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

Using SIR-based methods to determine the number of explanatory variables effecting on the blood pressure data.

By

Dheyaa Alaboudi [1] and Prof. Dr. Ali Alkenani [2]

1,2Department of Statistics, College of Administration and Economics,
University of Al-Qadisiyah, Al Diwaniyah, Iraq.
Correspondence: Ali Alkenani, E-mail: ali.alkenani@qu.edu.iq
http://orcid.org/0000-0001-5067-2321

1442 A.H. **2021 A.D**

Abstract

In some multiple regression applications, the number of predictors has become large, and for this reason, the sufficient dimension reduction (SDR) theory (Cook, 1998) has received much attention. The idea of sufficient dimension reduction (SDR) is to replace X with a low-dimensional orthogonal projection P_SX on the subspaces (S) without Loss of information about the $Y|$X distribution X without assuming any specific model. The target of the SDR is the central subspace S$(_{Y|X})$ many methods have been worked out to find $S_{Y|X}$ and one such method is the sliced inverse regression (SIR) (Li, 1991). Applied in different fields, SIR has proven robust for dimension reduction (DR) approach and is effective in handling high dimensional (HD) data and sufficient tool to deal with dimension reduction (DR) in conditional regression (Li and Yin, 2008). However, it does produce linear combinations (LCs) for all the original predictors. As a result, interpretation of SIR estimates can be difficult and sometimes misleading.

In this paper will we use methods that combine SIR work with the Lasso method. Ni et al.(2005) A note on shrinkage sliced inverse regression(SH-SIR), Li and Yin (2008) sliced inverse regression with regularization (RSIR) and Lin et al. (2018) sparse sliced inverse regression via Lasso (SIR-L) methods in analysis sample data for high blood pressure and the factors affecting it.

1. Introduction:

Dimension reduction methods is one of the important things that the researcher needs in analysis high-dimensional data(HD), especially in recent years. After the development of data collection methods and the development of data storage methods and storage capacity, many dimensional reduction(DR) methods have been proposed, and they can be divided into two types of methods. Classical reduction methods such as principal compounds analysis (PCA) method, factor analysis, and discriminant analysis, and methods others. These methods began to suffer in the analysis of HD. Therefore, the researcher Cook 1998 proposed the theory of sufficient dimensions reduction (SDR) Alkenani and Yu (2013). The well-known approaches of the SDR provide the tool for finding sufficient dimensions with no need for pre know the error distribution or specific a model. A lot of methods were done for finding central subspace $S_{Y|X}$ and one of these methods is SIR (Li, 1991).where those approaches replace original predictors with the linear combinations(LCs) of predictors in which they're low-dimensional. However, the explanation of resulting estimations isn't simple due to the fact that every one of the DR components is a linear combinations to the every original predictors. To get rid of this problem, regularization methods have been added to the dimension reduction methods solutions.

where SIR has been combined with some regularization methods to obtain parameter estimation and predictors selection simultaneously,

Under the framework of the SDR, Ni et al. (2005) have presented a shrinkage SIR(SH-SIR). Li and Nachtsheim (2006) have proposed another version of sparse SIR. Li (2007) proposed Sparse SIR. Wang and Yin (2008) suggested sparse MAVE (SMAVE) approach. Alkenani and Yu (2013) proposed SMAVE with the Adaptive Lasso, SCAD and MCP penalties. Alkenani and Reisan (2016) suggested SSIRQ. Doaa (2019) suggested QR with MAVE (QMAVE) and QMAVE with Lasso penalty (LQMAVE). Esraa (2020) suggested SMAVE with the Elastic-net and Adaptive Elastic-net.

The remainder of this paper is as follows. In Section 2, a brief review of SIR. Simple presentation of the methods of analysis used is in Section 3. in Section 4 Real data analysis . Finally, the conclusions are presented in Section 5.

2. Sliced Inverse Regression SIR:

The SIR method was suggested by Li (1991). The basis of this method is to reverse the relationship in the traditional (classical) regression analysis. Regression analysis study the correlation of the dependent variable (y) with the independent variables (X) represented by E(y| x).While SIR study this relationship through E (x | y). Then we divide into the model off to multiple sliced according to the values of (y), next we conduct different statistical processes for each sliced. For the problems of the regression with scalar response variable y and a p-dimensional predictor $X=(x_1, \ldots, x_p)^T$ according model below:

