Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 24, Issue 1 (2022) ISSN$_{Print}$: 1816-9171

Analyzing of diabetes data using sparse MAVE methods

Ali Alkenani
Naeem Abed Otaiwi Aljobori
College of Administration and EconomicsUniversity of Al-Qadisiyah, Iraq.

Corresponding Author: Naeem Abed Otaiwi Aljobori

Abstract: The SDR (sufficient dimension reduction) is one of the important topics in many scientific fields. It has attracted attention of researchers because it is considered a beneficial approach to address the problem of the high dimension HD that has emerged due to the explosion of large data in the last decades. Under SDR framework settings, many procedures are proposed to combine the ideas of SDR methods and regularization approaches. In this paper, we present some of these methods, the SMAVE-EN (sparse MAVE - elastic net), RSMAVE (robust sparse MAVE) and RSMAVE-EN (robust sparse MAVE - elastic net). Also, the diabetic data are analyzed through the mentioned methods**.**

Keyword: SDR, MAVE, MAVE-EN, RSMAVE, RSMAVE-EN.

1. Introduction:

Because of the explosion of large data over the past two decades, HD (high dimensional) problem appears in a lot of scientific fields. Therefore, the statistical analysis becomes difficult. A beneficial approach to remedy this problem is to reduce the p-dimensional predictors vector X without much loss of information on regression. This reduction has been obtained via the SDR [1];[2]. Moreover, the SDR methods can be provide us with approach to get sufficient dimensions without a parametric model. Let y is a response variable and $x = (x_1, x_2, \dots, x_p)^T$ 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 $\perp\!\!\!\perp$ refers to independence and dimension reduction subspace (DRS) is the column space spanned by θ. The intersection of all DRS is named the central subspace ($S_{y/x}$).The $S_{y/x}$ involves all regression information of y/x [3]. A lot of methods were introduced for obtaining ($S_{y/x}$). For example, SIR [1], SAVE [4] and PHD [5]. Whereas, [6] presented the concept of central mean subspace (CMS) ($S_{E(y/x)}$). A number of DR approaches have been presented for estimate ($S_{E(y/x)}$), such as MAVE [7]. Actually, each DR component is considered a linear combination of all original predictors. This drawback makes the SDR methods suffer difficult to explain the resulting estimates. The goal of variable selection (V.S) ways is to select the best subset of predictors from all subsets of predictors. This means that, these ways of V.S play an important role in constructing a multiple regression model. Furthermore, the selection of a suitable subset of predictors improves the prediction accuracy. Also, the choice of a small subset of predictors makes interpretation of the results easier than a large set. There are some researchers who have been interested in V.S by penalizing the least squares, such as Lasso [8], SCAD [9], Elastic Net (EN) [10], adaptive Lasso [11], adaptive elastic net (ADEN) [12] and MCP [13]. On the other hand, new ideas have been introduced by some researchers when they are combined SDR methods and regularization methods. Such as, Li et al. (2005), Ni et al. (2005), Li and Nachtsheim (2006), Li (2007), Li and Yin (2008). Whereas, [14] proposed SMAVE, [15] introduced (P- MAVE), [16] introduced SCAD-MAVE, ALMAVE and MCP-MAVE, respectively. Along this line, [17] proposed SMAVE-EN. [18] suggested RSMAVE and [19] proposed RSMAVE-EN. This paper is organized as follows: the SDR and MAVE are presented in section 2. Section 3 describes the SMAVE-EN, RSMAVE and RSMAVE-EN methods. Real data analyses are analyzed in section 4. The conclusions are illustrated in section 5.

2. SDR and MAVE

2.1 SDR

The SDR is a beneficial approach to remedy the problem of HD. The idea of SDR approach works to replace the original HD of predictor vector with a suitable low dimensional projection without losing a lot of regression information. Assume the regression model as follows:

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

where y ϵ R^1 is a response variable, X= (x_1, x_2 , …. , x_p) is a $p \times 1$ predictor vector X and ε is the error term.

