University Of AL-Qadisiyah

Available online at www.qu.edu.iq/journalcm

JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS

ISSN:2521-3504(online) ISSN:2074-0204(print)

Dimension reduction analysis of Ischemic stroke by using MAVE and SMAVE methods

Mohammed .H.AL-Sharoot [a], ***Sanaa.J.Tuama*** [b,*]

a Department of Statistics, College of Administration and Economics, University of Al Qadisiyah, Al Diwaniyah, Iraq. mohammed.alsharoot@qu.edu.iq

b Department of Statistics, College of Administration and Economics, University of Al Qadisiyah, Al Diwaniyah, Iraq. sanaa.j.tuama@qu.edu.iq

ARTICLE INFO

Article history:
Received: 14 /11/2024
Rrevised form: 4 /12/2024
Accepted : 29 /12/2024
Available online: 30 /12/2024

Keywords:

dimension reduction, local linear, MAVE, SMAVE, ischemic stroke.

ABSTRACT

With the rapid development that the world has witnessed in all scientific fields, data analysis has become more complex due to the large size of collected data that needs to be analyzed. Because of that the data analyst needs a powerful tool to deal with such that kind of data and this is why the dimension reduction methods have developed. In this paper, we focuses on minimum average variance estimation as powerful dimension reduction tools to analysis of ischemic stroke for sample of patients under some covariates. The aim of this paper is to determine the most important covariates that affect the size of ischemic stroke as response variable.

And the results indicated that the SMAVE method is the best method for prediction of the future ischemic stoke volume.

https://doi.org/10.29304/jqcsm.2024.16.41801

1. Introduction

The curse of dimensionality is very popular situation which captured the attention of many data analyst in various fields of sciences, such as, genetics, biology, finance, biomedicine, and the social sciences. Recently, with the development in computational procedures, very large datasets are conventionally collected in different fields of sciences. These datasets usually are with high dimension. So, the dataset with more variables is classified as high dimension dataset, which is usually more than the observations. Because of high dimensionality problem, the traditional and classical statistical methods are no longer applicable to analyze these high dimensional dataset by reason of the complicated structure of high dimensional dataset which is yields poor fitted model in regression analysis, Sheng and Yin (2006), Cook and Ni (2005) .

*Corresponding author: *Sanaa.J.Tuama*

Email addresses: *sanaa.j.tuama@qu.edu.iq*

Communicated by ‘sub etitor’

So, to reduce the high dimension of the underlying dataset, the sufficient dimension reduction procedure is the solution. The key idea of the sufficient dimension reduction is made to reduce the high dimension of covariates while keeping all the important information in the fitted regression model, Li and Wang, (2007).

As long as we focus in this paper on the regression model analysis, we will review the sufficient dimension reduction topic within the context of regression analysis; it is well-known that any violated of Gauss-Markov assumptions, the ordinary least square method will yields high variance estimators in case of presence the high dimensionality problem. To overcome the problem of high dimension, subset selection procedure to the underlying datasets has applied in many studies. So, the feature extraction method that represented by sufficient dimension reduction will be used in this paper. Other variable selection methods have studied as dimension reduction methods, such as, lasso method, Tibshirani (1996). the adaptive lasso method, Zou, H. (2006), elastic net method, and so on. Also, ridge method by Hoerl and Kennard (1970). has studied by many authors to address the problem of linearly dependence between the covariates (multicollinearty) that caused by the high dimensionality in the underlying datasets.

Consider the following multiple linear regression model,

$$\mathrm{Y} = \mathrm{f}(\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{P};\ \beta) + \epsilon. \ \dots(1)$$

Model (1) used to extract the information from the underlying dataset based on the dependence $(\mathrm{Y}|\mathrm{X})$ of $\mathrm{Y}_{\mathrm{n}\times 1}$, the response variable, on p-dimension $\mathrm{X} = (\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{P})^\mathrm{T}$ vector of covariates, where ϵ is the error term satisfy, $\mathrm{E}(\epsilon|\mathrm{X}) = 0$.

