

Republic of Iraq
Ministry of Higher Education
And Scientific Research
University of Al-Qadisiyah
College of Administration and Economics
Statistics Department



Robust Estimation and Variable Selection through Sparse MAVE with Applications

A thesis submitted to
The Council of the College of Administration
and Economics at University of Al-Qadisiyah
as partial Fulfillment of the requirements For
the M.S.C in Statistics

By
Naeem Abed Otaiwi Aljobori

Supervised by
Prof. Dr. Ali J. Kadhim Alkenani

2022 A.D.

1443 A.H.

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

وَالرَّاسِخُونَ فِي الْعِلْمِ يَقُولُونَ آمَنَّا بِهِ كُلٌّ مِنْ عِنْدِ رَبِّنَا

وَمَا يَذَّكَّرُ إِلَّا أُولُو الْأَلْبَابِ ﴿٧﴾

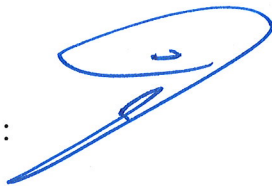
صَدَقَ اللَّهُ الْعَظِيمُ

سورة آل عمران آية (٧)

Supervisor's recommendation

I certify that the thesis entitled (Robust Estimation and Variable Selection through Sparse MAVE with Applications) has been under my supervision for the student (Naeem Abed Otaiwi Aljobori) in the Department of Statistics / College of Administration & Economics / University of Al-Qadisiyah as it is part of the requirements for master's degree in statistics sciences.

Signature :



Supervisor : Prof. Dr. Ali J. Kadhim Alkenani

Date :

Recommendation of the head of the Graduate Studies Committee

Based on the available recommendation , I would like to forward this thesis for discussion.

Signature :



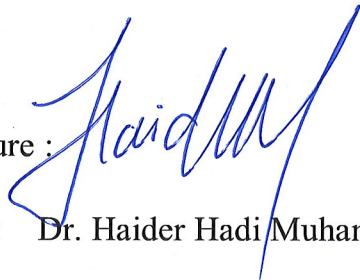
Name : Assist. Prof. Dr. Muhannad F. Al-Saadony

Chairman of the Higher Studies Committee in department of statistics

Approval of the linguistics expert

This is to hereby that the entitled (Robust Estimation and Variable Selection through Sparse MAVE with Applications) has been reviewed in terms of stylistics and linguistics (grammar and spelling). Therefore, after the modification of all recommended notes thesis has become free of all linguistics errors and ready to be defended and used as a scientific method to award the degree of master in statistical sciences.

Signature :



Name : Dr. Haider Hadi Muhammad

Date : 20.02.2022

(The Decision of the Committee)



University of Al-Qadisiyah

College of Administration and Economics

Statistics department

We are the heads and members of the defense committee certify that we have been looked at the thesis entitled (**Robust Estimation and Variable Selection through Sparse MAVE with Applications**) and we have debated the student (**Naeem Abed Otaiwi**). As a result , the student his defense his thesis and all its content. So that we have found that the thesis is worthy to be accepted to award a (excellence) master's degree in statistics science.

N.	Name	Scientific Title	Signature	Capacity
١	Salah Hamza Abed	Prof. Dr.		The head of the committee
٢	Muhammad F. Al-Saadony	Assist. Prof. Dr.		Member
٣	Mohammed Abed Alhussain	Assist. Prof. Dr.		Member
٤	Ali J. Kadhim Alkenani	Prof. Dr.		Member and supervisor

Approval of the College Committee

The council of the college of administration and economics at the University of Al-Qadisiyah have been approved on the decision of the defense committee.

Prof. Dr. Sawsan Kareem Hodan

The Dean of the college of administration and economics/Al-Qadisiyah University

Acknowledgements

Praise and thanks be to Allah, the Lord of the worlds, and thanks to the first teacher, the prophet of Allah peace and blessings of Allah be upon him and his honorable family.

I would like to express special thanks to my supervisor Prof. Dr. Ali Alkenani for his efforts towards the success of this work. He is really wonderful professor, who spared no efforts in my assistance.

I also gratefully acknowledge Prof. Dr. Tahir Reisan for his efforts and unlimited support to complete my thesis.

I sincerely express my thanks and gratitude to Dr. Mohammed Al-Sharoot, Dr. Muhannad Alsaadony, Dr. Rahim Alhamzawi, Dr. Ahmed Naeem, Dr. Hassan Sami Alshemary, Dr. Taha Alshaybawe, Dr. Fadil Alhusseini and Dr. Bahr Kadhim For their contribution in my teaching.

I particularly thank my brother Kareem for his endless support.

I would like to thank Qassim Noori for his assistance in collecting real data.

Finally, I owe a great dept to my father and mother, may God have mercy on them.

DEDICATION

To my supervisor Prof. Dr. Ali Alkenani.

To my respected teachers.

To my parents, may God have mercy upon them.

To those who left this world early, my brothers Rahim and his son Emad, Salem and Ghanem, may God have mercy upon them.

To my family: my brothers, my sisters, my wife, my sons and daughters.

To my dear friends.

This work is dedicated to them.

Abstract

The sufficient dimension reduction (SDR) is one of the important topics in many scientific fields. It has attracted researchers' attention because it is considered a beneficial approach to addressing the problem of the high dimension (HD). The problem of HD has been emerged due to big data in recent years. Many researchers came up with new ideas. They combined the SDR and regularization methods such as SMAVE-EN (sparse MAVE elastic net) among others. The SMAVE-EN is a model-free variable selection (V.S) approach. It mixes the minimum average variance estimation (MAVE) and elastic net (EN) approach. The SMAVE-EN is effective when the predictors are highly correlated under SDR settings. However, SMAVE-EN is not robust to outliers and it is sensitive to the presence of outliers in the data. In this thesis, we proposed the RSMAVE-EN. It is a robust model-free V.S approach. This approach works under different error distributions settings and gives robustness when there are outliers in both the dependent variable and the independent variables. We checked the behavior of proposed method via both simulation studies and a real data analysis.

Table of Contents

Acknowledgements	I
Dedication	II
Abstract	III

Chapter one

1.1 Introduction.....	2
1.2 Variable selection.....	4
1.3. Variable extraction.....	5
1.4. The Aim.....	6
1.5. Literature Review.....	6

Chapter two

2.1. Review for V.S methods	
Under the OLS.....	9
2.1.1. Traditional V.S methods	9
2.1.1.1. Stepwise selection	9
2.1.1.2. Forward selection	10
2.1.1.3. Backward elimination	10
2.1.1.4. Akaike information criteria (AIC).....	12
2.1.1.5. Bayesian information criteria (BIC).....	12
2.1.2. Regularization methods.....	13
2.1.2.1. Lasso	14
2.1.2.2. Adaptive Lasso	15

2.1.2.3. Elastic Net (EN)16

Chapter three

3.1. Brief Review of SDR and MAVE.....19

3.1.1. SDR19

3.1.2. MAVE20

3.2. SMAVE22

3.3. SMAVE-EN23

3.4. The proposed approach24

3.4.1. Robust SMAVE.... 24

3.4.2. Robust SMAVE-EN25

3.5 Tuning parameter selection28

Chapter four

4.1. Simulation study30

4.2. Boston Housing data44

4.3. Diabetes data.....48

Chapter five

5.1. Conclusions57

5.2. Recommendation and future work.....58

References59

List of abbreviations

Adaptive Lasso	Adaptive least absolute shrinkage and selection operator
AdEN	Adaptive elastic net
AIC	Akaike information criterion
ALMAVE	Sparse MAVE with adaptive Lasso penalty
Ave 0's	Average number of zero coefficients
BIC	Bayesian information criterion
CV	Cross validation
CD	Curse of dimensionality
CMF	Conditional mean function
CMS	Central mean subspace
DR	Dimension reduction
DRS	Dimension reduction subspace
EN	Elastic net
GR	Graphical regression
GCV	Generalized Cross validation
GSR	Graduate student rates

HD	High dimensional
IHT	Iterative Hessian Transformation
i.i.d	independent identically distributed
Lasso	Least absolute shrinkage and selection operator
LC's	Linear combinations
LS	Least squared
MAVE	Minimum average variance estimator
MCP	Minimax concave penalty
MCP-MAVE	Spars MAVE with MCP penalty
ME	Model selection
MSE	Mean squared error
OLS	Ordinary least squared
OP's	Oracle properties
OPG	Outer product of gradients
PACS	Pairwise absolute clustering and sparsity
PHd	principal Hessian directions
PE	Prediction error
P-MAVE	Penalized MAVE

RSIR	Regularized Sliced inverse regression
RSMAVE	Robust sparse MAVE
RSMAVE-EN	Robust sparse MAVE-elastic net
RSS	Residual sum squared
RSSIR	Robust Sparse sliced inverse regression
SAVE	Sliced average variance estimator
SCAD	Smoothly clipped absolute
SCAD-MAVE	Sparse MAVE with SCAD penalty
SD	Standard deviation
SDR	Sufficient dimension reduction
SIR	Sliced inverse regression
SMAVE	Sparse MAVE
SMAVE-EN	Sparse MAVE with EN penalty
SMAVE-ADEN	Sparse MAVE with ADEN
SSIQR	Sparse sliced inverse quintile regression
SSIR	Sparse sliced inverse regression
SSIR-AL	Sparse sliced inverse regression with adaptive Lasso
SSIR-EN	Sparse Sliced inverse regression with Elastic net penalty
SSIR-PACS	Sparse sliced inverse regression with PACS
V.S	Variable selection

List of Table

Table number	Title	Page number
4.1	Result of comparison for example 1 when $n=100$	33
4.2	Result of comparison for example 1 when $n=200$	35
4.3	Result of comparison for example 2 when $n=100$	37
4.4	Result of comparison for example 2 when $n=200$	39
4.5	Result of comparison for example 3 when $n=100$	41
4.6	Result of comparison for real data B.H based on NSV, MSE, RSE and $\text{adj. } R^2$	45
4.7	Result of comparison for diabetes data based on MSE and RSE	51
4.8	Result of comparison for diabetes data based on prediction error (Pe)	52
4.9	Result of comparison for diabetes data based on NSV	52

List of Figures

Figure number	Title	Page number
4.1	MSE for example 1 of four distributions when $n=100$	34
4.2	MSE for example 1 of four distributions when $n=200$	36
4.3	MSE for example 2 of four distributions when $n=100$	38
4.4	MSE for example 2 of four distributions when $n=200$	40
4.5	MSE for example 3 of four distributions when $n=100$	42
4.6	MSE for Boston housing data	46
4.7	The prediction error for diabetes data	53
4.8	MSE for diabetes data	54

Chapter one

Introduction, The Aim, Variable selection,
Variable extraction and Literature review

1.1. Introduction

If the number of variables is large, analysis of regression is a difficult process. In other word, increase the number of variables increase the complexity of regression models in analyzing the data set. This problem led researchers to work to reduce the high dimensions of data. In addition, due to “Curse of dimensionality” (Bellman, 1961) it is complex to formulate a parametric model for a large number of covariates. Thus, many approaches in statistical analysis do not work well. The sufficient dimension reduction (SDR) (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 $x = (x_1, x_2, \dots, x_p)^T$ 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 $\perp\!\!\!\perp$ 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 sliced inverse regression (SIR) (Li, 1991). 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 dimension reduction were introduced, for example the 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.

Under SDR framework settings, many procedures are proposed to combine the ideas of SDR methods and regularization methods by many researchers. For example Alkenani and Rahman (2020) proposed SMAVE-EN method. The authors combine a popular SDR method, MAVE (Xia et al., 2002) and Elastic Net (EN) penalty to produce sparse MAVE-EN. It is accurate estimates when the predictors are highly correlated under SDR settings. Moreover, variable selection and parameters estimation have been implemented simultaneously. In addition to the problem of the high dimension (HD), there is another problem is the presence of outliers in data. However, SMAVE-EN is not robust to outliers due to the use of least squares criterion. In this thesis, we proposed 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.