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

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

where $\perp\!\!\!\perp$ denotes independence and p (.) represents a projection operator . Subspaces which satisfying condition (2) are called mean DRS (dimension reduction subspace) [6]. Intersection of all DRS is called the central subspace $S_{y/x}$. The $S_{y/x}$ involves the regression information of y/x [3]. If d = dim(S) and $\theta = (\theta_1 , \theta_2 , …., \theta_d)$ is a basis for S, the predictor X can be replaced by the linear combinations

$$\theta_1^T X, \theta_2^T X, \dots, \theta_d^T X = f(\theta^T X) , d \leq p$$

Without loss of information on E(y|x). The intersection of all subspaces satisfying (2), that is called the central mean subspace $S_{E(y|x)}$ [6]. A number of methods have been introduced to estimate $S_{E(y|x)}$ such as iterative Hessian transformation [6] and MAVE [7] among other.

2.2 MAVE

The MAVE has been proposed by [7] as the matrix θ is solution of:

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

where $\theta^T\theta = I_d$ .The conditional variance given $\theta^T x$ is $\sigma_\theta^2(\theta^T x) = E[\{y – E(y|\theta^T x)\}^2 | \theta^T x]$. (4)

Thus,

$$\min E[y- E(y|\theta^T x)]^2 = \min E \{ \sigma_\theta^2 (\theta^T x)\} \qquad (5)$$

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

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

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

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 typically centered at $\theta^T x_0$. The selecting of the weights w_{ij} plays vital role in searching for the effective DR.

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

k_h represent the refined multidimensional Gaussian kernel, $h_{opt} = A(d)^{-1/(4+d)}$ 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 [7] for the more details.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.

3. Brief review of the methods used in the analysis

3.1 SMAVE-EN

Alkenani and Rahman (2020) proposed SMAVE-EN method. The authors combined the popular MAVE approach [7] and the EN penalty to produce a sparse and accurate estimate. The SMAVE-EN considered one of the efficient of the SDR methods that works with highly correlated predictors. The minimize of the SMAVE-EN is:

$$\left(\sum_{j=1}^{n}\sum_{i=1}^{n}[y_i - \{a_j + b_j^T\theta^T(x_i - x_j)\}]^2 w_{ij} \right) + \lambda_1 \|\theta_m\|_2^2 + \lambda_2\|\theta_m\|_1 , \quad (7)$$

where $\|\cdot\|_1$ and $\|\cdot\|_2^2$ are the L_1 norm and L_2 norm respectively, λ_1 and λ_2 are the tuning parameters which control the amount of shrinkage.

3.2 RSMAVE

[20] introduced a study about the sensitivity of MAVE to outlier values and suggested the robust enhancement to MAVE where, the local least squares have been replaced by local L- or M- estimation. The robust MAVE estimates can be written by minimizing:

$$\sum_{j=1}^{n}\sum_{i=1}^{n} p[y_i - \{a_j + b_j^T\theta^T(x_i - x_j)\}] w_{ij} , \quad (8)$$

where $p(.)$ represent the robust loss function. Under this setting, the robust SMAVE (RSMAVE) has been proposed by [18]. The authors added the *L1* penalty into the expression (8) as follows:

$$\left(\sum_{j=1}^{n}\sum_{i=1}^{n} p[y_i - \{a_j + b_j^T\theta^T(x_i - x_j)\}] w_{ij} \right) + \sum_{k=1}^{d} \lambda_k |\theta_k|_1 , \quad (9)$$

Where p (.) is a robust loss function, $|\cdot|_1$ is the L_1norm and $\{\lambda_k , k = 1, 2 ,\ldots , d\}$ are the nonnegative regularization parameters. [21] proposed robust variable selection in SIR using Tukey's biweight criterion and ball covariance (RSSIR).

3.3 RSMAVE-EN

The robust SMAVE-EN (RSMAVE-EN) has been introduced by Alkenani and Naeem (2021) when the authors have been combined the EN penalty and the expression (8). The RSMAVE-EN method can be obtained by the following minimizing:

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

where $p(.)$ represents a robust loss function, $\| \cdot \|_1$, $\|.\|_2^2$ are the L_1 norm and L_2 norm respectively, λ_1 and $\lambda 2$ are the tuning parameters. The RSMAVE-EN can exhaustively estimate directions in the regression mean function also select the informative covariates simultaneously. Moreover, the RSMAVE-EN considered a robust approach to the existence of possible outliers in both the dependent variable and independent variables.