So, our goal in using the sufficient dimension reduction technique in regression model (1) is to reduce on the p-dimension covariates into K-dimension, where $\mathrm{k} < \mathrm{p}$, without losing any important information from the underlying dataset. In this paper, we consider the central mean subspace method of feature extraction which is named, Minimum Average Variance Estimator (MAVE), and Sparse Minimum Average Variance Estimator (SMAVE).

Xia et al. in 2002, proposed MAVE method as adaptive method in semiparametric regression model to reduce the space of dimension in high dimension dataset. Also, Hastie and Tibshirani in 1986(3) and Friedman and Stuetzle in 1981 studied how the functional approximation of the underlying regression dimension model can reduce the high dimension of X. In 1991, Li proposed the sliced inverse regression as reduction of X without any assumption about the regression function $\mathrm{E}(\mathrm{Y}|\mathrm{X})$. Principle component analysis is one of the most popular dimension reduction method, see Jolliffe in 2002 more information. Alkenani and Rahman in 2021 introduced (SMAV-EN) as dimension reduction method, this proposed method employed the penalized variable selection elastic net method. Also, in 2020 Rahman and Alkenani proposed the SMAVE penalized with adaptive elastic net as dimension reduction method. Moreover, in 2013, Alkenani and Yu studied the curse of dimensionality through employing some of variable selection methods, such as, adaptive lasso, Lasso, SCAD, and MCP with MAVE method.

Cook and Li in 2002 studied the dimension reduction problem in mean regression analysis $\mathrm{E}(\mathrm{Y}|\mathrm{X})$ based on the theory of central mean subspace, see Li in 2007 about the sparse sufficient dimension reduction estimator in regression model with starinkage method.

Motivated by the recent literature review and development of dimension reduction methods MAVE and SMAVE, we analyzed the regression model for ischemic stroke data to identify the important Covariates that effect the volume of ischemic stroke in a sample of patients.

2. MAVE Method

MAVE method essentially proposed to solve the high dimension problem in model (1) under the condition $Y \perp X|X^TB$. In model (1) the regression function $f(.)$ is unknown function defined on p-dimension covariates space. Based on the notion of sufficient dimension reduction, we interested in finding $\beta_1, \ldots.. \beta_K \in R^P$ such that Y regress on X in linear combination $\beta_1^T X, \ldots. \beta_K^T X$. Here, the dimension reduction matrix is $B_{p\times k}$, where the matrix B represents the linear combinations $\beta_1^T X, \ldots., \beta_K^T X$ of predictor variables which are orthogonal. So, based on the definition of central Subspace, the column space of matrix B is dimension reduction space for the conditional mean E(Y|X). Consequently, the smallest such that column space is named central mean space, denoted by $S_{E(Y|X)}$. Here, we will focus on MAVE method to estimate $S_{E(Y|X)}$.

Again, in model (1) if there is no pre-asumptions about the form of the regression function $g(X) = E(Y|X = x)$, then the suitable choice to handle with the function $g(X)$ is the nonparametric technique. Now, we can rewrite the model (1) and in dimension reduction perspective as follows,

$$y = h(B^T X) + \epsilon \ldots\ldots..(2)$$

Where $h(.)$ is the unknown link function, and $B_{p\times k}$ orthogonal matrix with $k < p$ under $E(\epsilon|X) = 0$.

If the conditions of model (2) are met, the projection of p-dimension X onto K-dimensional subspace $B^T X$ takes all the informatio which provided by X, see Xia et al. (2002) for more information. The B estimation is the solution that can be define by,

$$\min\,[E\{y - E(Y|B^T X)\}^2] \quad \ldots(3)$$

With

$$var(B^T X) = [E\{y - E(Y|B^T X)\}^2|B^T X].$$

So, we can rewrite the minimization problem in (3) as follows,

$$B = \min E[var(B^T X)], \quad \ldots(4)$$

Subject to constraint the $B^T B$= I. The minimization problem in (4) is called MAVE.

Based on nonparametric methodology, one can use the local linear expansion to approximate the function $h(.) = E(y_i|B^T X)$ for the observations of sample (X, y).

