Republic of Iraq

Ministry of Higher Education

and Scientific Research

University of Al-Qadisiyah

College of Administration and Economics

Statistics Department

Analyzing of diabetes data Using SIR-based methods

By

Ali Alkenani[1] and Mohamed Abdulkadhim[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

1443 A.H. **2021 A.D**

Abstract

The SDR received great attention in high-dimensional regressions. Assume Y is a response variable and $\mathrm{X} = (\mathrm{x}_1, \dots, \mathrm{x}_p)^T$ is a predictor of p-dimensions. Without assuming any parametric model, the main idea of SDR is to replace X with a Low-dimensional orthogonal $P_s\mathrm{X}$ to S while retaining information about the Y | X distribution. The aim of SDR procedure is to find the central subspace $\mathrm{S}_{Y|\mathrm{X}}$, and that $\mathrm{S}_{Y|\mathrm{X}}$is the intersection of all subspaces such as $Y \perp\!\!\!\perp \mathrm{X} | P_s\mathrm{X}$. Where $\perp\!\!\!\perp$ denotes independence. Therefore, $P_\beta\mathrm{X}$ excerpts all the information from *X* about *Y*, where β is the base to $\mathrm{S}_{Y|\mathrm{X}}$. (Cook, 1998).

There are several proposed methods for finding $\mathrm{S}_{Y|\mathrm{X}}$, and one of the well-known methods is SIR (Li, 1991), SIR is applied in several fields including economics, and bioinformatics. SIR faces difficulties in interpreting the resulting estimates about SIR due to its production of linear combinations from all of the original predictors. To improve the interpretation of SIR analysis, it is necessary to decrease the number of non-zero coefficients which are also insignificant in the SIR directions.

The objective of our study is to reduce the number of nonzero coefficients in SIR directions for obtaining better interpretability. Through combining some of the regularization methods with the SIR method to produce sparse and accurate estimations.

in this paper will we employ methods that merge SIR work with the Lasso method. SSIR(Ni et al, 2005), RSIR (Li and Yin, 2008), SIR-LASSO Lin et al.((2018) methods in analyses sample data for diabetes.

1. Introduction

When the number of predictors is great, regression analysis in some applications is very difficult, and the high-dimensional analysis of data with a p×1 the outcome Y on a predictor vector was attracted the attention of many researches, and as a result a problem has arisen of what is known as the "dimensional curse"(Bellman, 1961). This problem occurs when the dimensions increase very quickly, and therefore the available data becomes sparse. The curse of dimensionality is a problem for most

statistical methods. Therefore, reducing the dimension of predictors is considered one of the useful tools which help to solve the problem of "dimensional curse".

In order to find the central subspace $S_{(Y|X)}$, several methods were chosen, and one of these methods is SIR (Li, 1991). , where these methods replace the original variables with linear combinations (LCs) of the predictors in which they are low dimensional. , and in order to get rid of this problem, regularization methods were added to the solutions of dimensionality reduction methods, where SIR was combined with some regularization methods to obtain parameter estimation and select predictors at same time.

In the framework of SDR, the SIR shrinkage estimator (SSIR) was also proposed by adding the Lasso penalty to the SIR least squares formulation by Ni et al. (2005) Li and Nachtsheim (2006) combined Lasso and LARS with SIR to produce SIR (SPSIR) in a scattered way Li (2007) combined a number of SDR methods with the concept of organization estimation, and that this strategy has been applied to SIR. Several SDR methods, and regulated SIR (RSIR) were proposed by Li and Yin (2008) in order to enable SIR to operate when p> n and the predictors are closely related, where p and n are the number of predictors and sample size respectively.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). Alkenani and Abdulkadhim (2020) suggested SSIR with the Elastic-net. Esraa (2020) suggested SMAVE with the Elastic-net and Adaptive Elastic-net.

The rest of the article was as follows, in the Section 2 we presented a brief review of SIR, in the Section 3 brief review of the methods of analysis used, and in the Section 4 Analysis Real data, while in the Section 5 Discussed conclusions.

2. SIR

For estimating the basis of $S_{y|\mathbf{x}}$, the SIR method was proposed by Li (1991). The SIR requires $Z = \Sigma^{-\frac{1}{2}}\big(X - E(X)\big)$, satisfy the condition $E(\mathrm{Z}|P_c\mathrm{Z}) = P_c\mathrm{Z}$, where $\Sigma_x = Cov(X)$ is the population covariance matrix of X and $c$is a basis of $S_{Y|\mathrm{Z}}$. This term binds $S_{Y|\mathrm{Z}}$ with the inverse regression of Z on Y. Symmetric kernel matrix of SIR is $M = cov\,[E(Z|Y)]$ and $Span(M) \subseteq S_{Y|\mathrm{Z}}$.

