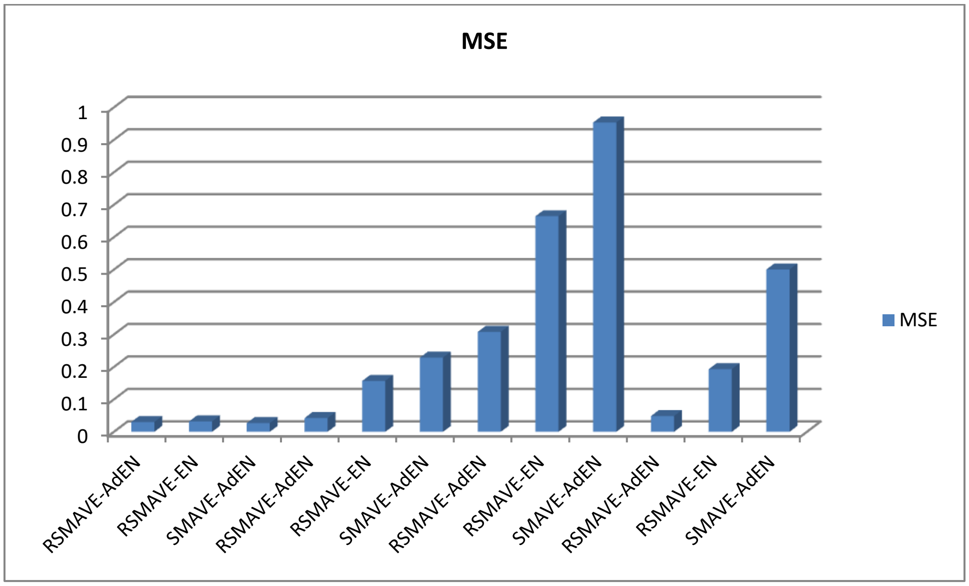

Figure (3) at n= 200 and a pollution rate of 5%

Table 4: Results of the first example, when (Sample volume) n = 200, p = 40 and contamination percentage is 10% for two distributions 3 and 4.

Distribution	Method	Ave.0's	MSE	$\|Corr (\theta^T x, \hat{\theta}^T x)\|$
1	RSMAVE-AdEN	15	0.040297	0.972215
	RSMAVE-EN	15	0.051849	0.977498
	SMAVE-AdEN	14	0.036728	0.978402
2	RSMAVE-AdEN	16	0.049932	0.983752
	RSMAVE-EN	16	0.159899	0.975321
	SMAVE-AdEN	11	0.524359	0.927181
3	RSMAVE-AdEN	16	0.407872	0.949726
	RSMAVE-EN	15	0.699467	0.925868
	SMAVE-AdEN	11	0.954373	0.923392
4	RSMAVE-AdEN	16	0.098588	0.973941
	RSMAVE-EN	15	0.199984	0.963446
	SMAVE-AdEN	12	0.587731	0.952597

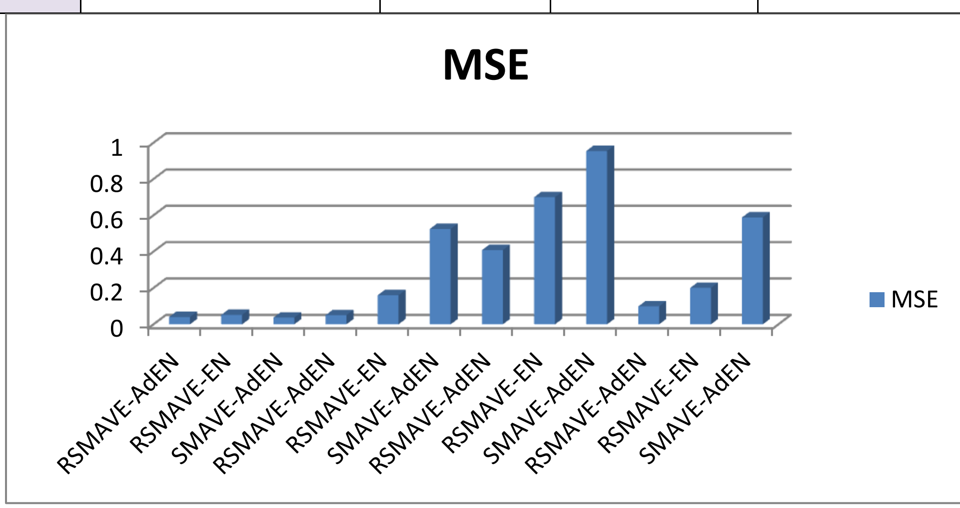

Figure (4) at n= 200 and a pollution rate of 10%

Table5: Results of the second example, when (Sample volume) n = 100, p = 40 and contamination percentage is 5% for two distributions. 3 and 4.