4. Real data

In this section, we used the SMAVE-EN, RSMAVE and RSMAVE-EN methods, in analysis diabetic patient's data. We collected the data of a sample of 105 persons who visited Imam Sadiq Hospital in Al-Hila city during March and April (2021). We considered y represents the reading of blood sugar. X includes 20 predictors as follows: x_1 is Urea (blood urea), x_2 is Creat. (Creatinine), x_3 is T.S.B (Total serum Bilirubin test), x_4 is HBA1c (Hemoglobin A1), x_5 is ALK (Alkaline phosphatase), x_6 is G.P.T (Glutamic pyruvic transaminase), x_7 is G.O.T (Glutamic oxaloacetic transaminase), x_8 is CHOl (Cholesterol), x_9 is T.G (Triglycerides test), x_{10} is U.ACID (Uric Acid), x_{11} is WBC (White blood cell), x_{12} is PCV (Packed cell Volume), x_{13} is HB (Hemoglobin), x_{14} is ESR (Erythrocyte Sedimentation Rate), x_{15} is S.Na (Serum Sodium), x_{16} is S.Ca (Serum Calcium), x_{17} is PLT (Platelet Count Test), x_{18} is Iron, x_{19} is S.K (Serum Potassium Levels) and x_{20} patient's age. To achieve the study objectives, we analyzed the data set by adding some outliers in x and y. Four cases are considered, no outlier and a percentage of 5%, 10% and 15% contaminated observations. To evaluate the estimation accuracy for mentioned methods, we conducted a comparison based on the mean squared error (MSE), residual standard error (RSE) and prediction error for real data. Also, we reported the number of selected variables by SMAVE-EN, RSMAVE and RSMAVE-EN.

Table1. Results of the comparison of estimation accuracy based on MSE and RSE for SMAVE-EN, RSMAVE and RSMAVE-EN.

Outliers	Method	MSE	RSE
No outlier	SMAVE-EN	0.9792	1.0040
	RSMAVE	0.6305	0.8056
	RSMAVE-EN	0.6147	0.7951
5%	SMAVE-EN	1.4583	1.2520
	RSMAVE	0.8937	0.9645
	RSMAVE-EN	0.8035	0.914
10%	SMAVE-EN	1.900	1.399
	RSMAVE	0.9596	0.9747
	RSMAVE-EN	0.8053	0.9105

15%	SMAVE-EN	2.0131	1.4400
	RSMAVE	1.0059	1.0186
	RSMAVE-EN	0.8482	0.9391

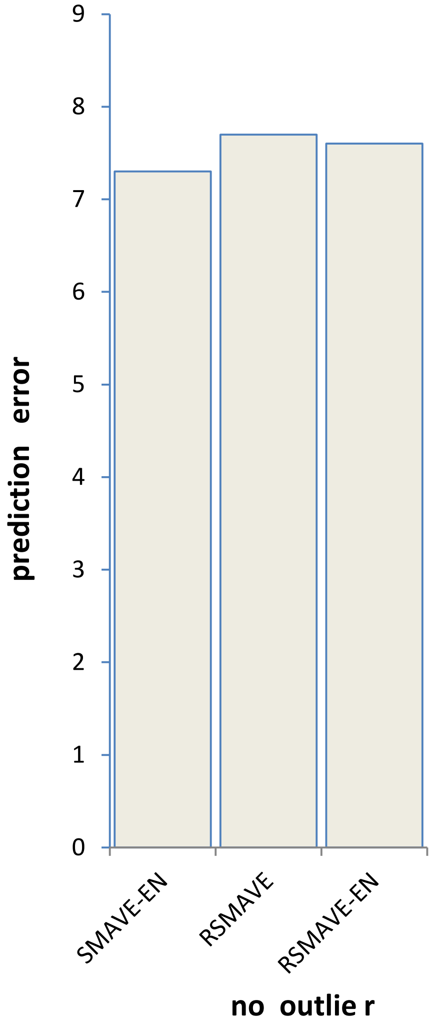
Table2. Comparison of the three methods based on prediction error

outliers	methods		
	SMAVE-EN	RSMAVE	RSMAVE-EN
No outlier	7.2686	7.6669	7.6249
5%	16.3609	9.7581	9.4074
10%	25.4998	16.0452	13.8306
15%	34.4269	20.1670	18.0690