For a specific value of an observation, X_0, the local linear expansion value of X_0 can be define as follows,

$$E(y_i|B^T X_i) \approx \alpha + b^T B^T (X_i - X_0), \ldots (5)$$

where $\alpha = h(B^T X_0)$ and $b^T = (b_{(1)}, \ldots\ldots, b_{(k)})$, such that $b_{(d)}$ is the partial derivative of $h_B(v_1, \ldots\ldots, v_d)$ with respect to v_d, where $\beta_1^T X = v_1, \ldots \beta_d^T X = v_d$, $d = 1, \ldots, k$.

Therefore, the minimization problem in (3) can be found based on (4) and (5) as follows,

$$B = min_{B:B^T B=1}\{\textstyle\sum_{j=1}^{n} \widehat{var_\beta}(B^T X_j)\ \}$$

$$= \min\{\textstyle\sum_{j=1}^{n}\sum_{j=1}^{n}(y_i - \hat{y})^2 w_{ij}\ \} \ldots \qquad (6)$$

where $\hat{y} = \alpha_j + b_j^T B^T (X_i - X_j)$, $b_j^T = (b_{j1} \ldots b_{jk})$, and

$$w_{ij} = \frac{K_h(B^T(X_i - X_j\,))}{\sum_{l=1}^{n} k_h(B^T(X_l - X_j)}$$

Here, $K(.)$ is the product kernel function under $h_{m\times 1}$ vector of bandwidth. See Xin et al. in 2002 and Shi, in 2009 for more information. Finally, the minimization problem in (6) represents the MAVE method.

3. SMAVE method

Sparse MAVE method is another dimension reduction method that adds the penalized lasso method to the minimization problem in (6). This method developed by Wang and Yin in 2008 to overcome the difficulty of interpretability in MAVE method. The Sparse MAVE minimization problem can be written as follows,

$$\mathrm{Min}_{B:B^T B=1}\left(\sum_{j=1}^{n}\sum_{j=1}^{n}(y_i - \{\widehat{\alpha_j} + \widehat{b_j}(\hat{\beta}_1, \ldots, \hat{\beta}_m, \beta_m\ )^T\big(X_i - X_j\big)\ \}\ ]^2 w_{ij} + \lambda|\beta_m|\right) \qquad \ldots (7)$$

Where $w_{ij} = K_h\big[\widehat{B}^T(X_i - X_j)\big] / \sum_{i=1}^{n} K_h\big[\widehat{B}^T(X_i - X_j)\ ]\big]$and $\lambda > 0$ is the shrinkage parameter which can be estimated by the gentralized cross validation. See Wang and Yin in 2008 for details about the sparse MAVE algorithm. Moreover, to estimate K-dimensiona) of X, the Bayesian information criterion has been used,

$$BIC_k = \ln\left(\frac{RSS_k}{n}\right) + \frac{1}{nh^k}\ln(n)k,$$

Where,

$RSS = \sum_{j=1}^{n}\sum_{j=1}^{n}(y_i - \{\widehat{\alpha_j} + \widehat{b_j}(\hat{\beta}_1, \ldots, \hat{\beta}_m, \beta_m\ )^T\big(X_i - X_j\big)\ \}\ ]^2 w_{ij}$, is the residual sum of squares.

4. Real data analysis

A simple random sample of 235 observations was taken from the laboratory division at the Al-Diwaniyah teaching Hospital. This sample contained the important information about the patients of ischemic stroke which identified according to the vision of doctors specializing in neurological specialization. From statistics point of the view, we can study the relationship between the size of ischemic stroke as dependent variable and some covariates (smoking, gender, age, weight, blood pressure, D-dimer analysis, cholesterol, Covid-19 infection, hospitalizations, place of residence, marital status, and educational attainment) and the goal is to identify the most influential covariates on the changing in dependent variable. The following tables and figure are obtained by implementing the MAVE and SMAVE algorithm through the R-packages. Table (1) summarized the results o parameters estimates that obtained by using MAVE and SMAVE methods.