Distribution	Method	Ave.0's	MSE	|Corr ($\theta^T x, \hat{\theta}^T$ x) |
1	RSMAVE-AdEN	22	0.204056	0.944004
	RSMAVE-EN	21	0.210365	0.937518
	SMAVE-AdEN	18	0.201173	0.950409
2	RSMAVE-AdEN	22	0.145295	0.956032
	RSMAVE-EN	22	0.207786	0.955059
	SMAVE-AdEN	18	0.165431	0.887389
3	RSMAVE-AdEN	22	1.445401	0.703849
	RSMAVE-EN	21	1.820911	0.735499
	SMAVE-AdEN	19	1.942613	0.713441
4	RSMAVE-AdEN	20	2.020744	0.849991
	RSMAVE-EN	20	2.276505	0.791489
	SMAVE-AdEN	15	2.883258	0.853346

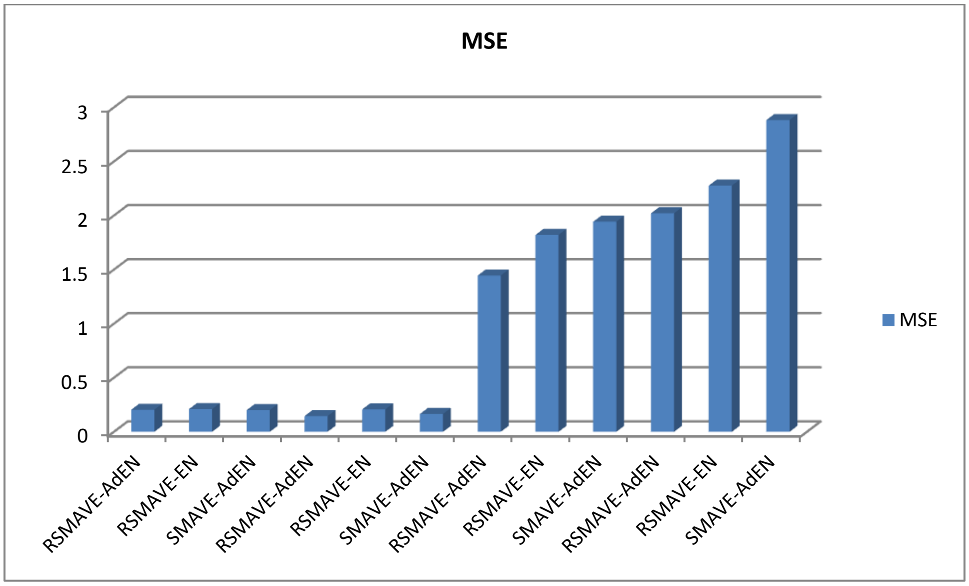

Figure (5) at n= 100 and a pollution rate of 5%

Table 6: Results of the second example, when(Sample volume) n = 100, p = 40 and contamination percentage is 10% for two distributions. 3 and 4.

Distribution	Method	Ave.0's	MSE	|Corr ($\theta^T x, \hat{\theta}^T$ x) |
1	RSMAVE-AdEN	23	0.248207	0.944208
	RSMAVE-EN	22	0.275993	0,949595
	SMAVE-AdEN	21	0.241677	0.950898
2	RSMAVE-AdEN	22	0.217998	0.935793
	RSMAVE-EN	21	0.259298	0.899450
	SMAVE-AdEN	19	0.292874	0.785773
3	RSMAVE-AdEN	22	1.007161	0.369866
	RSMAVE-EN	22	1.50184	0.260525
	SMAVE-AdEN	18	1.87707	0.418881
4	RSMAVE-AdEN	22	0.959013	0.894399
	RSMAVE-EN	22	1.266276	0.889945
	SMAVE-AdEN	18	1.443551	0.872766