Table3. Comparison of variable selection for the three methods based on number of selected variables.

outliers	methods		
	SMAVE-EN	RSMAVE	RSMAVE-EN
No outlier	8	12	12
5%	14	10	10
10%	12	11	10
15%	13	11	10

From outcomes of table1,2 and 3 for the previous three methods, the comparison demonstrated that, the three reported methods yielded similar results in case of standard normal distribution in estimation accuracy. Whereas in the three cases of data contamination, we can note that SMAVE-EN method was sensitive to contamination but rest 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. Depend on the above observations it is clear that under various settings


9
8
7
6
5
4
3
2
1
0
prediction error
SMAVE-EN
RSMAVE
RSMAVE-EN
no outlie r

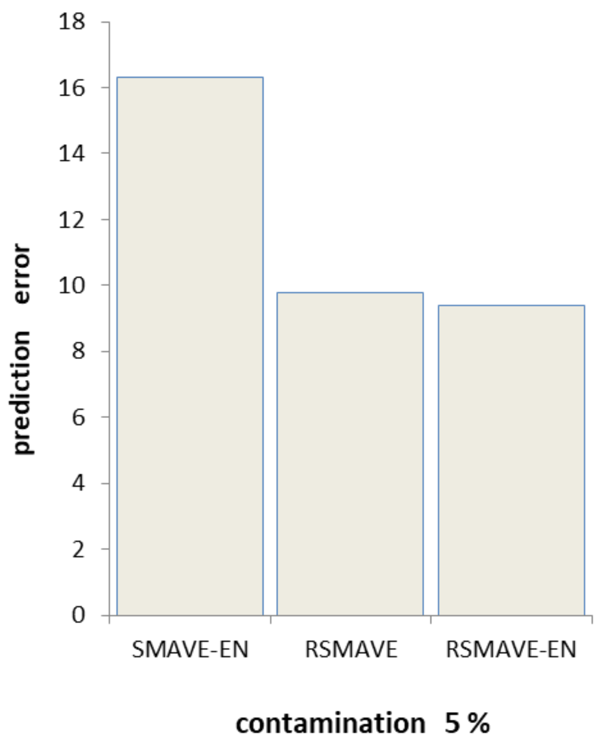


Figure1. Prediction error of the compared methods for the both contamination cases no outlier and 5% contamination based on the diabetes data.