In general the variable selection is divided into two types of methods, traditional and regularization methods.

1.2. Variable selection (V.S)

The V.S is a technique to identify the best subset among all possible subsets of predictors to include in a regression model. The task of selecting effective predictors among a larger set of all potential predictors is very important in constructing a regression model. Actually, unnecessary predictors make noise increase in model parameters estimates. Moreover, picking out the significant predictors can improve the prediction accuracy of the regression model. The redundant predictors should be removed. In other words, in regression analysis the smallest model that fits the data is the best model. In addition, removing the redundant predictors saves the cost and time by not measuring redundant predictors. Also, selecting the most significant small subset of predictors makes interpretation of results easier, lower cost models and gives a good understanding of dataset. A lot of V.S approaches have been proposed to achieve the mentioned goals.

These approaches are divided into two kinds of ways: traditional and regularization approaches. Traditional approaches such as stepwise selection (Efroymson, 1960), AIC (Akaike, 1970) and BIC (Schwars, 1978) among others. These methods have drawbacks such as instability, high variance, discrete shrinkage approaches and time consuming. Therefore, the outcomes of these approaches lack high prediction precision (Breiman, 1996). Many regularization methods have been suggested for V.S in the regression models. Such as Lasso (Tibshirani, 1996), EN (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006), ADEN (adaptive elastic net) (Zou and Zhang, 2009) and MCP (Yu and Zhu, 2010) among others. It is clear that these

approaches have higher stability compared to traditional approaches. Moreover, the process of V.S and parameters estimation is carried out simultaneously (Alkenani and Yu, 2013).

1.3. Variable extraction

The variable extraction aims to covert the (projection) variables into a new little number of variables. It is sharing objective of subset selection, the difference is that the outcomes should be specified in terms of all of the variables. Also, it refers to the process of finding the transformation that is projecting data from original space to the feature space. This approach is trying to enable data visualization through minimizing the p-dimensional predictor vector x without losing much of information. Many variable extraction methods have been introduced to reduce dimensionality without losing much information. They include factor analysis (Gorsuch, 1983), principal component analysis (PAC) (Jolliffe, 2002; Zhang and Olive, 2009), sliced inverse regression (SIR) (Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), principal Hessian directions (PHd) (Li, 1992), MAVE and Outer product of Gradients (OPG) (Xia et al., 2002, see also Xia, 2007, 2008) among others.

1.4. The Aim

The first objective of our study is to develop the SMAVE-EN method which suffers greatly from the existence of outliers in the data. Therefore, a model-free V.S and a robust approach RSMAVE-EN has been proposed to deal with outliers in predictors and response variable. Whereas the proposed approach can maintain its efficient properties.

The second objective of this study is to analyze diabetes data and choose the appropriate approach in data analysis for adopting it even with the data containing outliers.

1.5. Literature Review

Under SDR settings, several procedures have been suggested when the ideas of SDR approaches are combined with regularization approaches by a number of researchers. For example, Xia et al. (2002) introduced MAVE. Ni et al. (2005) introduced a shrinkage SIR, Li and Nachtsheim (2006) introduced sparse SIR approach, Li (2007) introduced SSIR. Whereas, Wang and Yin (2008) added the Lasso penalty into MAVE to produce SMAVE, Alkenani and Yu (2013) extended the SMAVE by mixing the MAVE approach with SCAD, adaptive Lasso and MCP to propose SCAD-MAVE, ALMAVE and MCP-MAVE respectively, Wang et al. (2013) introduced P-MAVE, Alkenani and Reisan (2016) suggested the SSIQR, Alkenani and Malik (2019) suggested QMAVE and LQMAVE which combines sparse QMAVE and the Lasso penalty, Alkenani and Rahman (2020) proposed

SMAVE-EN that combines the sparse MAVE and EN penalty, also the authors suggested SMAVE-ADEN which combines the SMAVE with adaptive EN. Alkenani and Salman (2021) suggested SSIR-AL, the authors combined adaptive Lasso with the SIR approach. Alkenani and Abdulkadhim (2020) suggested SSIR-EN when the authors combined EN and the SIR approach. However, the previous approaches are not robust and sensitive to the presence of outliers in variables. Therefore, several robust studies introduced such as, Cizek and Hardle (2006) introduced robust MAVE when the authors replaced local LS by local L- or M-estimation, Wang and Yao (2013) suggested a robust SMAVE (RSMAVE), Alkenani (2020) proposed RSSIR which is a robust V.S in SIR, also, Alkenani (2021) proposed robust SSIR-PACS (RSSIR-PACS).

The contribution in this thesis is a robust model-free V.S approach. It is called robust sparse MAVE elastic net (RSMAVE-EN). We replacing the least squares loss function in SMAVE-EN by robust loss function with Tukey's biweight function. It is an effective approach when the predictors are highly correlated under SDR settings. Furthermore, this approach works under different error distributions settings. We arranged the rest of the thesis as follows: In Chapter 2 review of some methods for V.S under the OLS settings. In Chapter 3, a brief review of SDR and MAVE. Also, SMAVE-EN was presented, the robust approach (RSMAVE-EN) has been proposed. In Chapter 4, the effectiveness of the suggested approach is verified through simulation studies and real data analysis. In Chapter 5, conclusions and future work were reported.

Chapter two

Review for V.S methods under the OLS:
Traditional V.S and Regularization methods

2.1. Review for V.S methods Under the OLS

The V.S technique is important in building multiple regression models. It helps to improve prediction accuracy and makes the interpretation easy. Moreover, it provides a low cost model (Guyon and Elisseeff, 2003).

2.1.1. Traditional V.S methods

A number of traditional approaches for V.S have been strengthened in the literature such as stepwise selection, AIC and BIC among others. These approaches have drawbacks such as instability, discrete procedures and high variance (Breiman, 1996).

2.1.1.1. Stepwise selection

It can be considered an evaluation of the forward selection way. It was established by (Efroymson, 1960) to improve its efficiency. The point of distinction between both ways is that all the independent variables at the end of every step are checked based on the choice of (F_{partial}), we re-assessed again because there are a strong relationships between the independent variables which introduced in the previous steps. Thus, this procedure was considered a good approach to choosing the best regression equation.

2.1.1.2. Forward selection

This procedure starts without independent variables in the model then adds independent variables are selected to the equation one by one. The most significant variable is added first based on the comparison (F_{partial}) for each variable with value (F_{tabular}). The largest value is chosen (F_{tabular}) for each step and after checking that value is larger than (F_{tabular}) the variable in question is inserted into the equation. This process continues to show the variables one by one until getting to the top (F_{partial}) lower than value (F_{tabular}) according to the following:

$$F^* = \frac{SSR(x_1)}{\frac{SSE(x_1)}{n-2}}, \quad (2.1)$$

where SSR is the deviations shown,

SSE is the unclarified deviations,

n is sample size.

2.1.1.3. Backward elimination procedure

This procedure is the reverse of the forward selection approach. It is considered one of the simplest methods of V.S, starting with a full model that considers all of the independent variables to be included in the equation. Variables then are deleted one after the other from the full model relying on

the value (F_{tabular}) till only the significant independent variables remain. The process is performed as follows:

1. Start with all variables in the regression model then calculate the values (F_{partial}) for each variable depending on the formula:

$$F_{i \text{ partial}} = \frac{SSR\left[\frac{x_i}{\text{all other explanatory variables}}\right]}{\frac{SSE(x_1, \dots, x_k)}{n-k-1}}, \quad (2.2)$$

And select the variable that has the least value (F_{partial}) then compare with (F_{tabular}). When ($F_{\text{partial}} < F_{\text{tabular}}$) the relevant variable is eliminated from equation and moved to the degree freedom of the numerator (1) and the denominator (n-k-1).

2. The variables excluding that eliminated in step 1 are included. (F_{partial}) for all the remaining variables from step 1, the least are selected and compared with (F_{tabular}) to d.f (1) for the numerator and (n-k-2) for denominator. If ($F_{\text{partial}} < F_{\text{tabular}}$), eliminates the variable and goes to next step. The process continue until get least value ($F_{\text{partial}} > F_{\text{tabular}}$) the process stop.

2.1.1.4. AIC (Akaike Information Criteria)

AIC has been suggested by (Akaike, 1973) which is a measure of relative quality of models. It is used to compare the quality of the models and determine which one of them is the most appropriate model for the data. This procedure estimates the quality of each model, relative to each of a rest models. Thus, it provides a tool for model selection. The best model according to AIC is the model with the lowest AIC and is illustrated as follows:

$$\text{AIC (K)} = -2Ln(L) + 2K, \quad (2.3)$$

where L is value of MLE of the model,

K is number of model parameters.

2.1.1.5. BIC (Bayesian Information Criteria)

The BIC has been suggested by (Schwarz, 1978) which is one of the traditional V.S approaches. It is a criterion for selecting a model from specific group of models. Actually, the BIC is similar to the AIC criteria, but the difference between them is that BIC includes the sample size which gives BIC the advantage over AIC (Carlos and Sergioc, 2012).

The best model according to BIC is the model with the lowest BIC value and it is illustrated as follows:

$$\text{BIC}(k) = -2\ln(L) + k \ln(n), \quad (2.4)$$

2.1.2. Regularization methods

The second type of the V.S methods are the regularization approaches. These approaches contribute significantly to solve the complexity of regression models. It can help to avoid the problem of model over fitting. The fitting means that the balance of variance and bias in the model. The model which has a high complexity tends to a large variance but low bias. Whereas the model that has low complexity tends to low variance but large bias. In other word, these approaches make complex regression models less complicated to eliminate the problem of over fitting. In order to control the complexity of the model, the regularization ways add the penalty term to the standard loss function, such as loss function of OLS. Donoho and Johnston (1994) introduced the first use of the regularization ways for V.S. In these ways the V.S is performed via the conduct parameter estimation (Wang and Yin, 2008). Examples of regularizations approaches are Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), EN (Zou and Hastie, 2005),

OSCAR (Bondell and Reich, 2008), MCP (Zhang, 2010) and PACS (Sharma et al., 2013) among others.

2.1.2.1. Lasso

The Lasso (least absolute shrinkage and selection operator) has been proposed by (Tibshirani, 1996) which is considered a beneficial approach for a simultaneous estimate of parameters and V.S. It is an effective and powerful approach to remedy HD data. In this approach RSS is minimized subject to $\|\theta_k\|_1$ being less than a constant. Therefore, the Lasso approach shrinks some of the coefficients and makes the other equal to zero. This estimator is obtained by adding a penalty function to the least squares loss as in the following equation:

$$\hat{\theta} (Lasso) = \arg \min \sum_{i=1}^n (y_i - \theta^T x_i)^2 + \lambda \sum_{k=1}^p |\theta_k|, \quad (2.5)$$

$\lambda > 0$ represents the tuning parameter, the largest value of λ provides higher shrinkage level (Alkenani and Yu, 2013), the value of λ is determined through the Generalized Cross Validation (GCV) as follows:

$$GCV = \frac{RSS}{n\{1 - p(\lambda)/n\}^2},$$

where $RSS = \sum_{i=1}^n (y_i - \theta^T x_i)^2$,

$P(\lambda)$ the effective number of parameters, larger value of $P(\lambda)$ cause more inflation (penalization)

$\sum_{k=1}^p |\theta_k| = \|\theta\|_1$ is the l_1 norm of θ , $k = 1, \dots, p$, p is the number of variables, $i = 1, \dots, n$ and n is sample size.