Let take a random sample of size n of (X, Y), which has a joint distribution. Let $\bar{X}$ is the sample mean of X. Also, assume that $\hat{Z} = \hat{\Sigma}^{-\frac{1}{2}}(X - \bar{X})$ is the sample version of Z, where $\hat{\Sigma}$ is the sample covariance matrix of X. Let h is the number of slices and n_y is the number of observations in the yth slice. Thus, $\widehat{M} = \sum_{y=1}^{h} \hat{f}_y \hat{Z}_y \hat{Z}_y^T$ is the sample version of M, where $\hat{f}_y = n_y/n$ and $\hat{Z}_y$is the average of Z in the slice y. Let $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \lambda_p \geq 0$are the eigenvalues corresponding to the eigenvectors$\hat{v}_1, \hat{v}_2, \ldots\ldots, \hat{v}_p$ of $\widehat{M}$. If the dimension d of $S_{Y|Z}$ is known, $span(\hat{\beta}) = span(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_d)$is a consistent estimator of$S_{Y|X}$, where$\hat{\beta}_i = \hat{\Sigma}^{-\frac{1}{2}}\hat{v}_i$.

The SIR provides an estimator $span(\hat{\beta})$ of $S_{Y|X}$. Usually, the elements of $\hat{\beta} \in \mathbb{R}^{p\times d}$ are nonzero. In the construction of 'sufficient predictors', only the important predictors are needed if the number of predictors is large or the predictors are highly-correlated. To this end, a number of regularizations methods were employed with SIR by many researchers to compress some rows of$\hat{\beta}$ to 0's.

To improve interpretability, the SIR was formulated as a regression type optimisation problem by Cook (2004) through minimising

$$F(A, C) = \sum_{y=1}^{h} \left\| \hat{f}_y^{1/2} \hat{Z}_y - AC_y \right\|^2, \ldots\ldots \qquad (1)$$

over $A \in \mathbb{R}^{p\times d}$ and $C_y \in \mathbb{R}^d$, with $C = (C_1, \ldots\ldots, C_h)$. Let $\hat{A}$ and $\hat{C}$are the values of A and C that minimise F. Then $span(\hat{A})$ equals the space spanned by the d largest eigenvectors of M. By focusing on the coefficients of the X variables, Ni et al. (2005) rewrite (1) as

$$G(B, C) = \sum_{y=1}^{h} \left(\hat{f}_y^{1/2} \hat{\Sigma}^{-\frac{1}{2}} \hat{Z}_y - BC_y \right)^T \hat{\Sigma} \left(\hat{f}_y^{1/2} \hat{\Sigma}^{-\frac{1}{2}} \hat{Z}_y - BC_y \right), \ldots\ldots (2)$$

where $B \in \mathbb{R}^{p\times d}$. The value of B which minimises (2) is exactly $\hat{\beta}$ and $span(\hat{\beta}) = span\left(\hat{\Sigma}^{-\frac{1}{2}} \hat{A}\right)$ is the estimator of $S_{Y|X}$.

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

3.1.SSIR

Ni et al. (2005) suggested a shrinkage SIR estimated (SSIR) of $S_{Y|X}$ is $span(diag(\tilde{\alpha})\hat{\beta})$, where the shrinkage indices $\tilde{\alpha} = (\tilde{\alpha}_1, \dots, \tilde{\alpha}_p)^T \in \mathbb{R}^p$ are determined by minimising

$$\sum_{y=1}^{h} \left\| \hat{f}_y^{1/2} \hat{Z}_y - \hat{\Sigma}^{\frac{1}{2}} diag(\hat{B}\hat{C}_y)\alpha \right\|^2 + \lambda \sum_{i=1}^{p} |\alpha_i| \ , \dots \dots \qquad (3)$$

where, $\hat{B}$ and $\hat{C} = (\hat{C}_1, \dots, \hat{C}_h)$ minimise (2.3).

A standard Lasso algorithm can be employed to carry out (2.3). To be specific, let $\tilde{Y} = vec(\hat{f}_1^{1/2}\hat{Z}_1, \dots, \hat{f}_h^{1/2}\hat{Z}_h) \in \mathbb{R}^{ph}$

and $\tilde{X} = \left(diag(\hat{B}\hat{C}_1)\hat{\Sigma}^{\frac{1}{2}}, \dots, diag(\hat{B}\hat{C}_a)\hat{\Sigma}^{\frac{1}{2}}\right)^T \in \mathbb{R}^{ph \times p}$,

where $vec(.)$ is a matrix operator that stacks matrix's columns to single vector. Then the vector α, exactly the estimated of Lasso for the regression $\tilde{Y}$ on $\tilde{X}$.