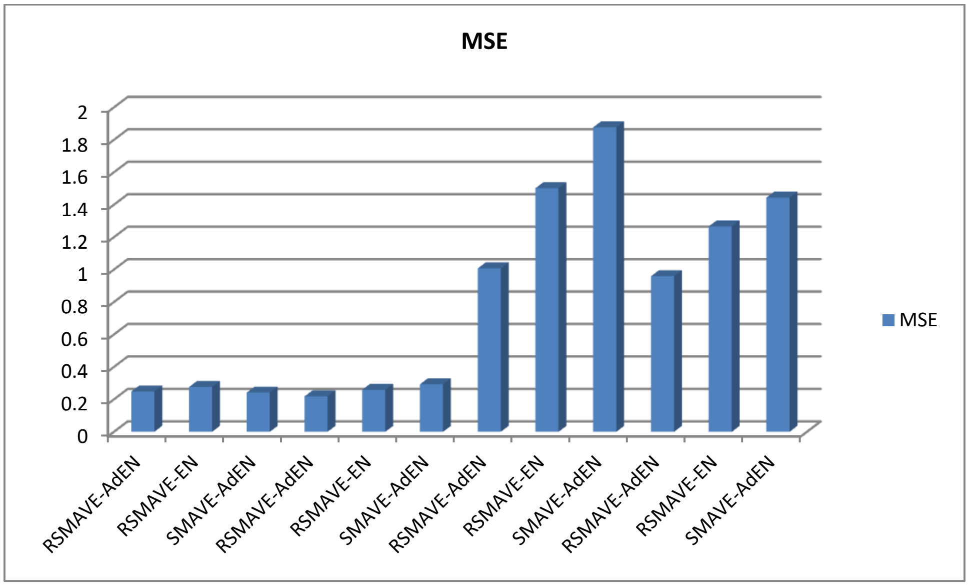

Figure (6) at n= 100 and a pollution rate of 10%

Table 7: Results of the second example, when (Sample volume) n = 200, p = 40 and contamination percentage is 5% for two distributions. 3 and 4.

Distribution	Method	Ave.0's	MSE	$\lvert Corr(\theta^T x, \hat{\theta}^T x) \rvert$
1	RSMAVE-AdEN	22	0.365259	0.915231
	RSMAVE-EN	22	0.367096	0.911084
	SMAVE-AdEN	18	0.327409	0.927016
2	RSMAVE-AdEN	22	0.362877	0.935677
	RSMAVE-EN	22	0.391598	0.925903
	SMAVE-AdEN	14	1.030230	0.847927
3	RSMAVE-AdEN	22	3.08434	0.456506
	RSMAVE-EN	22	3.30445	0.554282
	SMAVE-AdEN	19	3.72319	0.583353
4	RSMAVE-AdEN	23	4.577491	0.769963
	RSMAVE-EN	22	5.428073	0.769431
	SMAVE-AdEN	19	6.175188	0.814019

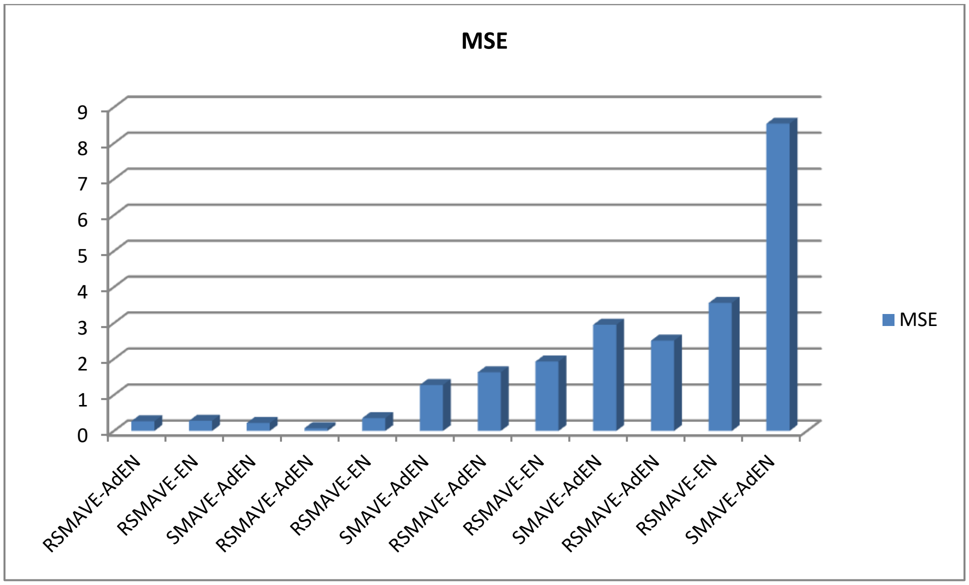

Figure (7) at n= 200 and a pollution rate of 5%

Table 8: **Results of the second example, when(Sample volume) n = 200, p = 40 and contamination percentage is 10% for two distributions. 3 and 4.**

Distribution	method	Ave.0's	MSE	$\lvert$Corr ($\theta^T x$,$\hat{\theta}^T$ x) $\rvert$
1	RSMAVE-AdEN	20	0.268495	0.942695
	RSMAVE-EN	21	0.280064	0.940263
	SMAVE-AdEN	20	0.222348	0.950288
2	RSMAVE-AdEN	20	0.078353	0.971359
	RSMAVE-EN	20	0.356162	0.961048
	SMAVE-AdEN	14	1.280471	0.886569
3	RSMAVE-AdEN	21	1.628861	0.628161
	RSMAVE-EN	21	1.928952	0.377362
	SMAVE-AdEN	14	2.95167	0.624616
4	RSMAVE-AdEN	20	2.510605	0.895776
	RSMAVE-EN	20	3.558851	0.615418
	SMAVE-AdEN	15	8.544191	0.685529