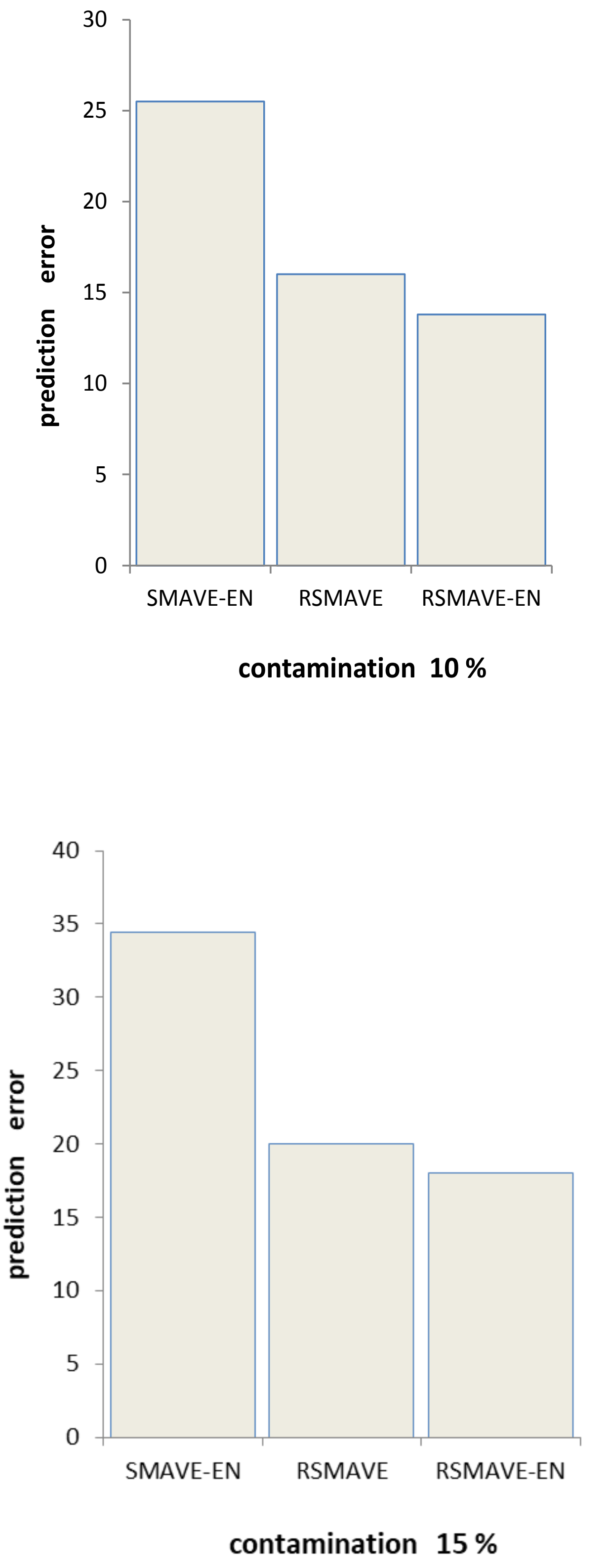
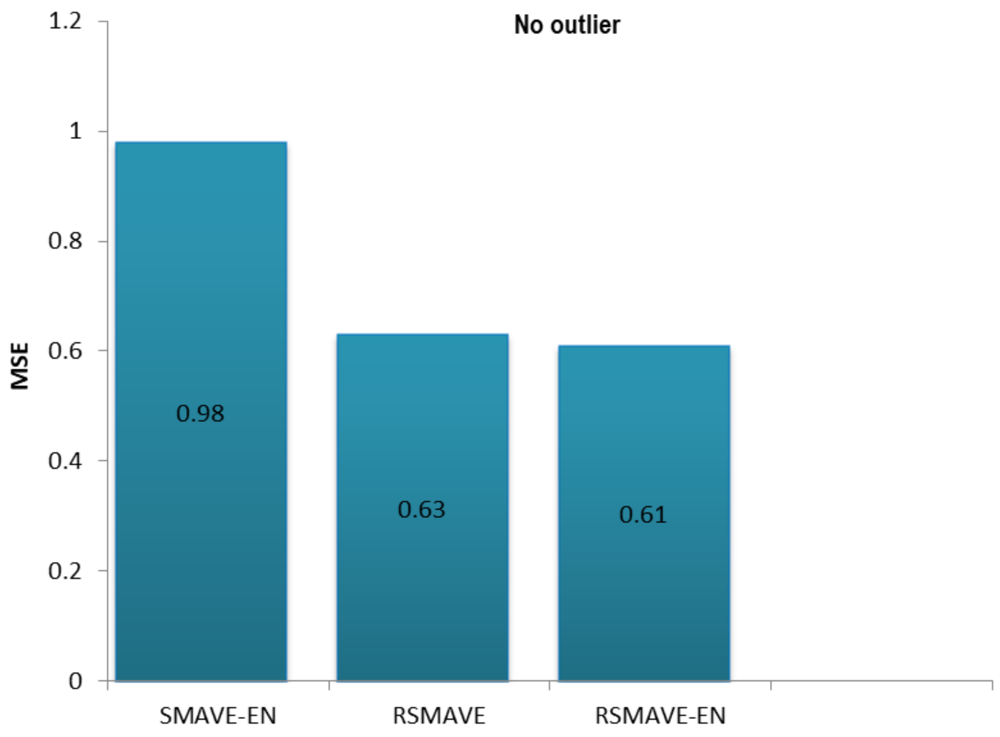


Figure2. Prediction error of the compared methods for the both contamination cases 10% contamination and 15% contamination based on the diabetes data.

.