Table 1 : Parameters estimates by (MAVE , SMAVE)

Variable	MAVE	SMAVE
Age	0.4558	0.3847
Gender	0.1023	0.0003
Weight	0.2089	0.1852
Marital Status	0.0078	0.0014
Smoking Status	0.3574	0.3011
Blood Pressure	0.5023	0.4537
Diabetes	0.2561	0.0005
Educational Attainment	0.0005	0.0024
Cholesterol	0.4057	0.3228
Location	0.0003	0.0006
D_DIMER	0.5582	0.5016
Hospitalizations	0.0008	0.0007

The age estimate is (0.4558) using the MAVE method, while SMAVE gave a slightly lower value (0.3847), both indicating a positive association with the response variable. For the gender, in MAVE method the estimated value was in average difference (0.1023), while in SMAVE nearly zero this variable (0.0003), which indicates that SMAVE considers the gender variable to be less influential. For weight variable, both methods gave positive values, with MAVE (0.2089) and SMAVE (0.1852), indicating a positive effect. For marital status, it was close to zero in both methods and that indicating a non-significant effect. Smoking status maintained relatively high values in both methods, showing a positive effect on the outcome, with MAVE (0.3574) and SMAVE (0.3011).

In contrast, blood pressure also had a positive effect, achieving the highest value (0.5023) using MAVE, while SMAVE gave a slightly lower but positive value (0.4537). For the diabetes variable, the two methods showed noticeable differences. MAVE gave a positive estimate (0.2561), while SMAVE almost eliminated its effect (0.0000), which reflects SMAVE's tendency to exclude this factor. Academic achievement was considered similarly ineffective, with values close to zero for both methods (0.0005 for MAVE and 0.0000 for SMAVE). Cholesterol showed a consistent positive effect, with MAVE providing a higher estimate (0.4057) compared to SMAVE (0.3228). For the place of residence variable, both methods yields estimates close to zero (0.0000 for both MAVE and SMAVE), indicating no significant effects on the response variable. D_DIMER was one of the most influential variables, showing high positive values in both methods (0.5582 for MAVE and 0.5016 for SMAVE), showing a strong association with the response variable. Finally, the number of hospitalizations yielded low values in both methods (0.0008 for MAVE and 0.0000 for SMAVE), suggesting a limited effect.

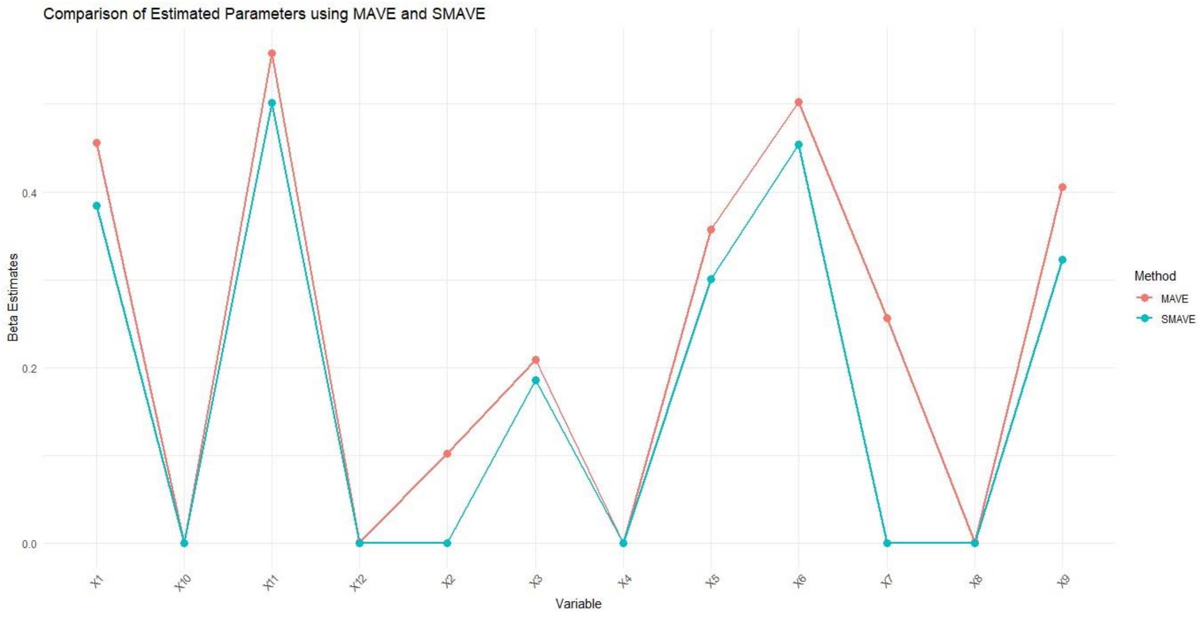

Figure 1 : Shows estimated parameters of variables using methods (MAVE , SMAVE)