3.2.RSIR

Li and Yin (2008) derived another least squares formula for SIR, which is equal to

$$\mathrm{V}(\mathrm{B}, \mathrm{C}) = \sum_{\mathrm{m}=1}^{\nu} \hat{f}_{\mathrm{m}} \left\| \overline{\mathrm{Z}}_{\mathrm{m}} - \mathrm{BC}_{\mathrm{m}} \right\|^2, \dots \dots (4)$$

and for the original prediction scale it was as follows:

$$\tilde{\mathrm{V}}(\mathrm{B}, \mathrm{C}) = \sum_{\mathrm{m}=1}^{\nu} \hat{f}_{\mathrm{m}} \left\| (\overline{\mathrm{X}}_{\mathrm{m}} - \overline{\mathrm{X}}) - \hat{\sum}_{\mathrm{x}} \mathrm{BC}_{\mathrm{m}} \right\|^2, \dots \dots (5)$$

then they proposed the ridge sliced inverse regression (RSIR) estimator by:

$$\mathrm{V}_{\Theta}(\mathrm{B}, \mathrm{C}) = \sum_{\mathrm{m}=1}^{\nu} \hat{f}_{\mathrm{m}} \left\| (\overline{\mathrm{X}}_{\mathrm{m}} - \overline{\mathrm{X}}) - \hat{\sum}_{\chi} \mathrm{BC}_{\mathrm{m}} \right\|^2 + \Theta vec(B)^T vec(B) \dots \dots (6)$$

where Θ is a non-constant constant Negative and vec(.), which is a matrix operator, is a packet that includes all the columns of the matrix in one vector, and to get a constant Θ, Li and Yin (2008) proposed an alternating least squares algorithm of least possible 4, and this algorithm mechanism is as follows: From given B, we can Get C through $\hat{C} = (\hat{C}_1, \dots, \hat{C}_\nu)$ Where $\hat{C}_m = (B^T \hat{\sum}_x^2 B)^{-1} B^T \hat{\sum}_\chi (\bar{X}_m - \bar{X})$,

$$m = 1, \dots, \nu$$

then rewrite 4 as least squares regression

$$\mathrm{V}_{\Theta}(\mathrm{B}, \mathrm{C}) = \left\| \tilde{K}^{1/2} \tilde{Y} - \tilde{K}^{1/2} \left(C^T \otimes \hat{\sum}_\chi\right) vec(B) \right\|^2 + \Theta vec(B)^T vec(B),$$

$$\dots \dots (7)$$

where $\otimes$ is Kronnecker.

product$\tilde{Y} = vec(\bar{X}_1 - \bar{X}, \dots, \bar{X}_v - \bar{X}), \tilde{K}^{1/2} D_f^{1/2} \otimes I_P$, and $D_f = diag(\hat{f}_1, \dots, \hat{f}_i)$. Given C , the solution of Bin (6) is

$$vec(\hat{B}) = (CD_f C^T \otimes \hat{\sum}_x^2 \Theta I_{pd})^{-1} (CD_f \otimes \hat{\sum}_x) \tilde{Y}$$
$$, \dots \dots (8)$$

This procedure will continue between minimizing B and C to a minimum until reached to convergence. Li and Yin (2008) derived the generalized cross-validation standard(GCV).

To determine the ridge parameter Θ in (4) is given by

$$GCV = \frac{\left\| (I_{pv} - \Gamma_\Theta \tilde{K}^{1/2} \tilde{Y} \right\|^2}{pv\{1 - trace(\Gamma_\Theta)/pv\}^2}, \dots \dots (9)$$

where

$$\Gamma_\Theta = (D_f^{1/2} C^T \otimes \hat{\sum}_x)(\hat{C} D_f \hat{C}^T \otimes \hat{\sum}_x^2 + \Theta I_{pd})^{-1} (\hat{C} D_f^{1/2} \otimes \hat{\sum}_x)$$
$$, \dots \dots (10)$$

are the ridge estimates of the SIR (RSIR) and they represent linear combinations of all predictors, and no variable selection is achieved. Li and Yin (2008) followed the estimator idea which is the contraction factor, as well as the lesser choice of the lasso coefficient idea of the RSIR estimator in order to induce variance in the estimated linear groups. Let's $(\hat{B}, \hat{C})$ which refers to the estimated RSIR. α which is the cut-off inverse regression estimator (SRSIR) for the central space $S_{Y|X}$ which is defined as $span(diag(\hat{\alpha})\hat{B})$ where the contraction index vector $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_p) \in \mathbb{R}^p$ can be obtained by minimizing

$$V_\lambda(\alpha) = \sum_{m=1}^{v} \hat{f}_m \left\| (\bar{X}_m - \bar{X}) - \hat{\sum}_X diag(\hat{\alpha}) \hat{B} \hat{C}_m \right\|^2, \dots \dots (11)$$

over where $\sum_{j=1}^{p} |\alpha_j| \leq \lambda$ is subject to some non-negative constant λ

Because$diag(\alpha)\hat{B}\hat{C}_m = diag(\hat{B}\hat{C}_m)\alpha, we\ have V_\lambda(\alpha) = \sum_{m=1}^{v} \hat{f}_m \left\| (\bar{X}_m - \bar{X}) - \hat{\sum}_X diag(\hat{B}\hat{C}_m)\alpha \right\|^2$

Let $\tilde{Y} = vec(\bar{X}_1 - \bar{X}, \dots, \bar{X}_i - \bar{X}) \in \mathbb{R}^{pv}$

$$\tilde{X} = (diag(\hat{B}\hat{C}_1)\hat{\sum}_X, \dots, (diag(\hat{B}\hat{C}_1)\hat{\sum}_X)^T \in \mathbb{R}^{pv \times p}$$