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

where f($x_1, x_2, \ldots, x_p$) = E (y |**x**) and E(ε |**x**) = 0. The objective of the SDR or function is exploring central subset ($S_{Y|X}$) on subspaces S of the space of the predictor in a way that:

$$y \perp\!\!\!\perp \mathbf{x} \mid P_S\mathbf{x} \quad (2)$$

where $\perp\!\!\!\perp$ represents the statistical independence and $P_{(.)}$ represents an operator of projection. Subspaces that satisfy (2) are referred to as the central subspaces $S_{y|\mathbf{x}}$ Alkenani and Yu (2013), Cook (1996, 1998a). Thus if $d = \dim(S)$ and B = ($\beta_1, \beta_2, \ldots, \beta_d$) is a base for the subspaces (S), **x** may be exchanged by (LCs) $\mathbf{x}^T \beta_1$, $\mathbf{x}^T \beta_2, \ldots, \mathbf{x}^T \beta_d$, $d \leq p$ without losing any information on the $P_S\mathbf{x}$ .That is, f($x_1, x_2, \ldots, x_p$) = f($\mathbf{x}^T \mathbf{B}$). In the case where all subspace intersections satisfy (2), it is referred to as central subspaces (CS) Cook (1996, 1998a) and is represented as $S_{y|\mathbf{x}}$ **.** The model on which the SIR relies is similar to the semi-parameter regression model:

$$y = f\,(\beta_1^T X, \beta_2^T X, \ldots, \beta_k^T X, \epsilon) \quad (3)$$

where (β_ks) unknown vector, ϵ random error independent of X and f represents a random unknown function on R^{k+1} .

where collecting all sectors information and obtaining the underlying roots. Then the largest of which will be selected to represent the effective dimension reduction (e d r) vectors of the (SIR) respectively. Which represents the new format of data that act as parameters (β_ks) ,where data is converted to reduced form and replaced with the original data for ease of handling, and in this method the (HD)problem is remedy in SIR method. Li (1991).

3. A brief review of the methods of analysis used:

Ni et al. (2005) suggested the SH-SIR adding the ℓ_1 penalty to a SIR loss function. Based on the formula proposed by Cook (2004). Which provides an approximate form of the eigenvalue problem for SIR as the least squares problem, so that we can add penal methods using the SIR method with adding what suits their method of working. Cook (2004) SIR wrote as a problem to the least squares as follows

$$\text{minimises } F(A,C) = \sum_{y=1}^{h} \left\| \hat{f}_y^{1/2} \hat{Z}_y - A C_y \right\|^2 \quad (4)$$

Written by Ni et al. (2005) with the following formula:

$$\sum_{y=1}^{h}\left\|\hat{f}_y^{1/2}\hat{Z}_y - \hat{\Sigma}^{1/2} diag(\hat{B}\hat{C}_y)\alpha\right\|^2 + \lambda\sum_{i=1}^{p}|\alpha_i| \quad (5)$$

since SIR provides $\hat{B}$ and $\hat{C}$, there is an ability for the adoption of Lasso method, for the purpose of obtaining the indices of shrinkage $\hat{\alpha}$ as an argument α . see Ni et al. (2005).

Li and Yin (2008) suggest (RSIR) a scale SIR approach based on the SIR least squares formulation. L2 settlement is introduced, and an alternating least squares algorithm was developed, to enable SIR to work with n < p and are highly correlated prediction. ℓ_1 penalty is also introduced to realize simultaneous reduction estimation and predictors selection. see(Li and Yin ,2008)

Lin et al. (2018) (SIR-L) also proposed a SIR scale approach based on the least squares formulation. ℓ_1 penalty was introduced to achieve simultaneous reduction estimation and predictors selection according to a given algorithm. See(Lin et al. 2018)

4. Analysis real data:

In this section to we used the of SH-SIR, SIR-L and RSIR, in analysis real data on the blood pressure degrees person's. It is one of the diseases that humanity suffers in general. We examined a sample of 82 persons who visited the cardiac advisory at Karama Hospital in Al Kut, in August 2019. Data have been analysis with the use of the R-code. We got the results mentioned in tables (1) , (2) and (3).

where we considered the dependent variable (y_i) represents the reading of the pressure degrees. The variables that affect this reading it predictor vector (X) and as follows (x_1: Age of a person),(x_2: father's age),(x_3: Mother's age),(x_4: the brothers number),(x_5: the sisters number),(x_6: body mass index), (x_7: the degree of sugar),(x_8: glycated hemoglobin).(x_9: the ratio of urea),(x_{10}: Creatinine ratio),(x_{11}: white blood cell count),(x_{12}: the effective percentage of white blood cells),(x_{13}: Hemoglobin),(x_{14}: blood viscosity),(x_{15}: Platelets),(x_{16}, x_{17}: the proteins that are portion of immune the system).