2.1.2.2. Adaptive Lasso

Although the Lasso is a popular approach for simultaneous V.S and estimation of parameters, whereas, for a large coefficients its estimates are biased and do not have the oracle property (OP's) because of bias problem (Fan and Li, 2001). Zou (2006) introduced a new version of the Lasso which is called (adaptive Lasso). The author stated that the (adaptive Lasso) approach has advantages over Lasso and the (OP's) for this approach is achieved. Also, the author demonstrated that the estimates of Lasso are biased because that, in Lasso approach all coefficients are subject to same constraint. Thus, the estimates are inconsistent. Whereas, in adaptive Lasso approach, the coefficients of different predictors can be impose different weights in penalty function. Hence, the bias of the estimates can be reduced when we are able to select the weights such that the predictors with large coefficients has smaller a weights. While keeping its sparsity property, adaptive Lasso estimates are defined as follows:

$$\hat{\theta} \text{ (adaptive Lasso)} = \arg \min \sum_{i=1}^n (y_i - \theta^T x_i)^2 + \lambda \sum_{k=1}^p \tilde{w}_k |\theta_k| \quad (2.6)$$

where $\lambda > 0$ represents the tuning parameter,

the weights vector is $\tilde{w}_k = \frac{1}{|\tilde{\theta}_k|^\delta}$, $k = 1, \dots, p$,

P is number of variables, $\tilde{\theta}$ is a non-penalized regression estimate, $\delta > 0$,

δ is the contraction parameter.

2.1.2.3. EN (Elastic Net)

The EN is a regression approach suggested by (Zou and Hastie, 2005). EN can be achieved when the penalty term of Lasso and Ridge regression are combined. The researchers pointed out some flaws in Lasso's work in some cases, such as:

1. If the number of variables is greater than sample size, i.e. $p > n$ then the Lasso chooses at most n variables.
2. In the event that there is a set of strongly related variables, Lasso chooses only one from the set.

The Ridge term reduces the coefficients of correlated predictors toward each other, while the Lasso term chooses one among the correlated predictors. Thus, the researchers have been shown that the EN approach outperforms Lasso.

The EN estimates are as follows:

$$\hat{\theta} (EN) = \arg \min \sum_{i=1}^n (y_i - \theta^T x_i)^2 + \lambda_1 \sum_{k=1}^p \theta_k^2 + \lambda_2 \sum_{k=1}^p |\theta_k|, \quad (2.7)$$

where $\lambda_1, \lambda_2 \geq 0$ represented the tuning parameters,

θ_k^2 , $|\theta_k|$ are the l_2 norm related with Ridge penalty and l_1 norm related with Lasso, respectively, $k = 1, \dots, p$, p is the number of variables.

Chapter three

Brief Review of SDR and MAVE, SMAVE,
SMAVE-EN and The proposed approach

3.1. Brief Review of SDR and MAVE

3.1.1. Sufficient dimension reduction (SDR)

Let the regression-type model of a response variable $y \in \mathbb{R}^1$ on a $p \times 1$ predictor vector \mathbf{x} and the error term ε and assume the following model:

$$y = f(x_1, x_2, \dots, x_p) + \varepsilon, \quad (3.1)$$

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

$$y \perp\!\!\!\perp E(y|\mathbf{x}) |_{\mathcal{P}_S \mathbf{x}}, \quad (3.2)$$

where $\perp\!\!\!\perp$ denotes independence and $\mathcal{P}(\cdot)$ is a projection operator. Subspaces satisfying condition (3.2) are called mean DRS (Cook and Li, 2002). If $d = \dim(S)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_d)$ is a basis for S , the predictor \mathbf{x} can be replaced by the linear combinations $\boldsymbol{\theta}_1^T \mathbf{X}, \boldsymbol{\theta}_2^T \mathbf{X}, \dots, \boldsymbol{\theta}_d^T \mathbf{X}$, $d \leq p$ without losing of information on the CMF. That is, $f(x_1, x_2, \dots, x_p) = f(\boldsymbol{\theta}^T \mathbf{x})$. If the intersection of all subspaces satisfies (3.2), that is called the central mean subspace (CMS) $S_{E(y|\mathbf{x})}$ (Cook and Li, 2002). Many methods have been suggested to estimate $S_{E(y|\mathbf{x})}$ and MAVE (Xia et al., 2002) is one of these methods.

3.1.2 Minimum average variance estimation (MAVE)

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

$$\min_{\theta} \{ E[y - E(y|\theta^T x)]^2 \}, \quad (3.3)$$

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

$$\sigma_{\theta}^2(\theta^T x) = E[\{y - E(y|\theta^T x)\}^2 | \theta^T x]. \quad (3.4)$$

Thus,

$$\min_{\theta} E[y - E(y|\theta^T x)]^2 = \min_{\theta} E\{ \sigma_{\theta}^2(\theta^T x) \} \quad (3.5)$$

For any given x_0 , $\sigma_{\theta}^2(\theta^T x_0)$ can be approximated as follows:

$$\begin{aligned} \sigma_{\theta}^2(\theta^T x_0) &\approx \sum_{i=1}^n \{y_i - E(y_i|\theta^T x_i)\}^2 w_{i0} \\ &\approx \sum_{i=1}^n [y_i - \{a_0 + b_0^T \theta^T (X_i - X_0)\}]^2 w_{i0}, \end{aligned}$$

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 at $\theta^T x_0$ with $\sum_{i=1}^n w_{i0} = 1$, the choice of the weights w_{ij} plays a vital role in searching for the effective DR direction.

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

where $k_h(\cdot) = h^d k(\cdot/h)$ and d is the dimension of $k(\cdot)$, $k(\cdot)$ is the refined multidimensional Gaussian Kernel as follows (Brillinger, 1983)

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$h_{opt} = A(d) n^{-1/(4+d)}$ is the bandwidth,

$$A(d) = \left\{ \frac{4}{(d+2)} \right\}^{1/(4+d)}$$

where h is the smoothing parameter, called the bandwidth which controls the smoothness, bias and variability of the estimate (Xia et al., 2002) so the problem of finding θ is by solving the following:

$$\min_{\theta: \theta^T \theta = I_d} \left(\sum_{j=1}^n \sum_{i=1}^n [y_i - \{a_j + b_j^T \theta^T (x_i - x_j)\}]^2 w_{ij} \right) \quad (3.6)$$

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 (3.6) resolves iteratively with respect to $\{ (a_j, b_j), j=1, \dots, n \}$, and θ separately.

3.2 Sparse MAVE (SMAVE)

Although MAVE is an efficient dimension reduction method, its outputs are still linear combinations of all original predictors. Therefore, it suffers as the same difficulty in interpretation as other DR methods do. Wang and Yin (2008) combine a variable selection method Lasso (Tibshirani, 1996) with MAVE to propose sparse MAVE (SMAVE). The authors incorporate an L_1 penalty term into the MAVE loss function in (3.6). SMAVE has advantages over Lasso because it extends Lasso to multidimensional and nonlinear settings without assuming any particular model. The SMAVE minimizes:

$$\min \left(\sum_{j=1}^n \sum_{i=1}^n [y_i - \{a_j + b_j^T \theta^T (x_i - x_j)\}]^2 w_{ij} + \lambda \sum_{k=1}^p |\theta_{m,k}| \right), \quad (3.7)$$

where $m = 1, \dots, d$, and d is known and it can be estimated by BIC, $\|\cdot\|_1$ is the L_1 norm and λ is nonnegative regularization parameter which controls the amount of shrinkage.

3.3. SMAVE-EN

The previous approach Alkenani and Rahman (2020) proposed (SMAVE-EN) approach. 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:

$$\left(\sum_{j=1}^n \sum_{i=1}^n [y_i - \{a_j + b_j^T \theta^T (x_i - x_j)\}]^2 w_{ij} \right) + \lambda_1 \|\theta_m\|_2^2 + \lambda_2 \|\theta_m\|_1, \quad (3.8)$$

where $\|\cdot\|_2^2$ is L_2 norm related with ridge penalty and $\|\cdot\|_1$ is L_1 norm related with Lasso penalty. λ_1 and λ_2 are the tuning parameters. That 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.

3.4. The proposed approach

3.4.1. Robust SMAVE

Although SMAVE-EN method has advantages over the existing methods, However, SMAVE-EN is not robust to outliers, due to use of least squares criterion. Cizek and Hardle (2006) introduced a study of the sensitivity of MAVE to outlier values and suggested 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:

$$\sum_{j=1}^n \sum_{i=1}^n p[y_i - \{ a_j + b_j^T \theta^T (x_i - x_j) \}] w_{ij} , \quad (3.9)$$

where $p(\cdot)$ represent the robust loss function. Under this setting, Wang and Yao (2013) proposed a robust SMAVE. The authors added an L_1 penalty into the expression (3.9) as follows:

$$(\sum_{j=1}^n \sum_{i=1}^n p[y_i - \{a_j + b_j^T \theta^T (x_i - x_j)\}]w_{ij}) + \sum_{k=1}^d \lambda_k |\theta_k|, \quad (3.10)$$

where $p(\cdot)$ is a robust loss function,

$|\cdot|_1$ is the L_1 norm and $\lambda_k, k = 1, 2, \dots, d$: regularization parameters.

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

3.4.2. Robust SMAVE-EN

In this section, we extend the robust estimation to V.S and proposed robust SMAVE-EN (RSMAVE-EN), which can 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. The (EN) penalty added into expression (3.9). RSMAVE-EN that proposed can be obtained by minimizing the following:

$$\sum_{j=1}^n \sum_{i=1}^n p[y_i - \{ a_j + b_j^T \theta^T (x_i - x_j) \}] w_{ij} + \lambda_1 \|\theta_m\|_2^2 + \lambda_2 \|\theta_m\|_1, \quad (3.11)$$

where $p(\cdot)$ represents a robust loss function, $\| \cdot \|_1$ is the L_1 norm, $\| \cdot \|_2^2$ is the L_2 norm and λ_1 and λ_2 are the tuning parameters. We choose $p(\cdot)$ as a Tukey's biweight function to obtain robust estimation in both independent variables and response variable, 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

not sensitive to outliers in x and y . The minimizing in (3.11) is an robust version of the minimizing in (3.8) by replacing the least squares loss function in (3.8) by robust loss function with Tukey's biweight function.

The function of Tukey's biweight is:

$$p_c(u) = \begin{cases} \left(\frac{c^2}{6}\right) \left\{1 - \left[1 - \left(\frac{u}{c}\right)^2\right]^3\right\} & \text{if } |u| \leq c \\ \frac{c^2}{6} & \text{if } |u| > c \end{cases} \quad (3.12)$$

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

The RSMAVE-EN algorithm is as follows:

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

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 $j=1, 2, \dots, n$, from the following minimization problem:

$$\min_{a_j, b_j, j=1, 2, \dots, n} \left(\sum_{j=1}^n \sum_{i=1}^n p[y_i - \{a_j + b_j^T \theta^T (x_i - x_j)\}] w_{ij} \right) \quad (3.13)$$

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

$$\begin{aligned} \min_{\theta: \theta^T \theta = I_m} \quad & \sum_{j=1}^n \sum_{i=1}^n p \left[y_i - \left\{ \hat{a}_j + \hat{b}_j^T (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{m-1}, \hat{\theta}_m)^T (x_i - x_j) \right\} \right] w_{ij} \\ & + \lambda_1 \|\theta_m\|_2^2 + \lambda_2 \|\theta_m\|_1 \end{aligned} \quad (3.14)$$

4. Replace the m th column of θ by $\hat{\theta}_{\text{mRSMAVE-EN}}$ and iterate between step 2 and 3 until convergence is attained.

