STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **11**, March 2023, pp 422–444.
Published online in International Academic Press (www.IAPress.org)

Comparison of Subspace Dimension Reduction Methods in Logistic Regression

Saeed Heydari , Mahmoud Afshari*, Saeed Tahmasbi , Morad Alizadeh

Department of Statistics, Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf University, Bushehr, Iran

Abstract Regression models are very useful in describing and predicting real world phenomena. The Logistic regression is an extremely robust and flexible method for dichotomous classification prediction. This model is a classification model rather than regression model. When the number of predictors in regression models is high, data analysis is difficult. Dimension reduction has become one of the most important issues in regression analysis because of its importance in dealing with problems with high-dimensional data. In this paper, the methods of diminishing the dimension of variables in logistic regression, which include the estimation of central subspace based on the inverse regression, the likelihood acquisition method and principal component analysis are considered. Using a real data associated with the dental problems the Logistic regression is fitted and the correct classification of the data computed. At the end, The simulation study is presented to compare the sufficient dimension reduction methods with each other. In the simulation, MATLAB software is used and the Programs are attached at the end of the article in appendix.

Keywords Dimension reduction, Likelihood acquired direction, Sliced average variance estimation, Sliced inverse regression.

AMS 2010 subject classifications 62G30, 62M20

DOI:10.19139/soic-2310-5070-1303

1. Introduction

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.In other word it is one of the most important statistics topics that examines the relationship between a response variable Y and a variable $\mathbf{X} = (X_1, ..., X_p)^T$. In parametric regression, the link function is a simple algebraic function of $\mathbf{X}$, and least squares or maximum likelihood methods (among others) can be applied in order to find the best global fit. In full generality, the goal of a regression is to infer about the conditional distribution of the univariate response variable Y given $p \times 1$ vector of predictors $\mathbf{X}$. The logistic regression model, is a statistical method for binary classification that can be generalized to multiclass classification which is very easy to realize and achieves very good performance with linearly separable classes. Logistic regression is a simple and more efficient method for binary and linear classification problems. When the dependent variable has two categories, then it is a binary logistic regression. When the dependent variable has more than two categories, then it is a multinomial logistic regression. When the dependent variable category is to be ranked, then it is an ordinal logistic regression (Afshari et. al. 2017). Logistic regression solves many problems faced in freemium product development that linear regression cant, because rather than predicting a numerical value, it predicts a discrete, dichotomous value . For this reason, logistic regression might more accurately be called logistic classification.In regression models, the number of predictor

*Correspondence to: Mahmoud Afshari (Email: afshar@pgu.ac.ir). Department of Statistics, Persian Gulf University, Bushehr, Iran.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

variables is a main problem. Today, advances in technology have led to the creation of data with many explanatory variables, which is difficult to analyze, so dimensional reduction methods can be a reasonable method in this area. Reducing the number of variables without losing information has met the purpose of this study. The problem of dimensionality also provides the ability to use regression graphs (Afshari, 2017).