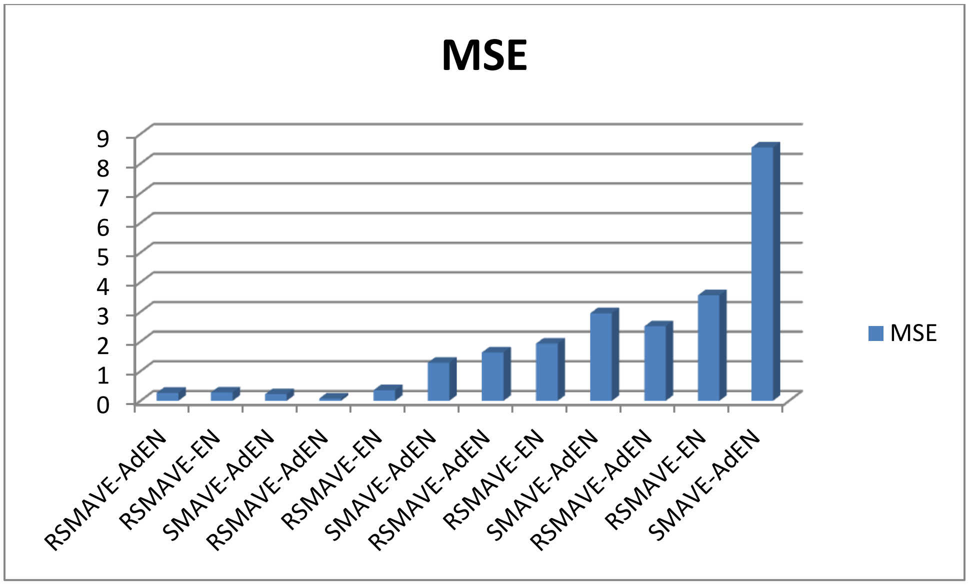

Figure (8) at n= 200 and a pollution rate of 10%

By comparing the results of the tables between the three methods used above, we note the following:

1- The proposed method R SMAVE-AdEN is the best and for all contamination cases (5%, 10%) and for the sample size (100,200), as it gave the lowest value for MSE, followed by the RSMAVE-EN method, as it gave the highest value for MSE. It also gave the highest correlation compared to (RSMAVE-EN) and (SMAVE-AdEN) method.

2-We note from the results of the tables that the MSE value increases with the increase in the pollution value.

6. Conclusion

Here are the most important conclusions:

1- The proposed method RSMAVE-AdEN in this research is a robust method for selecting variables and reducing dimensions at one time.

2- This method is efficient when the variables are highly correlated within the SDR numbers.

3- The proposed method has a good and consistent performance and gave the lowest value for MSE by comparing the three methods and is not affected by the presence of abnormal values for all contamination cases (5%, 10%) and for the sample size (100,200), but the higher the contamination percentage MSE value rose.

4- The proposed method works with high dimensions within the framework of SDR.

Abbreviations

SDR : Sufficient dimension reduction
AdEN : Adaptive elastic net
DR : Dimension reduction
EN : Elastic net
Lasso : Least absolute shrinkage and selection operator
MAVE : Minimum average variance estimator
MSE : Mean squared error
OLS : Ordinary least squared
OP's : Oracle properties
V.S : Variable selection
SMAVE: Sparse minimum average variance estimator

References

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

[2] Alkenani, A. (2021). Robust group identification and variable selection in sliced inverse regression using Tukey's biweight criterion and ball covariance. Gazi University Journal of Science, DOI: 10.35378/gujs.735503

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

[4] Alkenani, A. and E. Rahman. (2020,). Sparse minimum average variance estimation via the adaptive elastic net when the predictors correlated. J.Physics Conference Series , 1591, 012041.

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

[6]Alkenani and A.(2021). ROBUST SPARSE MAVE THROUGH ELASTIC NET PENALTY. J. Agricult. Stat. Sci. 17, Supplement 1, 2039-2046 Budapest.

[7]Cizek,P. and Hardle, W. (2006). Robust estimation of dimension reduction space. Computational Statistics and data analysis, 51, 545-555.

[8] Cook, R. (1998). Regression Graphics: Ideas for studying the Regression through Graphics. New York, Wiley.

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

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

[11] Tukey, J.W. (1960). A survey of sampling from contaminated distributions. Contributions to Probability and Statistics, 2, 448-485.

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

[13] Wang, Q. and W. Yao (2013). Robust Variable Selection Through MAVE. Computational Statistics and Data Analysis, 63, 42-49.
[14] Xia, Y., H. Tong, W. Li and L. Zhu (2002). An adaptive estimation of dimension reduction space. J. Royal Statistical Society, Series B 64, 363-410.
[15]Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476), 1418–1429.

[16] Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. J. Royal Statistical Society,Series B 67, 301-320.
[17]Zou, H., and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics, 37(4), 1733.