5. Update θ by $(\hat{\theta}_{\text{1RSMAVE-EN}}, \hat{\theta}_{\text{2RSMAVE-EN}}, \dots, \hat{\theta}_{\text{mRSMAVE-EN}}, \hat{\theta}_0)$, 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, $h_{\text{opt}} = A(d) n^{-1/(4+d)}$ 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.

3.5 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 :

$$AIC = n \log (RSS / n) + 2 p (\lambda) \quad (3.15)$$

$$BIC = n \log (RSS / n) + \log (n) p (\lambda) \quad (3.16)$$

$$RIC = \{n - p (\lambda)\} \log (RSS / n - p (\lambda)) + p (\lambda) \{ \log (n) - 1 \} + \{ 4 / (n - p (\lambda)) \}, \quad (3.17)$$

where $p (\lambda)$ denotes the number of non-zero coefficients,

RSS is defined as follows:

$$RSS = \sum_{j=1}^n \sum_{i=1}^n [y_i - \{ \hat{a}_j + \hat{b}_j^T (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{m-1}, \hat{\theta}_m)^T (x_i - x_j) \}]^2 w_{ij}, \quad (3.18)$$

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

$$RRIC = \{n - p (\lambda)\} \log (RRSS / n - p (\lambda)) + p (\lambda) \{ \log (n) - 1 \} + \{ 4 / (n - p (\lambda)) \}, \quad (3.19)$$

$$\text{Where } RRSS = \sum_{j=1}^n \sum_{i=1}^n p [y_i - \{ a_j + b_j^T \theta^T (x_i - x_j) \}] w_{ij}. \quad (3.20)$$

Chapter four

Simulation study and Real data

4.1. Simulation study

The purpose of this section is to assess the finite sample performance of the proposed RSMAVE-EN method through simulation studies. We compare the suggested approach with the methods described in the third Chapter of this thesis, the SMAVE-EN and RSMAVE methods. The comparison has been conducted to show the behavior of RSMAVE-EN approach in terms of the prediction accuracy and V. S. To check the efficiency of the V. S technique for the proposed approach, the average number of zeros the coefficients (Ave0's) was reported. In addition, we reported the absolute correlation between the estimated predictor $\hat{\theta}^T x$ and the true one $\theta^T x$. Whereas, the prediction accuracy was evaluated by calculating the mean squared error (MSE) for each example. For the tuning parameter, we employed a robust RIC (RRIC). This robust version of RIC was proposed by Alkenani (2020) which is explained in the third Chapter. The R code for SMAVE-EN is made by Alkenani and Rahman (2020). The RSMAVE was computed using R code made by Wang and Yao (2013). While, RSMAVE-EN is computed using R code made by Alkenani and Aljobori (2021). The estimation outcomes are based on 200 data replications. Moreover, we considered the distribution of x and ε for each of the following examples is as follows:

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. (Wang and Yao, 2013)

Example 1: Consider the model: $y = 1 + 2(\theta^T x + 3) \times \log(3|\theta^T x| + 1) + \varepsilon$,
where $d = 1, p = 40, n = 100, 200$

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

where $\text{corr}(i, j) = 0.5$ for all i and j (Alkenani and Rahman, 2021).

Example 2: We adopt the same model as the previous example 1,

$$y = 1 + 2(\theta^T x + 3) \times \log(3|\theta^T x| + 1) + \varepsilon,$$

where $d = 1, p = 40, n = 100, 200$ and consider $\theta = (\underbrace{3, \dots, 3}_{15}, \underbrace{0, \dots, 0}_{25})^T,$

the predictors X are :

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

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

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

$x_i, \quad i = 16, \dots, 40$, where z follows the same distribution as x and ε .

$\text{Corr}(i, j) = 0.8$ for all i and j .

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 (Alkenani and Rahman, 2021).

Example 3: Let $d = 2$, $p = 8$ and $n = 100$. The data are generating 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 $r = 0.7$ (Alkenani and Rahman, 2021).

Table 4.1: results for example 1, based on Ave0's, MSE and the absolute of correlation between $(\theta^T x, \hat{\theta}^T x)$ when size $n = 100$, $p = 40$.

dist.	method	Ave.0's	MSE	$ \text{Corr}(\theta^T x, \hat{\theta}^T x) $
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

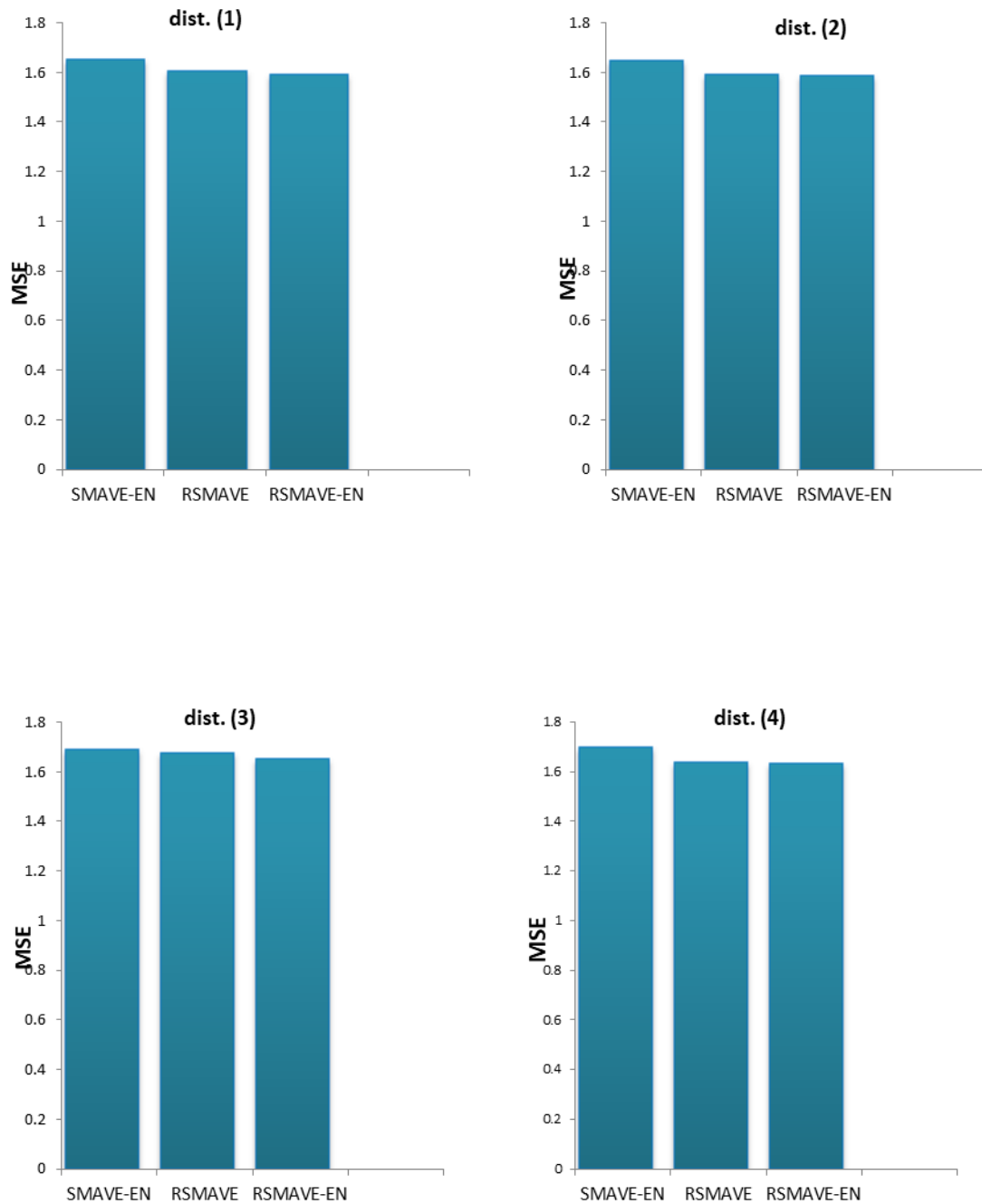


Figure 4.1: The MSE for the considered methods in example 1. when $n=100$, for the four distributions.

Table 4.2: results for example 1, based on Ave0's, MSE and the absolute of correlation between $(\theta^T x, \hat{\theta}^T x)$ when size $n = 200$, $p = 40$.

dist.	method	Ave.0's	MSE	$ \text{Corr}(\theta^T x, \hat{\theta}^T x) $
1	SMAVE-EN	12	1.649	0.995
	RSMAVE	13	1.596	0.996
	RSMAVE-EN	13	1.589	0.998
2	SMAVE-EN	10.5	1.648	0.979
	RSMAVE	12	1.589	0.999
	RSMAVE-EN	13.5	1.586	0.999
3	SMAVE-EN	11.5	1.657	0.978
	RSMAVE	12	1.637	0.988
	RSMAVE-EN	13	1.618	0.994
4	SMAVE-EN	11.5	1.697	0.981
	RSMAVE	12	1.614	0.996
	RSMAVE-EN	13	1.608	0.997

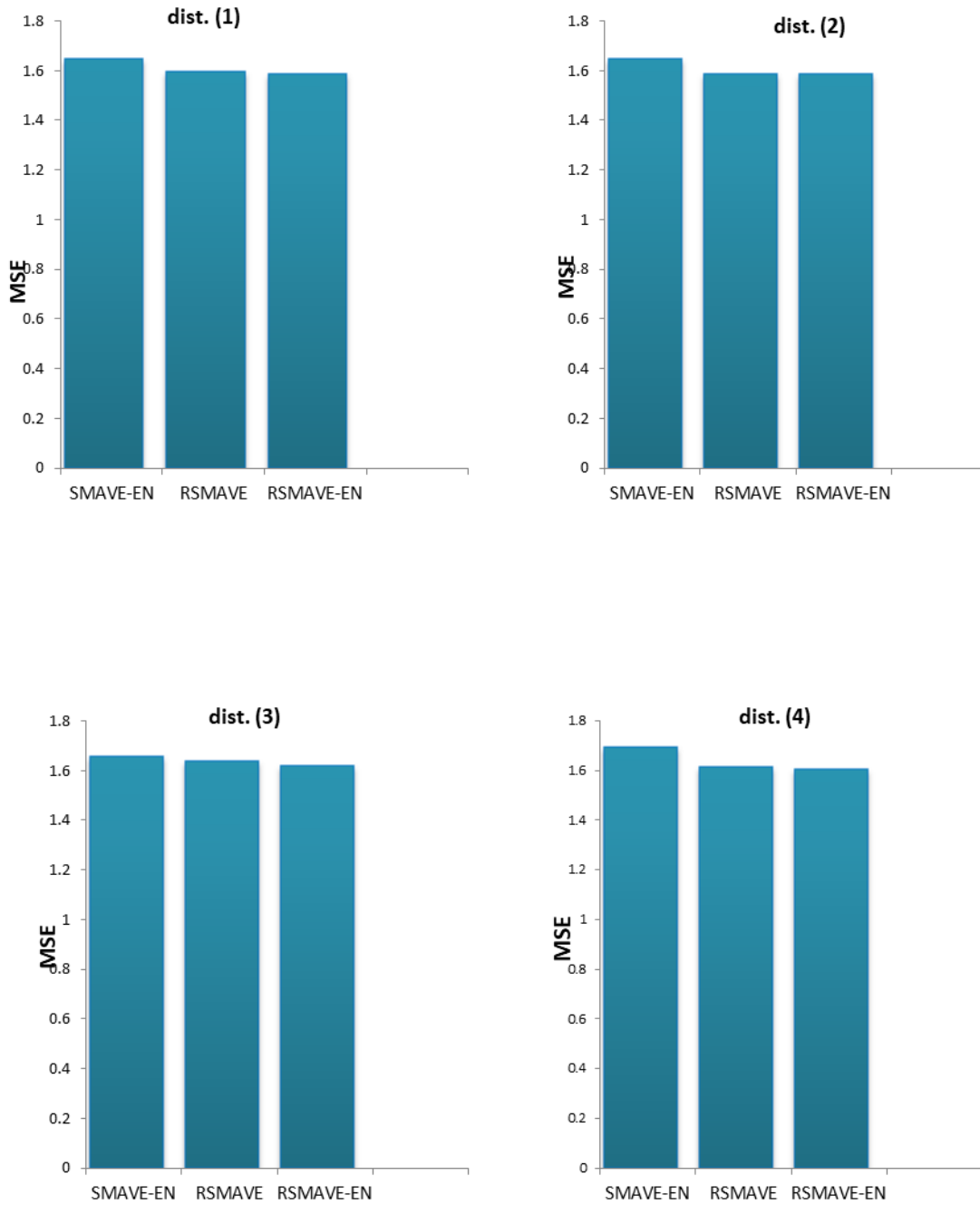


Figure 4.2: The MSE for the considered methods in example 1. when $n=200$, for the four distributions.