Since in many statistical applications the dimension p is large, the statistical analysis becomes difficult. Therefore, it is very important to reduce the dimension p without much loss of information on regression. One of the proposed solution in dealing with this problem is to reduce the number of these variables. This has been achieved through the development of sufficient dimension reduction methods. The goal is to find the appropriate subspace with a small dimension. Since the subspace depends on parameters, it should be estimated. Real-world data, such as speech signals, digital photographs, or FMRI scans, usually has a high dimensionality. In order to handle such real-world data adequately, its dimensionality needs to be reduced. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data (Fukunaga, 1990). Dimensionality reduction is important in many domains, since it mitigates the curse of dimensionality and other undesired properties of high-dimensional spaces (As a result, dimensionality reduction facilitates, among others, classification, visualization, and compression of high-dimensional data, see for instant, Jimenez and Landgrebe, 1997, for more information).
Dimension reduction for regression, as pioneered by such authors as Duan and Li (1991), Li (1991, 1992), Cook and Weisberg (1991), and Cook (1994, 1998), is aimed at reducing the dimension of a vector-valued predictor $\mathbf{X}$, while preserving its regression relation with a real-valued response Y. Research into dimension reduction has gained considerable momentum in recent years due to the rapidly increasing data volume and dimension, which demand preprocessing techniques to reduce their scope. Li, et. al. (2003) considered the analysis of multivariate response data with multivariate regressors. Methods for reducing the dimensionality of response variables developed, with the goal of preserving as much regression information as possible. Yin and Cook (2005), proposed a general dimension-reduction method that combines the ideas of likelihood, correlation, inverse regression and information theory. Yin et. al. (2008) proposed a dimension reduction method for estimating the directions in a multiple-index regression based on information extraction and showed that under the assumption of elliptical predictors, the estimation of multiple-index regressions can be decomposed into successive single-index estimation problems. Zhu et. al. (2010), offered a complete methodology of cumulative slicing estimation to sufficient dimension reduction and proposed three methods that are termed, respectively, as cumulative mean estimation, cumulative variance estimation, and cumulative directional regression. Luo et. al. (2014) introduced a new sufficient dimension reduction framework that targets a statistical functional of interest, and proposed an efficient estimator for the semiparametric estimation problems of this type.
Li (1991) and Duan and Li (1991) introduced a link-free and distribution-free method for estimation of central subspace called Sliced inverse regression (SIR) method to estimate the subspace. The basic principle of SIR is to reverse the role of Y and X and to study the geometric property of the first conditional moment $E(\mathbf{X}|Y)$. In fact, the Sliced Inverse Regression (SIR) is an effective method for dimension reduction in high-dimensional regression problems. The original method, however, requires the inversion of the predictors covariance matrix. Hsing and Carroll (1992) have derived the asymptotic properties of this procedure for the special case where each slice contains only two observations. Li et al. (1999) extended sliced inverse regression (SIR) of Li (1991) to the setting which allows for censoring in the data. Bura and Cook (2001), considered the assumptions on the predictor distribution, under which the chi-squared test was proved to apply, are relaxed, and the result is extended. A general weighted chi-squared test that does not require normal regressors for the dimension of a regression is given. Simulations show that the weighted chi-squared test is more reliable than the chi-squared test when the regressor distribution digresses from normality significantly. The term Sliced refers to the fact that a slicing is realized on the response variable Y to facilitate the estimation of the inverse conditional expectation. Li and Yin (2008), proposed a regularized SIR approach based on the least-squares formulation of SIR. An alternating least-squares algorithm developed, to enable SIR to work when the number of predictors, p, exceeds the sample size, n, and highly correlated predictors.