The shrinkage vector α is exactly the lasso estimator for regression $\tilde{Y}$ with *pv* with the observations on the $\tilde{X}$-dimensional data matrix. RIC, BIC, AIC methods are used to determine λ. Li and Yin (2008) are based on a criterion proposed by Zuh et.al. (2006) to estimate $d = \dim(\ S_{Y|X})$ Zhu et al.(2006) suggested d that can be estimated by:

$$\hat{d} = arg \max_{0 \le e \le p-1} \left\{ \frac{n}{2} \sum_{i=1+\min(\Phi\gamma, e)}^{p} \log(\hat{\varrho}_i) + 1 - \hat{\varrho}_i) - \frac{O_n(2p - e + 1)}{2}, \right\}, \dots\dots (12)$$

where, the matrix an $\Pi\zeta = cov(E(X/Y) + I_P$)d$\hat{\varrho}_1, \dots, \hat{\varrho}_p$ Where it shows the eigenvalues of the sample estimate $\hat{A}$ of A , Φ is the number of$\hat{\varrho}_1 \geq 1$, andO_n , and is a penalty constant taken to be $O_n = (\log(n)v/n)$.

3.3.SIR-LASSO

In this part we will introduce the lasso effective variable from the SIR of the multi-indicator model (1) with the general covariance matrix Σ, considering primarily the single-indicator model $y = f(\beta^{\tau}x, \epsilon)$. Let η be the vector The eigenvalue associated with the largest eigenvalue var($E[x|y]$). where $\beta \propto$Σ^(-1) η, and to estimate the area extended by β there are two methods. The first approach as discussed by Lin et al[2015] for estimations $\Sigma^{-1}$and η separately (see logarithmic 1), and the second approach avoids direct estimation of $\Sigma^{-1}$by solving the penalized least square problem: $\| \frac{1}{n} XX^{\tau}\beta - \eta \|_2^2 + \mu \| \beta \|_1$where X is the matrix of the variable p × n sampled (see Algorithm 2). However, as with most L_1penalty methods for nonlinear models, the theoretical basis for this approach is not understood. This is because these two approaches provide good estimates compared to previous approaches (eg, Li (1991), Li and Nachtsheim (2006), Li (2007), and as described in Lin et al. (2015) and supplementary materials.

3.3.1. Sparse SIR for High Dimensional Data

Throughout this part we will adopt the following symbols. As for the matrix V, we call the space created by the column vectors the column area and denote it by the column $col(V)$. As i-th and j-th for the first and second rows of the matrix, we denote them by the symbol V_i,∗ and $V\ *,_j$, respectively. And for the (column) vectors x and $\beta \in R^p$,we denote their intrinsic product $\langle x, \beta \rangle\ by\ x(\beta)$,and the k-th entry of x by x

(k). And in the case of two positive numbers a and b, we use $a \vee b$ and $a \wedge b$ to denote max$\{a, b\}$ and min$\{a, b\}$ respectively; We also use C', C'', C_1 and $C_2$to refer to general absolute constants, although the actual value may vary from case to case. For the sequences $\{a_n\}\, and\, \{b_n\}$, we denote $a_n \succ b_n$ and $a_n \prec b_n$if there are positive constants C', and C'', as $a_n \geq Cb_n$ and $a_n \leq C''b_n$ respectively. We denote $a_n \asymp b_n$ if both $a\_n \succ b_n$ and $a_n \prec b_n$hold. The base $(1, \infty)$ norm and (∞, ∞) norm of matrix A are defined as $\| A \|_{1,\infty} = max_{1\leq j\leq p} \sum_{i=1}^{p} |A_{i,j}|$ and $max_{1\leq j\leq n}|A_{i,j}|$ respectively. To simplify the discussion, we assume that $\frac{s\, log(p)}{n\lambda}$is small enough. We emphasize again that the data for our *X* variable is a $p \times n$ rather than the traditional $n \times p$ matrix.(Lin et al, 2018)

3.3.2. Diagonal Thresholding -SIR.

When p ≫ n and by marginal sorting of all variables across the diagonal elements of $\widehat{\Lambda}H$ the diagonal threshold (DT) check method continues (Lin et al. , 2015) and then apply SIR to those retained variables to obtain an estimate for column (B). This procedure is shown to be constant if the number of non-zero entries in each row of Σ is restricted.

Algorithm 1 (DT-SIR)

1- To determine the set of important predictors z̧, with| z̧ |= x o(n) We use the diagonal element magnitudes of $\widehat{\Lambda}H$

2- To estimate a subspace $\widehat{S_I}$.ˆ. SIR is applied to the data $(y, x_{\text{z̧}})$

3: By filling in 0′s for non-significant predictions. We expand $b\, \widehat{S_I} to\, a$ into a subspace in R^p.