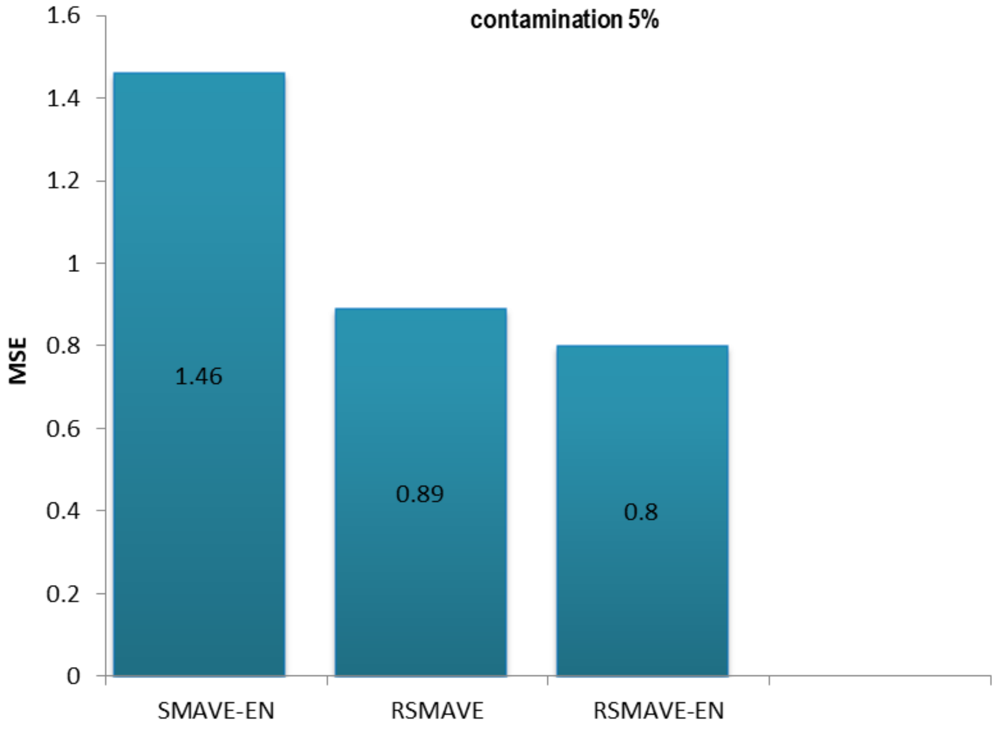
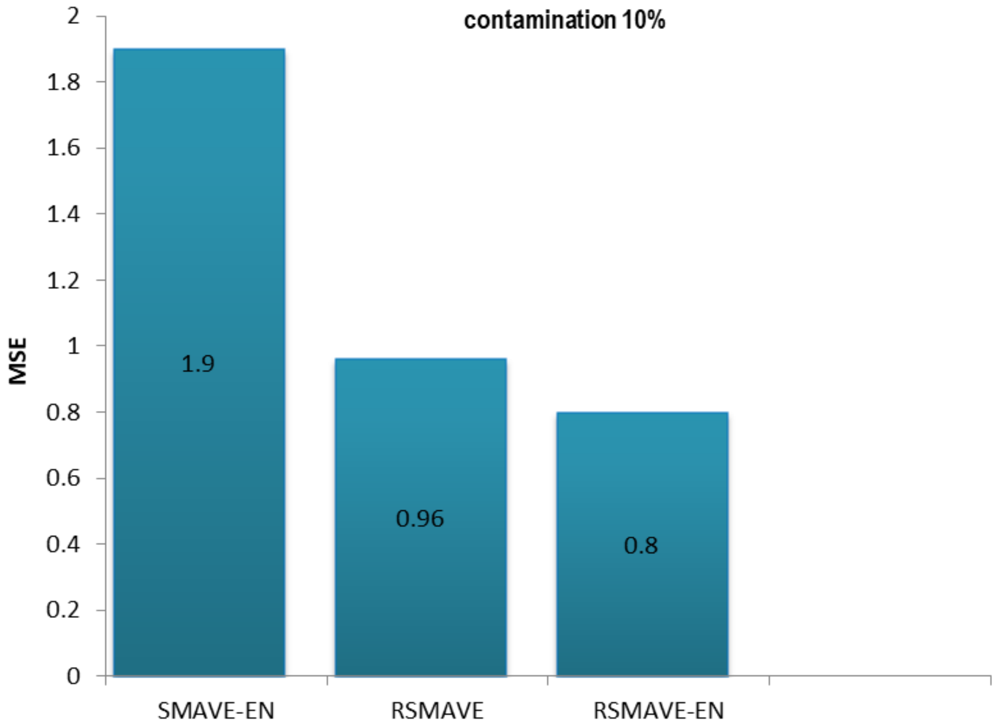


Figure3. Mean square error of the compared methods for the both contamination cases no outlier and 5% contamination based on the diabetes data.