Cook and Weisberg (1991) proposed another method that use second conditional moment $E(\mathbf{X}|Y)$ names sliced average variance estimation (SAVE) for estimating the central subspace. This method has good performance in finding quadratic forms and fail to find linear trend and in this method, the response variable should be discrete or categorical. Ye and Weiss (2003) introduced the broad classes of dimension reduction candidate matrices, and distinguished estimators of the matrices from the matrices themselves. Also they proposed bootstrap methodology to select among candidate matrices, estimators and dimension, and in particular we investigate linear combinations of different methods. Zhu and Zhu (2007), used the kernel method to estimate the SAVE and proved that this estimator is both asymptotically normal and root n consistent. Examples and real data presented for illustrating our method. Li and Wang (2007) introduced the directional regression method to estimate the central subspace, which combines the methods of dimension reduction based on the first two conditional moments. Zhu et. al. (2007) presented a further investigation for the hybrid methods of inverse regression-based algorithms and a set of simulations for several typical models were carried out to guide the selection of coefficient in the hybrids. Lue (2008), considered the SAVE for censored data. Based on the weight adjustment, he developed the modification of sliced average variance estimation for estimating the lifetime central subspace without requiring a prespecified parametric model. The simulation results reported and comparisons made with the sliced inverse regression of Li et al. (1999). Cook and Forzani (2009) introduced the likelihood acquired directions methods, which is more desirable than other methods. Also, principal component analysis is one of the oldest methods for dimension reduction. In this paper, we introduce these methods in details and use them in clinical data and compare them with each other.
Sliced inverse regression (SIR) and sliced average variance estimator (SAVE), which are based on the first two conditional moments $E(\mathbf{X}|Y)$ and $E(\mathbf{X}\mathbf{X}^T|Y)$, are among the most commonly used dimension reduction estimators. They have well-known limitations, however. In particular, SIR is known to fail when the response surface is symmetric about the origin, whereas SAVE is not very efficient in estimating monotone trends for small to moderate sample sizes. Li and Wang (2007) introduced a natural and simple principle for dimension reduction, called directional regression (DR), that synthesizes the dimension reduction methods based on first two conditional moments and achieves substantial improvement in accuracy. They developed the asymptotic distribution of the DR estimator, and from that a sequential test procedure to determine the dimension of the central space. Like contour regression, DR is derived from empirical directions, but achieves higher accuracy and requires substantially less computation. Yu et al. (2014), extended directional regression to a general family of estimators via the notion of general empirical directions and developed a new methodology for nonlinear dimension reduction.
Principal component analysis (PCA) is probably the oldest and certainly the most popular technique for computing lower-dimensional representations of multivariate data. The technique is linear in the sense that the components are linear combinations of the original variables (features), but non-linearity in the data is preserved for effective visualization. The technique can be presented as an iterative computation of the direction of highest variation followed by projection onto the perpendicular hyperplane. This quickly provides a few perpendicular directions that account for the majority of the variation in the data, giving a low dimensional representation of the data. A complete set of principal components can be viewed as a rotation in the original variable space. See, for example Joliffe (1986) for a comprehensive treatment and history of principal component analysis.
The rest of this paper is organized as follows. In Section 2, the main definitions on the dimension reduction subspace are given. In Section 3, the estimation methods of central subspace based on inverse regression consist of sliced inverse regression, sliced average variance estimation and directional regression are presented. In Section 4, the maximum likelihood estimation of central subspace which does not depend on the condition of being normal is obtained. In Section 5, the application of the presented method using a real example associated with the dental problems and the correct classification of the data are computed. In Section 6, the simulation study is presented to compare the sufficient dimension reduction methods with each other. Concluding remarks are given in Section 7.

2. Dimension reduction subspace

Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. Ideally, the reduced representation should have a dimensionality that corresponds to the intrinsic dimensionality of the data.In this section we want to discuss about Dimension reduction subspace:

When we want to reduce the dimension of predictor variables from p to a smaller value d without loss of information, the equation

$$Y = f(\boldsymbol{\beta}_1^T\mathbf{X}, \boldsymbol{\beta}_2^T\mathbf{X}, ..., \boldsymbol{\beta}_d^T\mathbf{X}, \boldsymbol{\epsilon}), \tag{1}$$

makes it possible. In equation (1), $\boldsymbol{\beta}_j,\ j = 1, ..., d$, is an unknown parameter, f is an unknown function on $\mathbb{R}^{d+1}$ and also $\boldsymbol{\epsilon}$ is independent of random vector $\mathbf{X}$. When the equation (1) holds, the $\boldsymbol{\beta}^T\mathbf{X}$ gives all information we need to know about the variable Y.

Definition 1
The term $\boldsymbol{\beta}^T\mathbf{X}$ is sufficient reduction for the regression Y on $\mathbf{X}$ if:

$$Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}. \tag{2}$$

The equation (2) means that given $\boldsymbol{\beta}^T\mathbf{X}$, the variables Y and $\mathbf{X}$ are independent. Thus the subspace spanned by the columns of $\boldsymbol{\beta}$, $S(\boldsymbol{\beta})$, is called a dimension reduction subspace (DRS) for $Y|\mathbf{X}$. Since (2) holds for $\boldsymbol{\beta} = \mathbf{I}_p$, so DRS is not necessarily unique. In applications we prefer to use graphical tools to show data. For this purpose we reduce a dimension of subspace and thus the idea of minimum dimension reduction subspace is appeared.(Cook, 1998).