3.3.3. Matrix Lasso

By solving the L_1 penalty problem, we can override the estimation and reflection of Σ because Σ col(B)=col(Λ), Li (1991). is held at the population level, and by solving a sample version of the equation with a regularization term suitable for dealing with high dimensionality a reasonable estimate of $col(B)$ can be obtained. And suppose that $\hat{\eta}_1, \cdots, \hat{\eta}_d$are the eigenvectors associated with the largest

eigenvalues of $\widehat{\Lambda}H$. Replacing Σ with its typical version $\frac{1}{n}XX^{\tau}$and imposing the L_1 penalty (Lin et al, 2018), we get a sample version of Σcol(B)=col(Λ), Li (1991).

$$\| \| \frac{1}{n}XX^{\tau}\beta - \hat{\eta}_i \| \|_2^2 + \mu \| \| \beta \| \|_1, \dots \dots (13)$$

for some appropriate μ_i's.

Algorithm 2 (Matrix Lasso)

1: Let $b_1, \dots, b_d$be the eigenvectors associated with the largest d eigenvalues of $\widehat{\Lambda}H_i$

2: For $1 \leq I \leq d$, let $b_i$be the minimizer of equation (13);

3: Estimate the central space col(B) by $col(b_1, \dots, b_d)$.

To produce scattered estimates of $\boldsymbol{\beta_i}$, this simple procedure can be easily implemented. Experimentally, it works reasonably well, so we set it as another benchmark to compare it to. Since we later noticed that its numerical performance was always worse than that of the main SIR-LASSO algorithm, we did not further investigate its theoretical properties (Lin et al, 2018).

3.3.4. The SIR-LASSO algorithm.

First we think of the single index model

$$y = f(x^{\tau}\beta_0, \epsilon), \dots \dots (14)$$

Suppose that $((x_i, y_i), i = 1, \dots, n,$ are arranged in such a way that $y_1 \leq y_2 \leq \cdots \leq y_n$. and without loss of generality. Construct an n × H matrix $M = I_H \otimes 1_c$, where $1_c$is the $c \times 1$ vector with all entries being 1. Then, according to the definition of X_H, we can write $X_H = XM/c. Let\ \hat{\lambda}$ be the largest eigenvalue of $\widehat{\Lambda}_H$=$\frac{1}{H}X_H X^{\tau}H$ and let $\hat{\eta}$ be the corresponding eigenvector of length 1 That is,

$$\hat{\lambda}\hat{\eta} = 1_H X_H X^{\tau}{}_H \hat{\eta} = \frac{1}{nc}XMM^{\tau}X^{\tau}\hat{\eta}.$$

Thus, by defining

$$\tilde{y} = \frac{1}{c\hat{\lambda}}MM^{\tau}X^{\tau}\hat{\eta}, \dots \dots (15)$$

We have $\hat{\eta}=\frac{1}{n}X\tilde{y}$ and note that a key in estimating the central space $col(\boldsymbol{\beta})$ of SIR is the equation $\boldsymbol{\eta} \propto \boldsymbol{\Sigma\beta}$ If approximating $\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\eta}}$ and $\frac{1}{n}XX^{\tau}$ respectively. This equation can be written as $\frac{1}{n}X\tilde{y} \propto \frac{1}{n}XX^{\tau}\beta$. To restore vector sparse β^∝β , and we can consider the following optimization problem $min \parallel \beta \parallel_1$, subject to $\|X(\tilde{y}-X^{\tau}\beta)\|_{\infty} \leq \mu$,

which is known as the Dantzig selector (Candes and Tao, 2007). A related formulation is the Lasso regression, where $\boldsymbol{\beta}$ is estimated by the minimizer of

$$\mathscr{b}_{\beta} = \frac{1}{2n} \parallel \tilde{y} - X^{\tau}\beta \parallel_2^2 + \mu\|\beta\|_1, \ldots\ldots (16)$$

As described by Bickel et al. (2009), the Dantzig determinant is asymptotically equivalent to Lasso for linear regressions, so we propose and study the Lasso-SIR algorithm in this part:

Algorithm 3 (SIR-LASSO for single index models)

1: Let $\hat{\lambda}$and $\hat{n}$be the first eigenvalue and eigenvector of $\widehat{\Lambda}_H$, respectively.

2: Let $\bar{y} = \frac{1}{cb}MM^{\tau}X^{\tau}\hat{n}$and solve the Lasso optimization problem

$$\hat{\beta}(\mu) = argmin\ \mathscr{b}_{\beta}, where\ \mathscr{b}_{\beta} = \frac{1}{2n} \parallel \tilde{y} - X^{\tau}\beta \parallel_2^2 + \mu\|\beta\|_1.$$

where $\mu = C\sqrt{\frac{log(p)}{n\lambda}}$ for sufficiently large constant C;

3. Estimate P_{β} by $P_{\widehat{\beta}(\mu)}$.

There is no need for an inverse estimate of Σ in SIR-LASSO. Moreover, since the optimization problem (16) was well studied for linear regression models (Tibshirani, 1996, Efron et al. 2004, Friedman et al. , 2010), we may officially 'transfer' their results to index models. Practically, we use the *R* glmnet package to solve the optimization problem, where the adjustment parameter μ is chosen by using cross validation last but not least, Lasso-SIR can be easily generalized to the multiple index model (1).