ISSN$_{Online}$: 2312-9883
ISSN$_{Print}$: 1816-9171



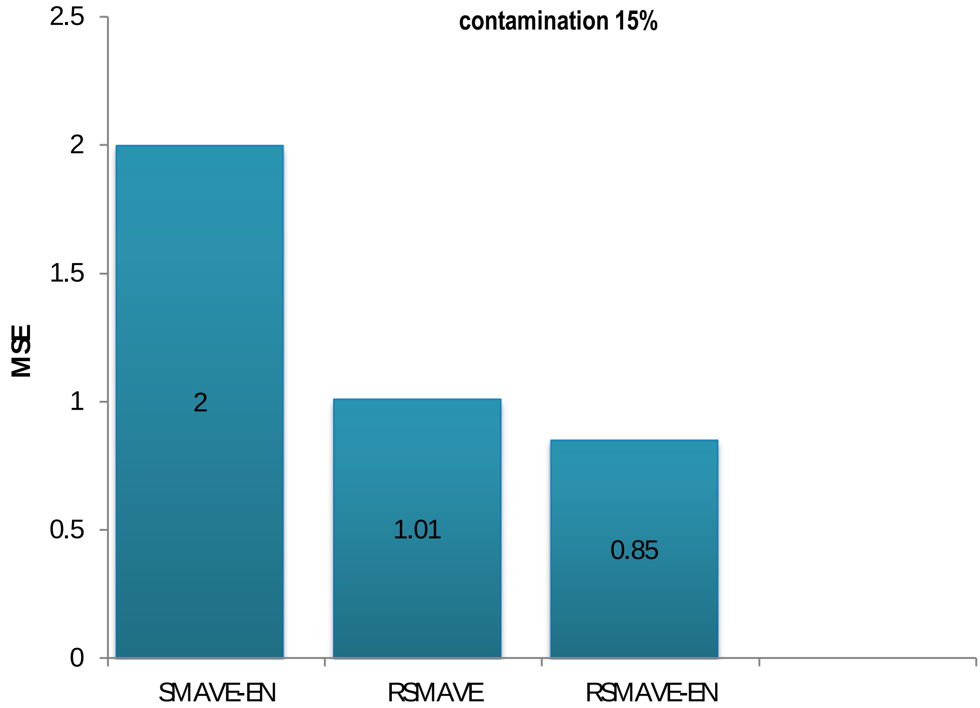


Figure4. Mean square error of the compared methods for the both contamination cases 10% contamination and 15% contamination based on the diabetes data.

6. Conclusion

In this paper, we have been presented the SMAVE-EN, RSMAVE and RSMAVE-EN. Also, we used these methods in analysis diabetic data and the factors affecting it. The outcomes of numerical study for real data analysis have shown that the RSMAVE-EN has more effective in a variable selection and estimation accuracy even with the outliers exist in predictors x and response variable y. Thus, the RSMAVE-EN outperformed the competitors SMAVE-EN and RSMAVE for various cases. Therefore, we recommend using the RSMAVE-EN method in analysis the data set especially when there are outliers in the data set.

References

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

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

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

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

[5] 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.

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

[7] Xia, Y., Tong, H., Li, W. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society Series B 64, 363–410.

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

[9] 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.

[10] 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.

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

[12] Zou, H., and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. Annals of statistics, 37(4), 1733.

[13] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics 38, 894–942.

[14] 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.

[15] Wang, T., Xu, P. and Zhu, L. (2013). Penalized minimum average variance estimation. Statist. Sinica 23 543–569.

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

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

[18] Wang, Q. Yao, W. (2013). Robust variable selection through MAVE. Computational Statistics and Data Analysis 63,42-49.

[19] Alkenani, A. and Aljobori, N. (2021). Robust sparse MAVE through elastic net penalty. International journal of Agricultural and Statistical Sciences. Submitted.

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

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

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