Definition 2
The subspace S is called minimum dimension reduction subspace (min DRS) for regression Y on $\mathbf{X}$ if the following relations hold:
1) The subspace S be a DRS.
2) For each S_{drs}, $dimS \leq dimS_{drs}$.

Note that, the minimum DRS is also not necessarily unique. The following example shows this subject.

Example 1
Let $p = 2$ and $\mathbf{X} = (x_1, x_2)$ be uniformly distributed on the unit circle, then $||\mathbf{X}|| = 1$. Set $Y|\mathbf{X} = x_1^2 + \epsilon$, where ϵ is an independent error. Since $x_1^2 + x_2^2 = 1$, either x_1 or x_2 has full information about the $Y|\mathbf{X}$, then

$$Y|\mathbf{X} = x_1^2 + \epsilon = (1 - x_2^2) + \epsilon.$$

Thus $S((1,0)^T)$ and? $S((0,1)^T)$ are both min DRS (Cook, 1998).

Minimum dimension reduction subspaces are not unique and one regression has several min DRS with the same dimensions, this dimension is called structural dimension (d) of the regression. When the subspace is not unique, we face with problem in estimating it, so the idea of a central DRS would be helpful (Cook, 1998).

Definition 3
(Cook, 1998). The subspace S is called minimum dimension reduction subspace (min DRS) for regression Y on $\mathbf{X}$ if the following relations hold:
1) The subspace S be a DRS.
2) For each S_{drs}, $S \subset S_{drs}$.

The central DRS is denoted by $S_{Y|\mathbf{X}}$. The central subspace exists if and only if the intersection of all DRS be a DRS, i.e. $S_{Y|\mathbf{X}} = \bigcap S_{drs}$. Since the central subspace depends on the parameter, it should be estimated. In the next sections, we explain about the estimation methods of central subspace based on inverse regression.

3. The estimation of central subspace based on inverse regression

Now we want to Estimate centeral subspace based on inverse regression Based on three subsection as below:

3.1. Sliced inverse regression

Li (1991), introduced sliced inverse regression method for estimation of central subspace. Let $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - E(\mathbf{X}))$ and $\mathbf{\Sigma}$ be a covariance matrix of vector $\mathbf{Z}$, we use inverse regression $\mathbf{Z}|Y$ to estimate central subspace. Recall that in this method, the response variable should be discrete or categorical.

Theorem 1
Let $\boldsymbol{\eta}$ be a basis for $S_{Y|\mathbf{Z}}$, $\mathbf{P}_{\boldsymbol{\eta}}$ be a projection matrix on $S_{Y|\mathbf{Z}}$ and $E(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z}) = \mathbf{P}_{\boldsymbol{\eta}}\mathbf{Z}$, then we have

$$E(\mathbf{Z}|Y) = \mathbf{P}_{\boldsymbol{\eta}}E(\mathbf{Z}|Y), \tag{3}$$

and

$$E(\mathbf{Z}|Y) \in S_{Y|\mathbf{Z}}, \tag{4}$$

or equivalently

$$S_{E(\mathbf{Z}|Y)} \subset S_{Y|\mathbf{Z}}. \tag{5}$$

Proof
To proof (3), we know that $Y \perp\!\!\!\perp \mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z}$ and $\mathbf{P}_{\boldsymbol{\eta}} + \mathbf{Q}_{\boldsymbol{\eta}} = \mathbf{I}_p$. Therefore