And suppose that $\hat{\lambda}_i, 1 \leq i \leq d$, be the d-top eigenvalues of $\widehat{\Lambda}_H$ and $\hat{\eta}$ =($\hat{\eta}_1$,⋯,$\hat{\eta}_d$) form the corresponding eigenvectors. Similar to defining a 'pseudo-response variable'

for the single index model, we define a multivariate spurious response $\tilde{Y}$as

$$\tilde{Y} = \frac{1}{c} MM^{\tau} X^{\tau} \hat{\eta}\, diag\left(\frac{1}{\hat{\lambda}_1}, \cdots, \frac{1}{\hat{\lambda}_d}\right), \ldots \ldots (17)$$

.

To obtain the corresponding estimate, we apply a Lasso to each column of the pseudo-response matrix.

Algorithm 4 (SIR-LASSO: for multiple index model)

1: Let $\hat{\lambda}_i$ and $\hat{\eta}_i, i = 1, \ldots, d$ be the top d eigenvalues and eigenvectors of $\widehat{\Lambda}_H$ respectively.

2: Let $\tilde{Y} = \frac{1}{c} MM^{\tau} X^{\tau} \hat{\eta}\, diag(\frac{1}{\hat{\lambda}_1}, \cdots, \frac{1}{\hat{\lambda}_d})$.

For each $1 \le i \le d$, solve the Lasso optimization problem

$$\hat{\beta}(\mu) = argmin\ \ell_{\beta}, where\ \ell_{\beta} = \frac{1}{2n} \| \tilde{y} - X^{\tau}\beta \|_2^2 + \mu\|\beta\|_1.$$

where $\mu = C\sqrt{\frac{log(p)}{n\lambda}}$ for sufficiently large constant C;

3: Let $\hat{\beta}$ be the matrix of $\hat{\beta}_1, \ldots, \hat{\beta}_d$. The estimate of P_{β} is given by $P_{\hat{\beta}}$, and the number of directions d plays an important role when implementing algorithm 4, and it is common practice to determine the maximum gap between the ordered eigenvalues of the matrix $\widehat{\Lambda}_H$, which does not work well under HDLSS settings, and there is also a gap between the adjusted $\hat{\lambda}_i^a = \hat{\lambda}_i \| \hat{\beta}_i \|_2$ where $\hat{\beta}_i$i is the first output of algorithm 4. Motivated by this, we estimate d according to the following algorithm:

Algorithm 5 Estimation of the number of directions d

1: Apply Algorithm 4 by setting d = H;

2: For each i, calculate $\hat{\lambda}_i^a = \hat{\lambda}_i \| \hat{\beta}_i \|_2$

3: Apply the k-means method on $\hat{\lambda}_i^a$

with k being 2 and the total number of points in the cluster with larger $\hat{\lambda}^a$isthe estimated value of d. (Lin et al, 2018).

4.Analysis Real data:

In this section, diabetic data were analyzed by SSIR-EN, SIR-LASSO, RSIR and SSIR methods. We studied the most important factors in the test data, as well as the

most important factors that affect the sugar level, and the study included data collection from Thi Qar governorate / Thi Qar health directorate / diabetes and Endocrinology Center. Registration 2013. In Thi Qar, 22 variables were studied. The data was analyzed by code (R). After analyzing the data, we got the results in Tables 1 and 2.

The diabetic major data include n=186 trials. response Y Is the percentage of sugar. X_1 (blood type), X_2 (gender), X_3 (age), X_4(place of residence), X_5(family medical history), X_6(kinship of parents), X_7(marital status), X_8(profession), X_9(spleen disease), X_{10}(heart disease), X_{11}(growth retardation), X_{12}(osteoporosis), X_{13}(hepatitis), X_{14}(final state of view), X_{15}(height), X_{16}(weight) , X_{17}(religion), X_{18}(smoking), X_{19}(the number of family members to which the affected individual belongs), X_{20} (income), X_{21}(age of the father), X_{22}(age of the mother).

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

Table 1: The adjusted R-square values for the model fit depending on the real data

		SSIR	SIR- LASSO	RSIR
Model Fit	*Linear*	0.74	0.88	0.77
	Quadratic	0.84	0.90	0.88
	Cubic	0.90	0.92	0.90
	Quartic	0.90	0.92	0.90

Table 1: Shows the superiority of the SIR-LASSO method, which had the largest values than the rest of the methods when we used the criterion R-square and in all the models used in the analysis, and this shows the superiority of the SIR-LASSO method.

Table 2: The prediction error of the cubic fit for the studied methods depending on the real data

Methods	Prediction error
SSIR	**1.1028**
SIR-LASSO	**0.7417**
RSIR	**0.8900**

Table 2 : Shows the error criterion for the purpose of comparing the real data analysis methods, and we note that the SIR-LASSO method contains the least prediction error, and this proves its superiority over the rest of the methods.