Table 4.3: results for example 2, based on Ave0's, MSE and the absolute of correlation between $(\theta^T x, \hat{\theta}^T x)$ when size $n = 100$, $p = 40$.

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

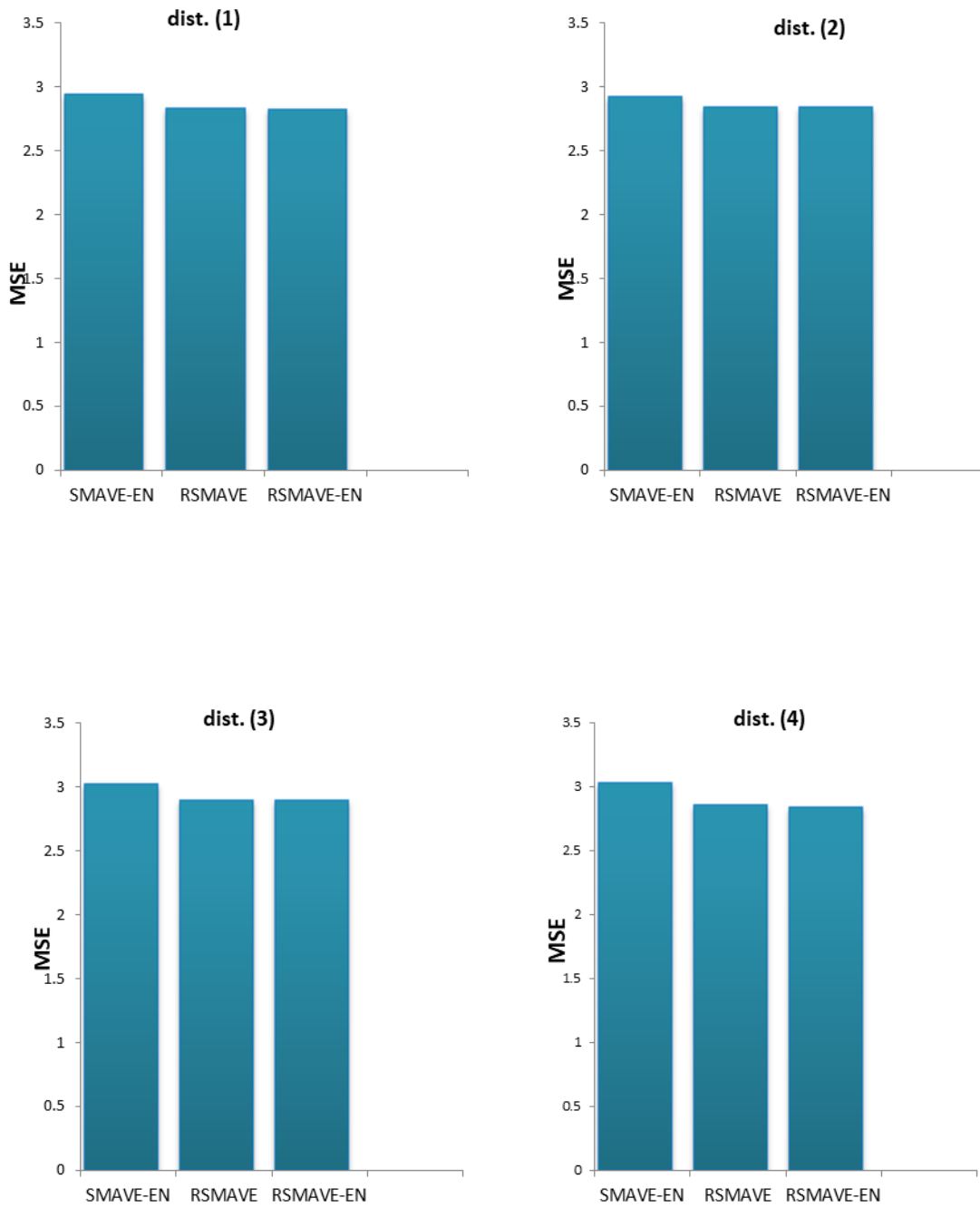


Figure 4.3: The MSE for the considered methods in example 2. when $n=100$, for the four distributions.

Table 4.4: results for example 2, based on Ave0's, MSE and the absolute of correlation between $(\theta^T x, \hat{\theta}^T x)$ when size $n = 200$, $p = 40$.

dist.	method	Ave.0's	MSE	$ \text{Corr}(\theta^T x, \hat{\theta}^T x) $
1	SMAVE-EN	14.5	2.929	0.836
	R SMAVE	11.5	2.825	0.981
	R SMAVE-EN	15	2.819	0.992
2	SMAVE-EN	14	2.904	0.845
	R SMAVE	12	2.822	0.978
	R SMAVE-EN	15	2.819	0.984
3	SMAVE-EN	13.5	3.009	0.831
	R SMAVE	12	2.883	0.951
	R SMAVE-EN	15	2.881	0.956
4	SMAVE-EN	13.5	3.021	0.742
	R SMAVE	12.5	2.846	0.961
	R SMAVE-EN	15	2.831	0.963

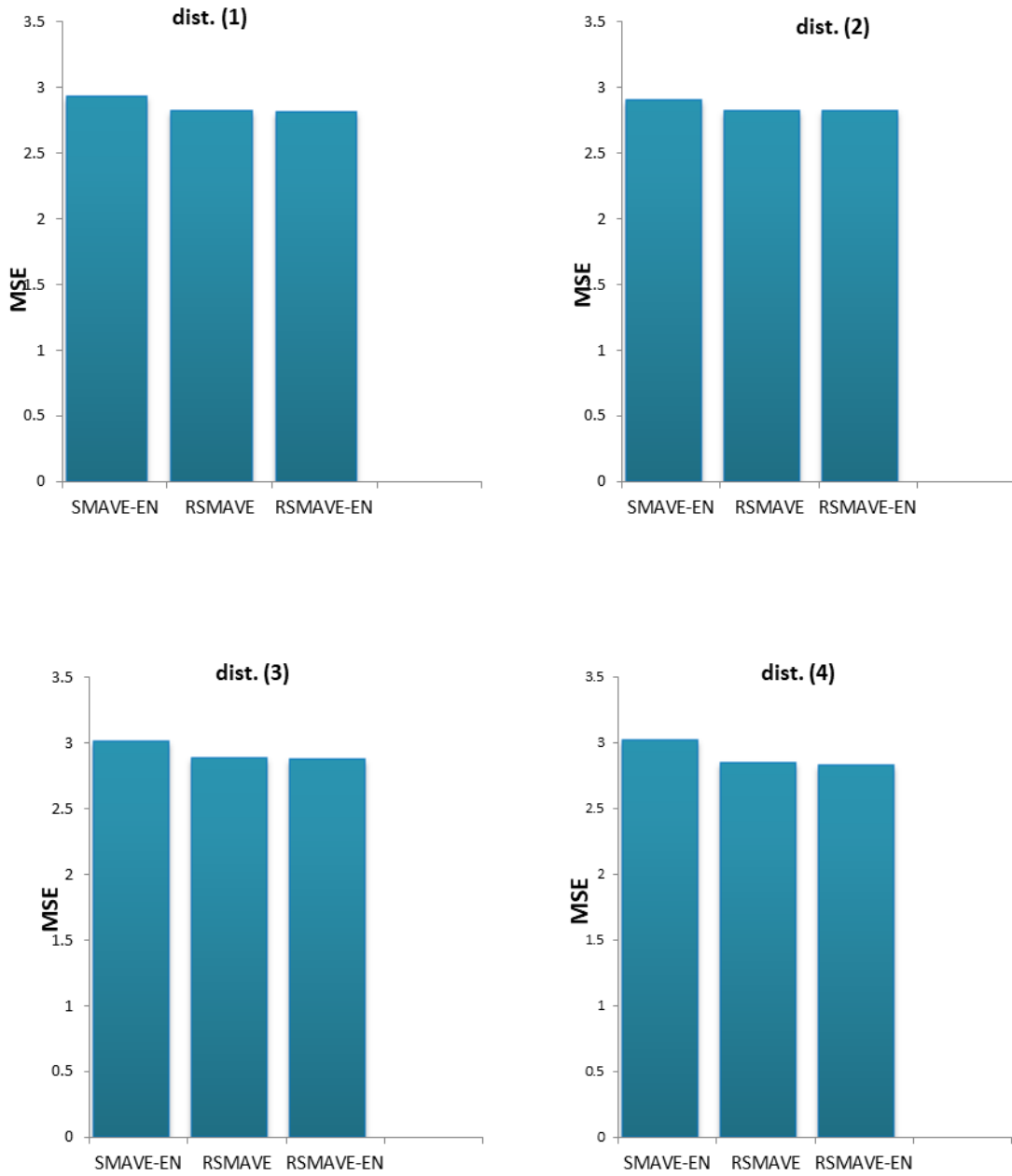


Figure 4.4: The MSE for the considered methods in example 2. when $n=200$, for the four distributions.

Table 4.5: Results of comparisons for example 3, based on the (Ave 0's), MSE, $r_1 = \text{corr}(\theta_1^T x, \hat{\theta}_1^T x)$ and $r_2 = \text{corr}(\theta_2^T x, \hat{\theta}_2^T x)$, when $n = 100$.