$$\begin{aligned} E(\mathbf{Z}|Y) &= E(E(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z}, Y)|Y) = E(E(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})|Y) = E(E((\mathbf{P}_{\boldsymbol{\eta}} + \mathbf{Q}_{\boldsymbol{\eta}})\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})|Y) \\ &= E(E((\mathbf{P}_{\boldsymbol{\eta}}\mathbf{Z} + \mathbf{Q}_{\boldsymbol{\eta}}\mathbf{Z})|\boldsymbol{\eta}^T\mathbf{Z})|Y) = E(\mathbf{P}_{\boldsymbol{\eta}}E(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z}) + \mathbf{Q}_{\boldsymbol{\eta}}E(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})|Y) \\ &= E((\mathbf{P}_{\boldsymbol{\eta}}\mathbf{P}_{\boldsymbol{\eta}}\mathbf{Z} + \mathbf{Q}_{\boldsymbol{\eta}}\mathbf{P}_{\boldsymbol{\eta}}\mathbf{Z})|Y) = E((\mathbf{P}_{\boldsymbol{\eta}} + \mathbf{Q}_{\boldsymbol{\eta}})\mathbf{P}_{\boldsymbol{\eta}}\mathbf{Z}|Y) = E(\mathbf{P}_{\boldsymbol{\eta}}\mathbf{Z}|Y) \\ &= \mathbf{P}_{\boldsymbol{\eta}}E(\mathbf{Z}|Y). \end{aligned}$$

Since $\mathbf{P}_{\boldsymbol{\eta}}$ is a projection matrix on $S_{Y|\mathbf{Z}}$ and $E(\mathbf{Z}|Y) = \mathbf{P}_{\boldsymbol{\eta}}E(\mathbf{Z}|Y)$, then we have$E(\mathbf{Z}|Y) \in S_{Y|\mathbf{Z}}$. □

According to theorem 1, we estimate $E(\mathbf{Z}|Y) \in S_{Y|\mathbf{Z}}$ using the following algorithm.

Algorithm 1
1. Let $\mathbf{Z} = \hat{\mathbf{\Sigma}}_{xx}{}^{-1/2}(\mathbf{X} - \bar{\mathbf{X}})$ be standard form of vector $\mathbf{X}$ where $\hat{\mathbf{\Sigma}}_{xx}$ is covariance matrix.
2. We partition range of Y into H slices $I_1, ..., I_H$. Let the proportion of the $Y_i, i = 1, 2, \ldots, n$, that falls in slice r be $\hat{p}_r$, so

$$\hat{p}_r = 1/n \sum_{i=1}^{n} I_r(Y_i), \tag{6}$$

where $I_r(Y_i)$ is indicator function of Y_i in each slice.
3. Compute the sample mean of $\hat{\mathbf{Z}}$ within each slice:

$$\bar{\mathbf{Z}}_r = \frac{\sum_{i=1}^{n} \hat{\mathbf{Z}}_i I_r(Y_i)}{\sum_{i=1}^{n} I_r(Y_i)}. \tag{7}$$

The slice mean converges almost surely to the population mean:

$$E(\mathbf{Z}|Y = h) \in S_{E(\mathbf{Z}|Y)} \subset S_{Y|\mathbf{Z}} = \mathbf{\Sigma}^{1/2}S_{Y|\mathbf{X}}.$$

4. Find the eigenvalues and eigenvectors of the weighted sample covariance matrix:

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{h=1}^{H} n_h \bar{\mathbf{Z}}_h \bar{\mathbf{Z}}_h^T.$$

5. Let $d = dim[S_{E(\mathbf{Z}|Y)}]$, the SIR estimation of subspace $S_{E(\mathbf{Z}|Y)}$ is attained by

$$\hat{S}_{E(\mathbf{Z}|Y)} = S(\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\eta}}_d),$$

where $\boldsymbol{\eta}_i$ is the ith largest eigenvector of $\hat{\mathbf{V}}$. The SIR estimate of $S_{Y|\mathbf{X}}$ is

$$\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{S}_{E(\mathbf{Z}|Y)} = S(\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_d).$$

The proposed algorithm has a good performance for determining the linear trend. Assume that response variable is binary and let $\boldsymbol{\Sigma}_j = var(\mathbf{Z}|Y = j)$, $f = P(Y = 1)$, $\boldsymbol{\nu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ and $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0$.(Li, 1991).

3.2. Sliced average variance estimation

Cook and Weisberg (1991) proposed another method names sliced average variance estimation(SAVE) for estimating the central subspace. This method has good performance in detecting a quadratic forms. Again recall that in this method, the response variable should be discrete or categorical.