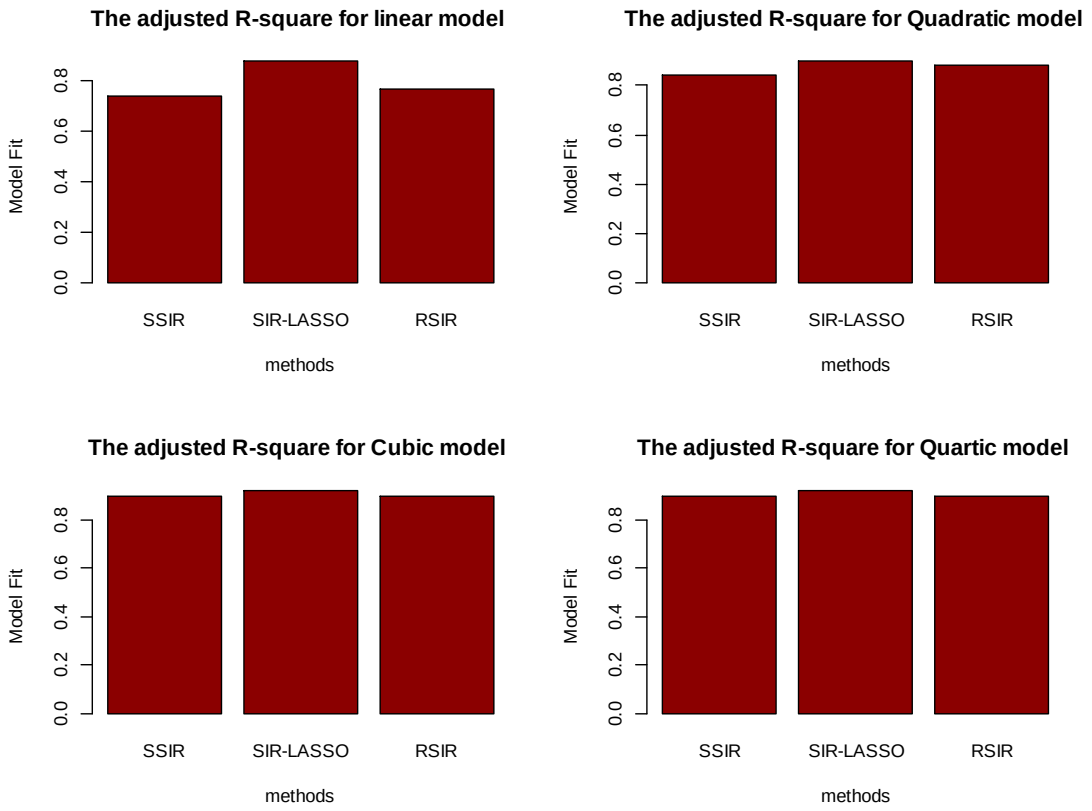

Figure1: Model fit for the considered methods According to the values of the adjusted R square.

Figure1: Shows the accuracy of method SIR-LASSO compared to the methods which used in analyzing real data for prediction error as shown in the graph.

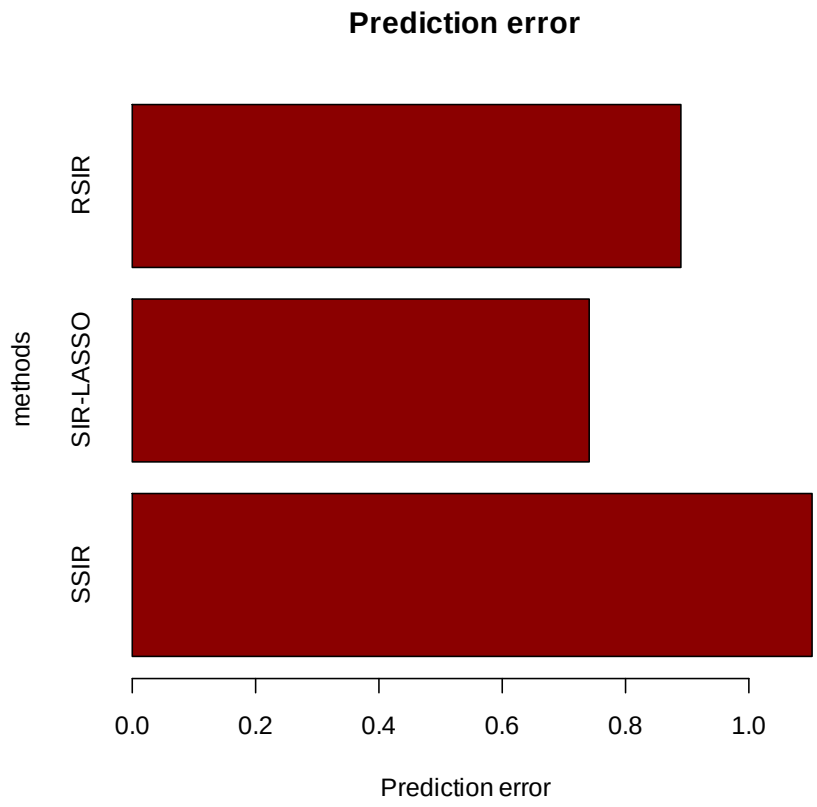

Figure 2: PE for the considered methods.

Figure 2: Shows for criterion PE the superiority of the method SIR-LASSO over the rest of the methods under study in analyzing real data, as shown in the graph.