Dist.	method	Ave.0's	MSE	$ r_1 $	$ r_2 $
1	SMAVE- EN	8	1.642	0.848	0.389
	R SMAVE	8	1.631	0.828	0.579
	R SMAVE-EN	8	1.617	0.852	0.618
2	SMAVE- EN	7	1.727	0.814	0.190
	R SMAVE	7	1.714	0.796	0.534
	R SMAVE-EN	8	1.685	0.813	0.595
3	SMAVE- EN	6.5	1.784	0.477	0.242
	R SMAVE	7	1.643	0.811	0.533
	R SMAVE-EN	7.6	1.640	0.852	0.609
4	SMAVE- EN	6.5	1.775	0.746	0.288
	R SMAVE	7	1.723	0.761	0.362
	R SMAVE-EN	8	1.707	0.769	0.377

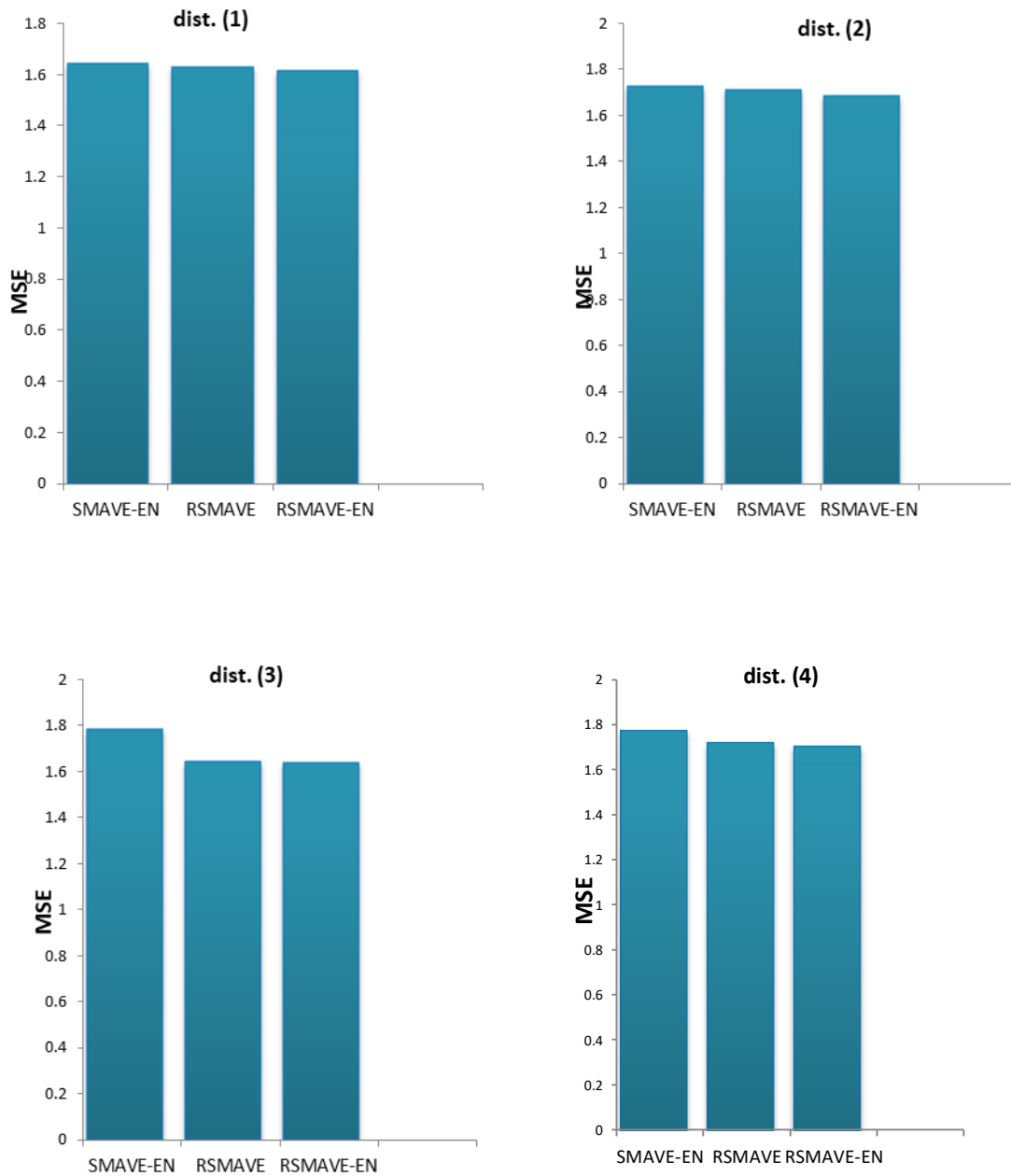


Figure 4.5: The MSE for the considered methods in example 3. when $n=100$, for the four distributions.

From outcomes of tables 4.1, 4.2, 4.3, 4.4 and 4.5 for the previous three examples, the comparison demonstrated that, the three reported methods yielded similar results in case of standard normal distribution (first distribution in tables) in both V.S and estimation accuracy. It is clear that the average number of zeros the coefficients (Ave.0's), mean squared error (MSE) and the absolute of correlation between $(\theta^T x, \hat{\theta}^T x)$ have a similar results. However, It can be seen that there is a slight outperform for the suggested approach where it has a lower MSE than the rest approaches also it has a bigger values based on Ave.0's and the absolute of $\text{Corr}(\theta^T x, \hat{\theta}^T x)$. 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, we can see that the performance of RSMAVE-EN outperformed RSMAVE method in terms of V.S and estimation accuracy. Depending on the above observations it is clear that under various settings, the proposed RSMAVE-EN has a good performance in terms of variable selection and estimation accuracy.

The Figures 4.1, 4.2, 4.3, 4.4 and 4.5 for the previous three examples, show that the MSE value for RSMAVE-EN is less than its values for RSMAVE and SMAVE-EN. This means that the suggested RSMAVE-EN has a better performance than the rest methods depending on the MSE of simulation studies.

4.2. Boston housing data

This data was collected by (Harrison and Rubinfeld, 1978), the data set includes $n = 506$ observations and $p = 14$ predictor, where y is medv (median value of owner occupied homes in \$ 1000's). X includes 13 predictors. 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, Four cases were considered in this analysis, no outliers and three a percentage of 5%, 10% and 15% contaminated observations. The data has been contaminated by replacing the predictors and response variable values by C value which equal to 100. Tables 3.7 and 3.8 explain that, to evaluate the estimation accuracy for proposed approach, we conducted a comparison based on the MSE (mean squared error), RSE (residual square error) and Adj. R^2 (Adjusted R-squared). Also, we reported the NSV (number of selected variables) by SMAVE-EN, RSMAVE and RSMAVE-EN methods.

Table 4.6: Results of comparisons for real data, based on Number of Selected Variables (NSV), MSE, RSE and Adj.R²

Contamination percentage	Methods	NSV	MSE	RSE	Adj.R ²
No outlier	SMAVE-EN	11	0.236	0.489	0.760
	RSMAVE	11	0.238	0.491	0.758
	RSMAVE-EN	11	0.235	0.488	0.760
5%	SMAVE-EN	13	0.289	0.541	0.707
	RSMAVE	12	0.242	0.495	0.755
	RSMAVE-EN	11	0.241	0.494	0.761
10%	SMAVE-EN	13	0.303	0.554	0.692
	RSMAVE	13	0.260	0.512	0.738
	RSMAVE-EN	11	0.259	0.511	0.739
15%	SMAVE-EN	13	0.318	0.567	0.678
	RSMAVE	12	0.271	0.522	0.727
	RSMAVE-EN	11	0.269	0.520	0.734

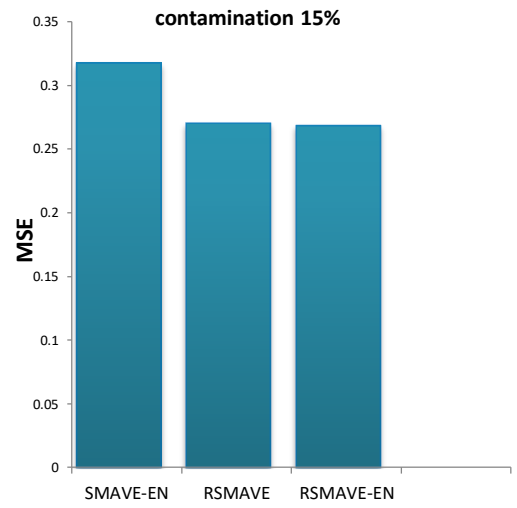
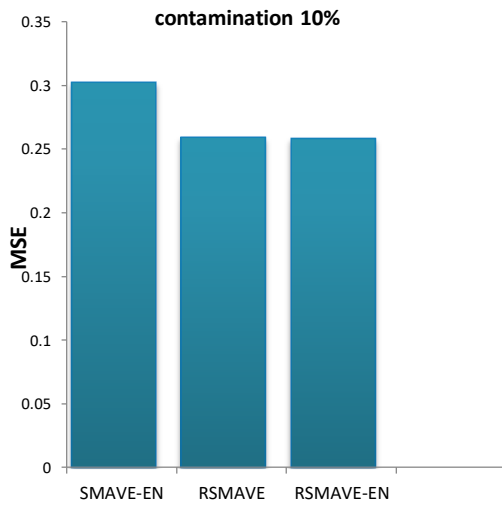
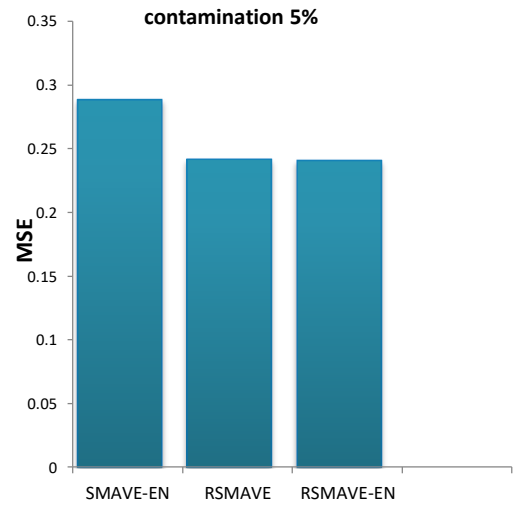
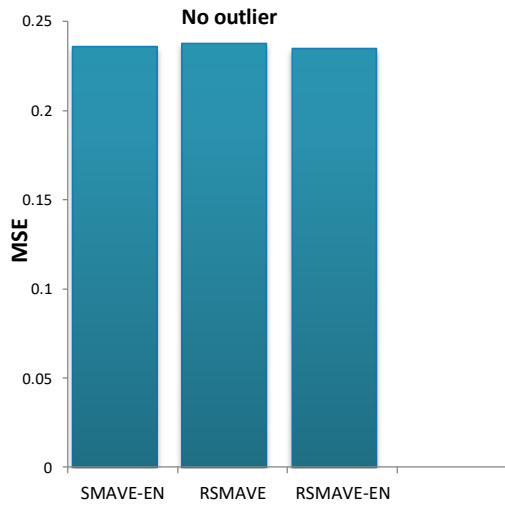


Figure 4.6: The MSE for the considered methods in real data, the four cases of contamination.

From the results of tables 4.6, depending on the number of selected variables (NSV), MSE, RSE and Adj.R^2 , it is obvious that the implement of RSMAVE-EN, RSMAVE and SMAVE-EN are yielded similar result 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 V.S. 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 for real data prove that the performance of proposed RSMAVE-EN approach is very good. In addition, the RSMAVE-EN gives stable results even with all cases, with contamination or without contamination.

The Figure 4.6 for the previous Boston housing data shows that the MSE value for RSMAVE-EN is less than its values for RSMAVE and SMAVE-EN. This means that the suggested RSMAVE-EN has a better performance than the RSMAVE and SMAVE-EN methods depending on the MSE of real data.