Theorem 2
Let $\boldsymbol{\eta}$ be a basis for $S_{Y|\mathbf{Z}}$, and
$1) E(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z}) = \mathbf{P}_{\boldsymbol{\eta}}\mathbf{Z}$.
$2) Var(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z}) = \mathbf{Q}_{\boldsymbol{\eta}}$,
where $\mathbf{P}_{\boldsymbol{\eta}}$ is a projection operator on $S_{Y|\mathbf{Z}}$ and $\mathbf{Q}_{\boldsymbol{\eta}} = \mathbf{I_p} - \mathbf{P}_{\boldsymbol{\eta}}$, then we have:

$$\boldsymbol{\Sigma}_{\mathbf{Z}|Y} = \mathbf{Q}_{\boldsymbol{\eta}} + \mathbf{P}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{Z}|Y}\mathbf{P}_{\boldsymbol{\eta}}, \tag{8}$$

and

$$S(\mathbf{I_p} - \boldsymbol{\Sigma}_{\mathbf{Z}|Y}) \subset S_{Y|\mathbf{Z}}. \tag{9}$$

Proof
Consider that

$$\boldsymbol{\Sigma}_{\mathbf{Z}|Y} = E[var(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z}, Y)|Y] + var[E(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z}, Y)|Y].$$

Since $Y \perp\!\!\!\perp \mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z}$, then we have

$$\boldsymbol{\Sigma}_{\mathbf{Z}|Y} = E[var(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})|Y] + var[E(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})|Y].$$

According to the first assumption of the theorem, we have

$$\boldsymbol{\Sigma}_{\mathbf{Z}|Y} = E[var(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})|Y] + \mathbf{P}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{Z}|Y}\mathbf{P}_{\boldsymbol{\eta}}. \tag{10}$$

By using the second assumption,

$$\boldsymbol{\Sigma}_{\mathbf{Z}|Y} = \mathbf{Q}_{\boldsymbol{\eta}} + \mathbf{P}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{Z}|Y}\mathbf{P}_{\boldsymbol{\eta}}. \tag{11}$$

To estimate the relation (9), consider that

$$\mathbf{I_p} - \boldsymbol{\Sigma}_{\mathbf{Z}|Y} = \mathbf{P}_{\boldsymbol{\eta}} + \mathbf{P}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{Z}|Y}\mathbf{P}_{\boldsymbol{\eta}}.$$

Thus

$$\begin{aligned}\mathbf{P}_{\eta}(\mathbf{I_p} - \boldsymbol{\Sigma}_{\mathbf{Z}|Y})\mathbf{P}_{\eta} &= \mathbf{P}_{\boldsymbol{\eta}}\mathbf{P}_{\boldsymbol{\eta}}\mathbf{P}_{\boldsymbol{\eta}} + \mathbf{P}_{\boldsymbol{\eta}}\mathbf{P}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{Z}|Y}\mathbf{P}_{\boldsymbol{\eta}}\mathbf{P}_{\boldsymbol{\eta}} = \mathbf{P}_{\boldsymbol{\eta}} + \mathbf{P}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_{\mathbf{Z}|Y}\mathbf{P}_{\boldsymbol{\eta}} \\ &= (\mathbf{I_p} - \boldsymbol{\Sigma}_{\mathbf{Z}|Y}).\end{aligned}$$

□

According to the assumptions 1 and 2 of theorem, we get

$$span\{E(\mathbf{I}_p - var(\mathbf{Z}|Y))^2\} \subseteq S_{Y|\mathbf{Z}}, \tag{12}$$

which is a basis for SAVE.
The SAVE algorithm is purposed in 5 steps to estimate $E(\mathbf{I}_p - var(\mathbf{Z}|Y))^2$.

Algorithm 2
(Cook and Weisberg, 1991)

1. Let $\mathbf{Z} = \hat{\mathbf{\Sigma}}_{xx}{}^{-1/2}(\mathbf{X} - \bar{\mathbf{X}})$ be standard form of vector $\mathbf{X}$ where $\hat{\mathbf{\Sigma}}_{xx}$ is covariance matrix.