5. Conclusion

In this research, we used the methods SSIR, RSIR and SIR-LASSO, when compared the results obtained, Showed the superiority of the method proposed by Lin et al. (2018) SIR-LASSO on the rest for the methods., and this leads us to a recommendation when analyzing the data. High dimensions using SIR-LASSO method, because it gives better 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). A comparative study for robust canonical correlation methods. Journal of Statistical Computation and Simulation, 83(4), 692–720.

3. Bellman, R. E. (1961). Adaptive Control Processes. Princeton University Press, Princeton, New Jersey.

4. Bickel PJ, Ritov Y, and Tsybakov AB. Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics, 37(4): 1705–1732, 2009. [Google Scholar]

5. Candes E and Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6) :2313–2351, 2007. [Google Scholar]

6. Cook, R. (1998). Ideas for Studying Regressions Through Graphics. Wiley, New York.

7. Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. The Annals of Statistics, 32(3), 1062–1092.

8. Efron B, Hastie T, Johnstone I, and Tibshirani R. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. [Google Scholar]

9. Friedman J, Hastie T, and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1): 1, 2010. [PMC free article] [PubMed] [Google Scholar]

10. 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.

11. Li, K.-C. (1991). Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86(414), 316–327.

12. Li, L. (2007). Sparse sufficient dimension reduction. Biometrika, 94(3), 603–613.

13. Li, L., and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. Technometrics, 48(4), 503–510.

14. Li, L., and Yin, X. (2008). Sliced inverse regression with regularizations. Biometrics, 64(1), 124–131.

15. Lin Q, Zhao Z, and Liu JS. On consistency and sparsity for sliced inverse regression in high dimensions. *arXiv* preprint arXiv:1507.03895, 2015. [Google Scholar]

16. Lin, Q., Zhao, Z., and Liu, J. S. (2018). Sparse sliced inverse regression via lasso. Journal of the American Statistical Association, 114(528), 1726–1739.

17. 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.

19. Ni, L., Cook, R. D., and Tsai, C.-L. (2005). A note on shrinkage sliced inverse regression. Biometrika, 92(1), 242–247.

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

الملخص

تحظى الـ SDR اهتماماً كبيراً في الانحدارات عالية الأبعاد، ولنفترض أنY هو متغير استجابة، وان $\mathrm{X} = (\mathrm{x}_1, \dots, \mathrm{x}_p)^T$ هو متنبئ الأبعاد، وبدون افتراض أي نموذج حدودي، إذ أن الفكرة الرئيسية لـ SDR هي استبدال X مع متعامد منخفض الأبعاد $P_S\mathrm{X}$ الى S، مع الاحتفاظ بالمعلومات عن توزيع Y | X . وان الهدف الرئيسي من إجراء SDR هو لإيجاد الفضاء الجزئي المركزي $S_{Y|\mathrm{X}}$ وهو يمثل تقاطع جميع المسافات الفرعية$Y \perp\!\!\!\perp \mathrm{X}|P_S\mathrm{X}$ ، حيث ان$\perp\!\!\!\perp$ تدلُّ على الاستقلال، لذلك فأنّ $P_\beta\mathrm{X}$ هي تمثل مقتطفات لجميع المعلومات من X حول Y ، حيث β هي الأساس لـ $S_{Y|\mathrm{X}}$ . .(Cook, 1998)

وتوجد هناك عدة طرق مقترحة لإيجاد $S_{Y|\mathrm{X}}$ ، وإحدى الطرق المعروفة هي الـ SIR من قبل (Li, 1991) اذ يتم تطبيق هذه الطرق في العديد من المجالات بما فيها الاقتصادية والمعلوماتية الحيوية، وان طريقة الـ SIR تواجه صعوبات خصوصاً في تفسير التقديرات الناتجة عن SIR، وذلك بسبب إنتاجه تركيبات خطية من جميع المتنبئين الأصليين، ومن اجل تحسين التفسير الخاص بـ SIR فانه لابد من تقليل عدد المعاملات غير الصفرية، والتي تعتبر غير مهمة في اتجاهات الـ SIR.

الهدف من دراستنا هو تقليل عدد المعاملات غير الصفرية في اتجاهات SIR للحصول على تفسير افضل. من خلال دمج بعض طرق Regularization مع طريقة SIR لإنتاج تقديرات متفرقة ودقيقة.

في هذا البحث سوف نستخدم طرقاً لدمج عمل SIR مع طريقة LASSO والمقترحة من قبل

RSIR (Li and Yin, 2008), SIR–LASSO(Lin et al, 2018), SSIR(Ni et al, 2005)

في تحليل عينة لمرضى السكري.

جمهورية العراق

وزارة التعليم العالي والبحث العلمي

جامعة القادسية

كلية الإدارة والاقتصاد

قسم الإحصاء

تحليل بيانات مرض السكري باستخدام الأساليب القائمة على SIR

بحث مقدم من قبل الطالب

محمد عبد الكاظم حصيل

أشراف

الأستاذ الدكتور

علي جواد كاظم الكناني

١٤٤٣ هـ

٢٠٢١ م