4.3 Diabetes data

In this section, to verify the performance of the proposed RSMAVE-EN method, we used the SMAVE-EN, RSMAVE and RSMAVE-EN methods in analysis diabetic patient's data. The data was collected by the authors from Al-Imam Al-Sadiq Hospital in Al-Hila city, Babylonian governorate. The test included a sample of 105 patients who visited the hospital during the months of February, March and April in the year 2021. We considered the response variable y represents the reading of blood sugar, where the blood sugar level is read by a blood glucose meter. The normal value of blood glucose level is: (3.9-6.5) mmol/L (mill mole / liter). X includes twenty predictors as follows:

x_1 is Urea (blood urea), normal value range (2.5 – 7.5) mmol / L.

x_2 is Creat. (Serum Creatinine Test), normal value range (61–132) mmol /L.

x_3 is T.S.B (Total serum Bilirubin Test), normal value range (0.3 – 1.2) mg /dL (milligrams per deciliter).

x_4 is HBA1c (Hemoglobin A1), normal value range (4.2 – 6.2) %.

x_5 is ALK (Alkaline Phosphatase Test), normal value range (21 – 92) u / L (unit / liter).

x_6 is G.P.T (Glutamic Pyruvic Transaminase Test), normal value range (7 – 56) u / L(unit per liter of blood serum).

x_7 is G.O.T (Glutamic Oxaloacetic Transaminase Test), normal value range (8 - 45) u / L (unit per liter of blood serum).

x_8 is CHOL (Cholesterol Test), total cholesterol level of less than 200 mg / dL or (5.17 mmol / L) is normal.

x_9 is T.G (Triglycerides Test), normal value range is less than 150 mg / dL or (1.7 mmol / L).

x_{10} is U.ACID (Uric Acid), the normal value ranges for male: 4.0 – 8.5 mg /dL and for female: 2.7 – 7.3 mg/dL.

x_{11} is WBC (White Blood Cell), normal value is 4500 – 11000 WBCs per microliter.

x_{12} is PCV (Packed Cell Volume Test) it measures the ratio of the volume of red blood cells to the volume of all components of the blood together. Generally, the normal value range of PCV is considered to be: For male, 38.3 to 48.6 percent. For female, 35.5 to 44.9 percent.

x_{13} is HB (Hemoglobin), normal value range is considered to be: For male, 13.8 - 17.2 g/dL and for female, 12.1 to 15.1 g/dL.

x_{14} is ESR (Erythrocyte Sedimentation Rate), normal value range is: For male, 0 – 20 mm/hr (millimeters per hour) and 0 – 29 mm/hr for female. These values can vary depending on the patient's age.

x_{15} is S.Na (Serum Sodium), normal value range (136 – 155) mmol/L.

x_{16} is S.Ca (Serum Calcium), normal value range (2.0 – 2.6) mmol/L.

x_{17} is PLT (Platelet Count Test), normal value range (150000 - 400000) platelets per microliter of blood.

x_{18} is Iron, normal value levels are generally (35.5 - 44.9) percent for adult women and (38.3 - 48.6) percent for adult men.

x_{19} is S.K (Serum Potassium Test), normal value range (3.5–5.0) mmol/L.

And x_{20} represents the patient's age.

In order to investigate the influence of outliers on the suggested RSMAVE-EN approach, we contaminated the data by adding some outliers in x predictors and y . Four cases are considered to be: no outlier, a percentage of 5%, 10% and 15% contaminated observations. The data has been contaminated by replacing x and y values by C value which equal to 100. To evaluate the estimation accuracy for mentioned methods, we conducted a comparison based on the MSE (mean squared error), RSE (residual standard error) and prediction error for diabetes data. Also, we reported the NSV (number of selected variables) by SMAVE-EN, RSMAVE and RSMAVE-EN.

Table 4.7: Results of the comparison of diabetes analyses based on MSE and RSE.

Contamination percentage	Methods	MSE	RSE
No outlier	SMAVE-EN	0.979	1.004
	RSMAVE	0.631	0.806
	RSMAVE-EN	0.615	0.795
5%	SMAVE-EN	1.458	1.252
	RSMAVE	0.894	0.965
	RSMAVE-EN	0.804	0.914
10%	SMAVE-EN	1.90	1.399
	RSMAVE	0.959	0.975
	RSMAVE-EN	0.805	0.911
15%	SMAVE-EN	2.013	1.440
	RSMAVE	1.006	1.019
	RSMAVE-EN	0.848	0.939

Table 4.8: Results of Comparison of diabetes data based on prediction error.

Contamination percentage	Methods		
	SMAVE-EN	RSMAVE	RSMAVE-EN
No outlier	7.629	7.667	7.625
5%	16.361	9.758	9.407
10%	25.450	16.045	13.831
15%	34.427	20.167	18.069

Table 4.9: Comparison of diabetes data for the three methods based on number of selected variables (NSV).

outliers	Methods		
	SMAVE-EN	RSMAVE	RSMAVE-EN
No outliers	11	12	12
5%	14	10	10
10%	12	11	10
15%	13	11	10

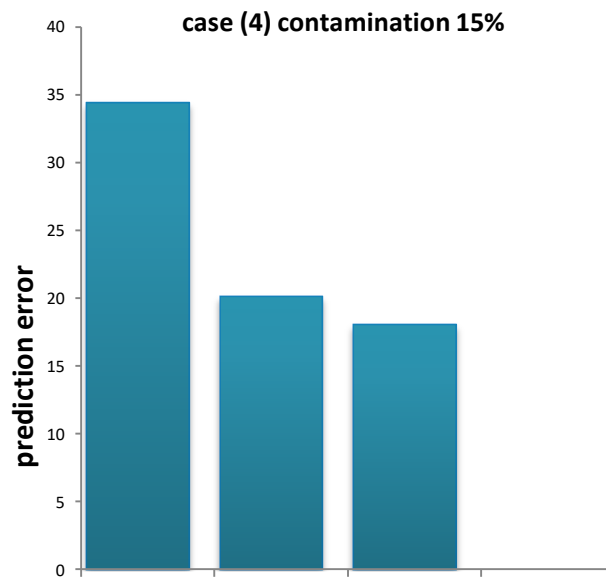
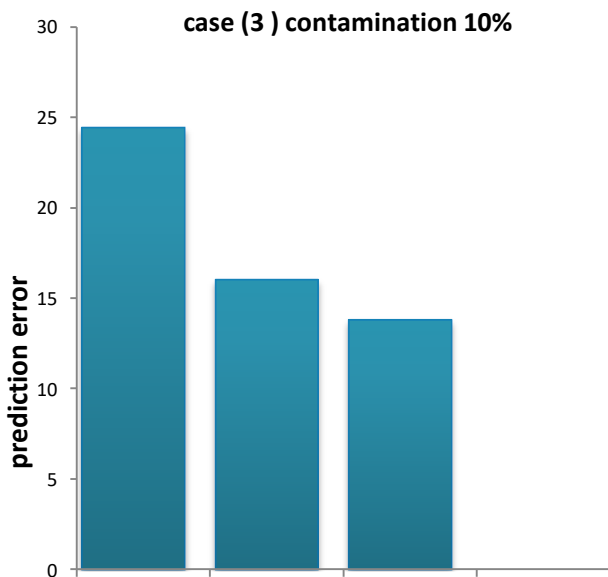
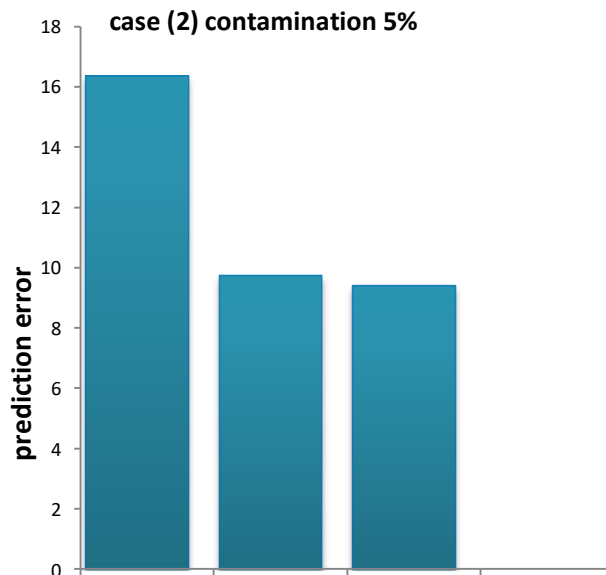
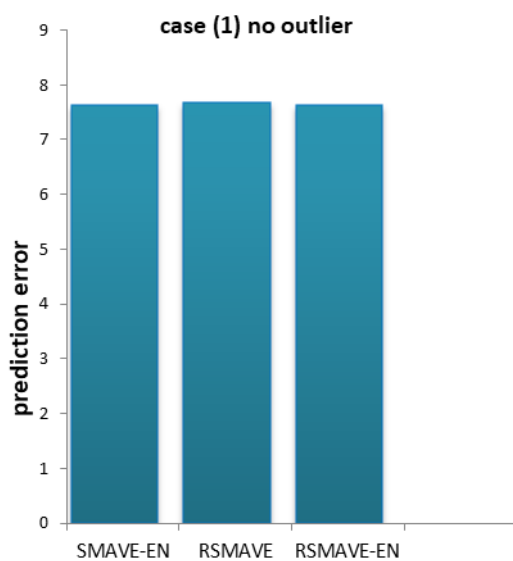


Figure 4.7: The prediction error for the considered methods in diabetes data, the four cases of contamination.

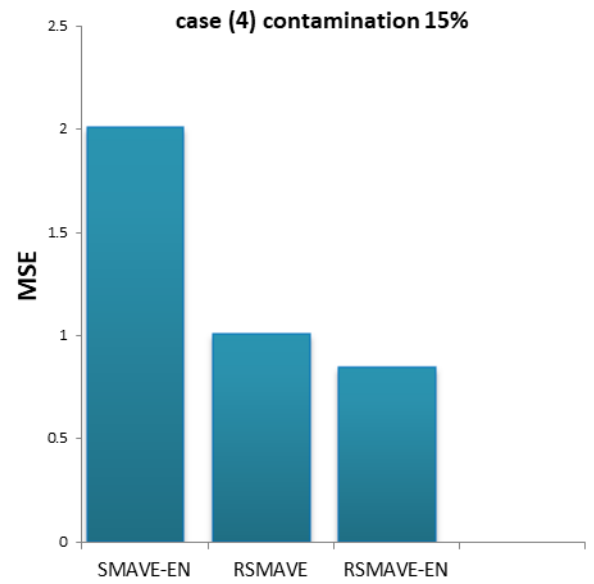
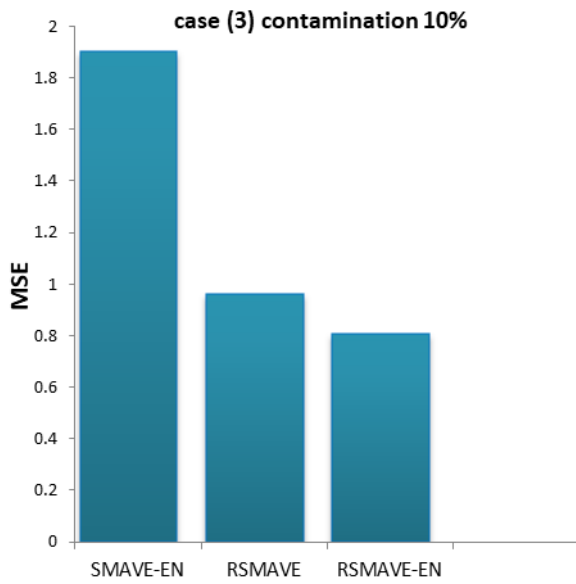
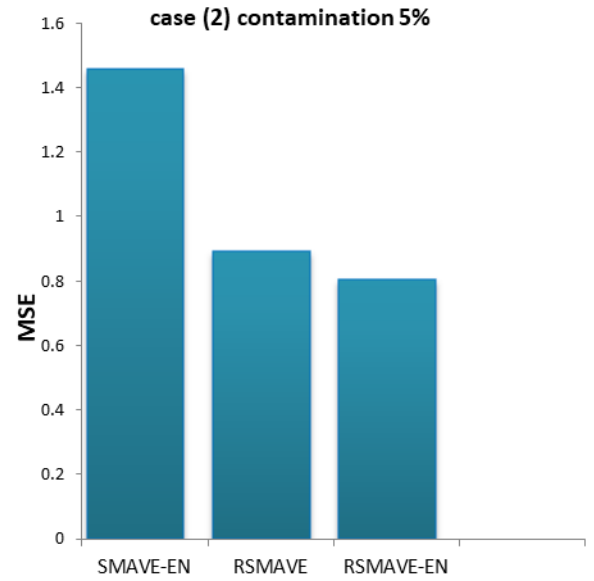
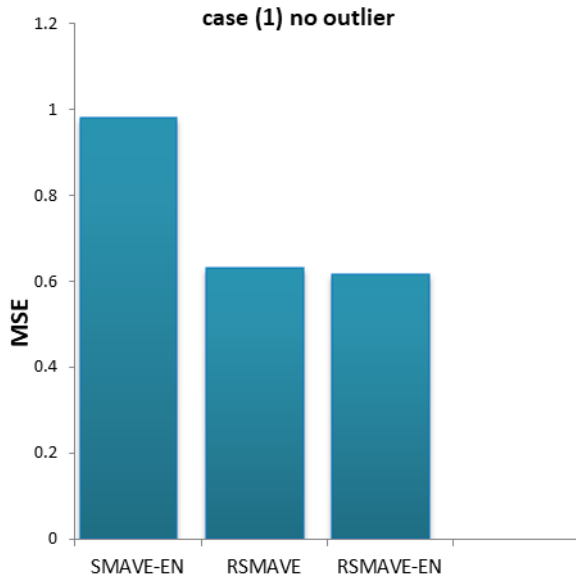


Figure 4.8: The MSE for the considered methods in real data, the four cases of contamination.