2. We divide the range of Y in to H slices and compute covariance matrix of $\hat{\mathbf{V}}_r$ for each slice. Note that $\hat{\mathbf{V}}_r$ is the estimation of $Var(\mathbf{Z}|\tilde{Y})$.

3. Let f_r is the sample fraction of observations in each slice and compute:

$$\mathbf{M} = \sum_{h=1}^{H} f_r(\mathbf{I} - \hat{\mathbf{V}}_r)^2.$$

4. Consider the jth predictor of SAVE, then

$$S_j = \boldsymbol{\eta}_j^T \hat{\mathbf{Z}}_i \quad j = 1, ..., p, \quad i = 1, .., n,$$

where $\boldsymbol{\eta}_j$denote the eigenvector corresponding to the jth-largest eigenvalue of $\mathbf{M}$.

5. Finally the SAVE estimate of $S_{Y|\mathbf{X}}$ is obtained by:

$$S(\hat{\mathbf{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_1, ..., \hat{\mathbf{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_d).$$

3.3. Directional regression

Li and Wang (2007) introduced the directional regression method to estimate the central subspace, which combines the methods of dimension reduction based on the first two conditional moments.

Lemma 1
Suppose that $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{Z}$ be random vectors, then following expressions are equivalent

$$\mathbf{U} \perp\!\!\!\perp \mathbf{W}|(\mathbf{Z}, \mathbf{V}) \qquad \mathbf{U} \perp\!\!\!\perp \mathbf{V}|\mathbf{Z} \tag{13}$$

$$\mathbf{U} \perp\!\!\!\perp \mathbf{V}|(\mathbf{Z}, \mathbf{W}) \qquad \mathbf{U} \perp\!\!\!\perp \mathbf{W}|\mathbf{Z} \tag{14}$$

$$\mathbf{U} \perp\!\!\!\perp (\mathbf{V}, \mathbf{W})|\mathbf{Z} \tag{15}$$

Theorem 3
Suppose that $\boldsymbol{\nu} \in \mathbb{R}^p$, $\boldsymbol{\nu} \perp S_{Y|\mathbf{Z}}$ and
1) $E(\boldsymbol{\nu}^T\mathbf{Z}|\mathbf{P}_{\boldsymbol{\eta}}\mathbf{Z})$ is a linear function of $\mathbf{Z}$.
2) $Var(\boldsymbol{\nu}^T\mathbf{Z}|\mathbf{P}_{\boldsymbol{\eta}}\mathbf{Z})$ be constant.
Then for each $(Y, \tilde{Y})$ columnar space $2I_p - A(Y, \tilde{Y})$ in $S_{Y|\mathbf{Z}}$ Placed.

Proof
Using lemma 1, $(\mathbf{Z}, Y) \perp\!\!\!\perp (\tilde{\mathbf{Z}}, \tilde{Y})$ results that

$$\mathbf{Z} \perp\!\!\!\perp \tilde{\mathbf{Z}}|(Y, \tilde{Y}), \quad \mathbf{Z} \perp\!\!\!\perp \tilde{Y}|Y, \quad \tilde{\mathbf{Z}} \perp\!\!\!\perp Y|\tilde{Y},$$

then $A(Y, \tilde{Y})$ extends as follows:

$$\begin{aligned} A(Y, \tilde{Y}) &= E(\mathbf{Z}\mathbf{Z}^T - \mathbf{Z}\tilde{\mathbf{Z}}^T - \tilde{\mathbf{Z}}\mathbf{Z}^T + \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T|Y, \tilde{Y}) \\ &= E(\mathbf{Z}\mathbf{Z}^T|Y) - E(\mathbf{Z}|Y)E(\tilde{\mathbf{Z}}^T|\tilde{Y}) - E(\tilde{\mathbf{Z}}|\tilde{Y})E(\mathbf{Z}^T|Y) + E(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T|\tilde{Y}). \end{aligned} \tag{16}$$