From Figure 1, we notice that the MAVE and SMAVE methods behave similarly. MAVE tends to maintain a small effect across a larger number of variables, while SMAVE shows a stronger tendency to zero out or reduce the effect of less important variables, which characterizes it with a greater ability to select influencing factors with greater emphasis. In table 2, we summarized the values of adjusted R-squared or different fit models.

Table 2 : Adjusted R-squared values or different models

Model	MAVE Adjusted R^2	SMAVE Adjusted R^2
Linear	0.725	0.785
Quadratic	0.845	0.895
Cubic	0.810	0.870
Quartic	0.820	0.880

From the table above, we note that the SMAVE method consistently outperforms the MAVE method across all models: linear, quadratic, cubic, and quaternary. In the quadratic model, SMAVE achieves a higher Adjusted R^2 value (0.895) compared to MAVE (0.845), indicating a better ability in explaining nonlinear structure of the relationship between the response variable and function form of the covariates in terms of SMAVE method. The prediction error values have listed in following table.

Table 3 : Prediction error of estimated models

Method	Prediction Error
MAVE	0.1075
SMAVE	0.0775

As we know the prediction error is a measure used to assess the predictive ability of the estimated model. Table 3, above shows the prediction error values for both the MAVE and SMAVE methods, where the SMAVE method shows a lower error of 0.0775 compared to the MAVE method which achieves a higher value of 0.1075. This reduction in prediction error reflects the efficiency of SMAVE in providing more accurate and reliable estimates with $\hat{\mathrm{f}}(.) = \mathrm{E}(y)$, this making SMAVE the preferred method in this analysis.

5. Conclusions

In this paper, we investigated the problem of high dimensionality that occurs in covariates. The MAVE and SMAVE have studied in some details as methods for sufficient dimension reduction. The goal of this study was to apply the sufficient dimension reduction methods (MAVE and SMAVE) on the relationship of ischemic stroke dataset, where the volume of the ischemic stroke is the response variable on some covariates to identify the most important predictor variables that effect the response variable. The results of parameters estimates under MAVE and SMAVE show sparse solutions and great interpretability of these methods for the underlying dataset. Also, the models (linear, quadratic, cubic, quartic) have employed to demonstrate the ability of MAVE and SMAVE in representing true model and therefore the quadratic model gives the highest adjusted R-squared measure. Finally, the prediction error measure has used to assess the prediction ability of the response variable, and the results indicated that the SMAVE method is the best method for prediction of the future ischemic stoke volume.

References

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

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

[3] Bura, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. Ann. Statist. 17, 435-555.

[4] Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. Ann. Statist. 30, 455-474.

[5] Cook, R. D., and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. Journal of the American Statistical Association 100, 470, 410-428.

[6] Friedman, J. and Stuetzle, W. (1981). Projection pursuit regression. J. Amer. Statist. Assoc. 76, 817-823.

[7] Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 1, 55-67.

[8] Jolliffe, I. T. (2002). Principal Component Analysis. Springer, New York.

[9] Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc. 86, 316-342.

[10] Li, L. (2007). Sparse sufficient dimension reduction. Biometrika 94, 603□613.

[11] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. J. Amer. Statist. Assoc. 102, 997-1008.

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

[13] Sheng, W., and Yin, X. (2006). Sufficient dimension reduction via distance covariance. Journal of Computational and Graphical Statistics 25, 1 , 91-104.

[14] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 1, 267-288.

[15] 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.

[16] Xia, Y. C., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of optimal regression subspace. J. R. Statist. Soc. B. 64, 363-410.

[17] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association 101, 476, 1418-1429.