We will analysis the real data the statistical methods mentioned above and using some statistical criteria to compare.

Table 1: Number of selected variables of methods.

Methods	Number of selected variables
SIR-L	13
SH-SIR	15
RSIR	15

In the first table, the real data for the degrees of blood pressure and the factors that affect them were analysis through the criterion for variables selection. The results indicate that the SIR-L method is better than the rest of the methods.

Table 2: The adjusted R-square values for the model fit.

		SIR-L	SH-SIR	RSIR
Model Fit	*Linear*	0.85	0.78	0.81
	Quadratic	0.90	0.85	0.88
	Cubic	0.91	0.86	0.88
	Quartic	0.90	0.88	0.88

In Table 2, we used adjusted R-square criterion for comparison between methods, so we found that SIR-L method have greater values than the rest of the methods and in all models used in the analysis. This proves the superiority of the method SIR-L.

Table 3: The prediction error of the Quadratic fit for the studied methods.

Methods	Prediction error
SIR-L	14.331
SH-SIR	16.829
RSIR	14.800

In Table 3, we use for the comparison (prediction error criterion) between the methods to analysis the data . We note that the SIR-L method is superior to the rest of the methods because it has the lowest prediction error of than the rest of the methods.

The figure1 demonstrates the superiority of the SIR-L method over the rest of the methods used in real data analysis, for both prediction error and adjusted R-square criterion. As shown in the graph.

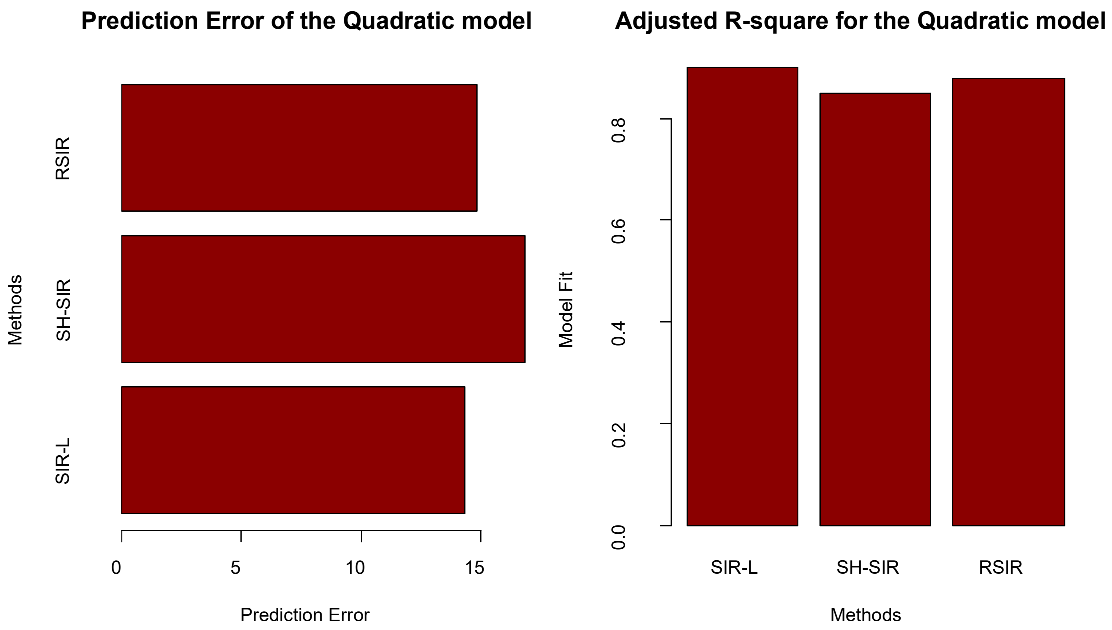

Figure 1: The prediction error criterion and the adjusted R-square criterion.

6. Conclusion:

In this paper, we used SH-SIR, RSIR and SIR-L methods that combine the Lasso method with (SIR) and compared the results obtained. Showed the superiority of the method suggested by Lin et al. (2018) SIR-L on the rest for the methods. This leads us to recommend the use of the SIR-L method in analysis high-dimensional data, as it gives more accurate results.

References

1. Alkenani, A. and Dikheel, T (2016). Sparse sliced inverse quantile regression. Journal of Mathematics and Statistics. Volume 12, Issue 3.

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

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

4. Cook, R. (2004). Testing predictor contributions in sufficient dimension reduction. Annals of Statistics 32,1061-92.

5. 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.

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

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

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

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

10. 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.

11. Ni, L., Cook, R. D. and Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. Biometrika 92, 242–247.

12. 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.