The tables 4.7, 4.8 and 4.9 show the outcomes of comparison for the diabetes data. The comparison demonstrated that, the reported methods yielded similar results in case of no outliers in estimation accuracy and V.S. Whereas, in the other three cases of data contamination, we can note that, the SMAVE-EN method is affected when adding outliers in the three cases of data contamination. While, the rest methods RSMAVE and RSMAVE-EN were not affected because they have the robustness. Also, we can see that the performance of suggested RSMAVE-EN outperformed RSMAVE method in terms of V.S and estimation accuracy. Depending on the above notes is obviously that, under various settings the suggested approach has a good behavior.

The Figures 4.7 and 4.8 for the diabetes data, show that the prediction error and the MSE values for RSMAVE-EN are the least than its values for RSMAVE and SMAVE-EN. This means that the suggested RSMAVE-EN has a better performance than the rest methods depending on the prediction error and the MSE values of diabetes data.

Chapter five

Conclusion, Recommendations and future work

5.1. Conclusion

We have proposed RSMAVE-EN method in this thesis. It is a robust approach to V.S and dimension reduction simultaneously. This approach has the efficiency when the predictors are highly correlated under SDR settings. 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 V.S and estimation accuracy even with the existence outliers 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 approaches. In addition, analytic results of diabetes data and Boston Housing (B.H) data showed that the suggested method has good and consistent results even with all contamination cases which considered through comparison with other RSMAVE and SMAVE-EN methods. Furthermore, the suggested approach maintains its properties in working with a nonlinear regression and multiple dimensions under SDR framework.

5.2. Recommendations and future work

We recommend using the suggested RSMAVE-EN approach in the analysis of the data set especially when there are existence outliers in the predictors and response. Also, we recommend the necessity of adopting electronic documentation to record the results of patient analyzes in clinics and hospitals. In order to, provide a database to facilitate the work of researchers.

The idea of robust suggested approach in this thesis can be extended for using in other SDR approaches. It is also possible to develop a similar work of suggested RSMAVE-EN for MAVE methodology to include the group V.S penalties such as, SMAVE - adaptive EN (SMAVE-ADEN) to produce a robust version of the estimate.

References

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In second International Symposium on Information Theory. Akademia Kiado, Budapest, pp. 267-281.
2. Alkenani, A. (2020). Robust variable selection in sliced inverse regression using Tukey biweight criterion and ball covariance. *Journal of Physics Conference Series*, 1664, 012034.
3. 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* 35 (2).
4. Alkenani, A. and Abdulkadhim, M. (2020). Regularized sliced inverse regression through the elastic net penalty. *Journal of Physics Conference Series*. Submitted.
5. Alkenani, A. and Aljobori, N. (2021). Robust sparse MAVE through elastic net penalty. *International journal of Agricultural and Statistical Sciences*, Vol.17, Supplement 1, 2039 - 2046.
6. Alkenani, A. and Dikheel, T. (2017). Robust Group Identification and Variable Selection in Regression. *Journal of Probability and Statistics* 2017, Article ID 2170816, 8 pages.

7. Alkenani, A. and Rahman, E. (2020). Sparse minimum average variance estimation via the adaptive elastic net when the predictors correlated, *Journal of Physics Conference Series*, 1591, 012041.
8. Alkenani, A. and Rahman, E. (2021). Regularized MAVE through the elastic net with correlated predictors, *Journal of Physics Conference Series*, 1897, 012018.
9. Alkenani, A. and Reisan, T (2016). Sparse sliced inverse quantile regression. *Journal of Mathematics and Statistics*. Volume 12, Issue 3.
10. Alkenani, A. and Yu, K. (2013). Sparse MAVE with oracle penalties. *Advances and Applications in Statistics* 34, 85–105.
11. Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton, New Jersey.
12. Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR," *Biometrics*, 64, 115-123
13. Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6), 2350–2383.
14. Brillinger, D. R. (1983). A generalized linear model with (Gaussian) regression variables. In *A Festschrift for Erich L. Lehmann* (eds P. J. Bickel, k. A. Doksum and J. L. Hodges, Jr), pp. 97-114. Belmont: Wadsworth.

15. Carlos, A. M. and Sergioc, C. S. (2012). Does BIC Estimate and Forecast Better than AIC?. Available at (<https://mpira.ub.uni-muenchen.de/42235/>).
16. Cizek, P. and Hardle, W. (2006). Robust estimation of dimension reduction space. *Computational Statistics and data analysis*, 51, 545-555.
17. Common, P. (1984). Independent component analysis, a new concept?. *Signal Processing*, 36(3), 287–314.
18. Cook, R. (1998). *Regression graphics: ideas for studying the regression through graphics*. New York, Wiley.
19. Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *The Annals of Statistics* 30, 455–474.
20. Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association* 86, 328–332.
21. Donoho, D. L., and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3), 425–455.
22. Efron, B. (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers*, 191–203.
23. Efron, B. et al. (2004). Least angle regression. *The Annals of Statistics* 32, 407–499.

24. 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.
25. Gorsuch, R. L. (1983). *Factor Analysis*, Hillsdale, New Jersey, L. Erlbaum Associates.
26. Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 1157 – 1182.
27. Hassel, M. (2021). Sparse sliced inverse regression via elastic net penalty with an application. Thesis submitted to college of administration and economics. University of Al-Qadisiyah. Iraq.
28. Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408), 986–995.
29. Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1–2), 71–120.
30. Jabbar, E. (2020). A non-linear multi-dimensional estimation and variable selection via regularized MAVE method. Thesis submitted to college of administration and economics. University of Al-Qadisiyah. Iraq.

31. Jolliffe, I. T. (2002). Principal components in regression analysis. *Principal Component Analysis*, 167–198.
32. Li, K. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* 86, 316–342.
33. 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.
34. Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika* 94, 603–613.
35. Li, L., Cook, R. D. and Nachtshiem, C. J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society Series B*, 67, 285–299.
36. Li, L., Li, B. and Zhu, L.-X. (2010). Groupwise dimension reduction. *Journal of the American Statistical Association*. 105, 1188–1201.
37. Li, L. and Nachtshiem, C. J. (2006). Sparse sliced inverse regression. *Technometrics* 48, 503–510.
38. Li, L. and Yin, X. (2008). Sliced Inverse Regression with regularizations. *Biometrics* 64, 124–131.

39. Malik, D. (2019). Sparse dimension reduction through penalized quantile MAVE with application. Thesis submitted to college of administration and economics. University of Al-Qadisiyah. Iraq
40. Ni, L. et al. (2005). A note on shrinkage sliced inverse regression. *Biometrika* 92, 242–247.
41. Powell, J. et al. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, 1403–1430.
42. Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of s-estimators. In *Robust and Nonlinear Time Series Analysis*, pages 256-272.
43. Salman, D. (2021). Sparse dimension reduction via regularized sliced inverse regression with application. Thesis submitted to college of administration and economics. University of Al-Qadisiyah. Iraq.
44. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
45. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

46. Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and statistics*, 2:448-485.
47. Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. Chapman and Hall, London. <http://dx.doi.org/10.1007/978-1-4899-4493-1>.
48. 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.
49. Wang, Q. and Yao, W. (2013). Robust Variable Selection through MAVE. *Computational Statistics and Data Analysis* 63, 42-49.
50. Wang, T. et al. (2013). Penalized minimum average variance estimation. *Statist. Sinica* 23 543–569.
51. Wang, T. et al. (2015). Variable selection and estimation for semi parametric multiple-index models. *Bernoulli* 21 (1), 242–275.10
52. Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, 35(6), 2654–2690.
53. Xia, Y. (2008). A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484), 1631–1640.

54. Xia, Y. et al. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B* 64, 363–410.
55. Yin, X. and Cook, R. D. (2005). Direction estimation in single index regressions. *Biometrika*, 92(2), 371–384.
56. Yin, X. et al. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99(8), 1733–1757.
57. Yu, Z. and Zhu, L. (2013). Dimension reduction and predictor selection in semi parametric models. *Biometrika*, 100, 641-654.
58. Zhang, C. H. (2010). Nearly unbiased variable selection under Minimax Concave Penalty. *Annals of Statistics* 38, 894–942.
59. Zhang, J. and Olive, D. J. (2009). Applications of a robust dispersion estimator. Southern Illinois University Carbondale.
60. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.
61. 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.
62. Zou, H., and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37(4), 1733.

الملخص:

يعد SDR (تقليل البعد الكافي) أحد الموضوعات المهمة في العديد من المجالات العلمية. لقد جذب إنتباه الباحثين لأنه يعتبر نهجاً مفيداً لمعالجة مشكلة البعد العالي (HD). ظهرت مشكلة البعد العالي بسبب البيانات الضخمة في السنوات الأخيرة. توصل العديد من الباحثين الى أفكار جديدة. لقد قاموا بدمج طرائق SDR مع طرائق "regularization" على سبيل المثال SMAVE-EN من بين مواضيع أخرى. ان طريقة SMAVE-EN هي طريقة إختيار متغير (V.S) بدون تحديد مسبق للنموذج. تدمج بين الحد الأدنى لمتوسط التباين (MAVE) و بين نهج EN. تكون SMAVE-EN فعالة عندما تكون المتنبئات شديدة الارتباط ضمن اعدادات SDR. ومع ذلك، فإن هذه الطريقة ليست حصينة تجاه القيم الشاذة وتنسم بالحساسية عند وجود القيم الشاذة في البيانات. في هذه الرسالة نقترح طريقة حصينة من طرائق اختيار المتغير هي RSMAVE-EN. يعمل هذا الاسلوب في ظل إعدادات توزيعات الخطأ المختلفة. ويعطي الحصانة اتجاه القيم الشاذة الموجودة في كل من المتغير التابع والمتغيرات المستقلة. تمّ التحقق من فعالية الطريقة المقترحة من خلال كلاً من دراسات المحاكاة وتحليل البيانات الحقيقية.



جمهورية العراق
وزارة التعليم العالي و البحث العلمي
جامعة القادسية / كلية الإدارة و الإقتصاد
قسم الإحصاء

التقدير الحمين وإختيار المتغير عن طريق تقدير الحد الأدنى لمتوسط التباين الصفري مع التطبيقات

رسالة مقدمة الى

مجلس كلية الإدارة والاقتصاد - جامعة القادسية
وهي جزء من متطلبات نيل شهادة الماجستير
في الإحصاء

من قبل

نعيم عبد عطوي الجبوري

بإشراف

أ.ر. علي جوار كاظم الكناني

٢٠٢٢ م

١٤٤٣ هـ