Now we just show that

$$S_{Y|\mathbf{Z}}^{\perp} \subset \{span(\mathbf{I}_p - A(Y, \tilde{Y}))\}^{\perp}. \tag{17}$$

Suppose that $\boldsymbol{\nu}_y \in S^{\perp}_{Y|\mathbf{Z}}$. According to assumption 1:

$$E(\boldsymbol{\nu}^T\mathbf{Z}|\mathbf{PZ}) = \boldsymbol{\alpha}^T\mathbf{PZ}, \qquad \exists\boldsymbol{\alpha} \in \mathbb{R}^p \tag{18}$$

then

$$\begin{aligned} 0 \le \boldsymbol{\alpha}^T\mathbf{P}\boldsymbol{\alpha} = (\boldsymbol{\alpha}^T\mathbf{PZ})(\boldsymbol{\alpha}^T\mathbf{PZ})^T = E(\boldsymbol{\nu}^T\mathbf{Z}|\mathbf{PZ})\mathbf{Z}^T\mathbf{P}\boldsymbol{\alpha} \\ = E(\boldsymbol{\nu}^T\mathbf{Z}\mathbf{Z}^T\mathbf{P}\boldsymbol{\alpha}|\mathbf{PZ}). \end{aligned}$$

Now with taking expectation value from both sides, we have

$$0 \le \boldsymbol{\alpha}^T\mathbf{P}\boldsymbol{\alpha} = E(\boldsymbol{\nu}^T\mathbf{Z}\mathbf{Z}^T\mathbf{P}\boldsymbol{\alpha}) = \boldsymbol{\nu}^T\mathbf{P}\boldsymbol{\alpha}.$$

Since $\boldsymbol{\nu}_y \in S^{\perp}_{Y|\mathbf{Z}}$ and $\mathbf{P}$ is a projection matrix on $S_{Y|\mathbf{Z}}$, then

$$\boldsymbol{\nu}^T\mathbf{P}\boldsymbol{\alpha} = 0, \tag{19}$$

So that $\boldsymbol{\alpha}^T\mathbf{P}\boldsymbol{\alpha} = 0$ and $\boldsymbol{\alpha}^T\mathbf{PZ} = 0$, and

$$E(\boldsymbol{\nu}^T\mathbf{Z}|\mathbf{PZ}) = 0. \tag{20}$$

According to assumption 2:

$$E((\boldsymbol{\nu}^T\mathbf{Z})^2|\mathbf{PZ}) = c + E^2(\boldsymbol{\nu}^T\mathbf{Z}|\mathbf{PZ}) = c. \tag{21}$$

Equal to conditional variance $\boldsymbol{\nu}^T\mathbf{Z}$ with condition $\mathbf{PZ}$. with taking expectation value from both side Phrase $c = \boldsymbol{\nu}^T\boldsymbol{\nu}$ Obtained. Therefore

$$E((\boldsymbol{\nu}^T\mathbf{Z})^2|\mathbf{PZ}) = \boldsymbol{\nu}^T\boldsymbol{\nu}.$$

We know that $Y \perp\!\!\!\perp \mathbf{Z}|\mathbf{PZ}$, then

$$E(\boldsymbol{\nu}^T\mathbf{Z}|Y) = E[E(\boldsymbol{\nu}^T\mathbf{Z}|\mathbf{PZ})|Y] = 0,$$

and

$$E[(\boldsymbol{\nu}^T\mathbf{Z})^2|Y] = E\{E(\boldsymbol{\nu}^T\mathbf{Z})^2|\mathbf{PZ})|Y\} = \boldsymbol{\nu}^T\boldsymbol{\nu}.$$

By replacing these relations in a relation (16) and considering that $(\mathbf{Z}, Y)$ and $(\tilde{\mathbf{Z}}, \tilde{Y})$ have same distribution, we get that

$$\boldsymbol{\nu}^T A(Y, \tilde{Y})\boldsymbol{\nu} = 2\boldsymbol{\nu}^T\boldsymbol{\nu}, \tag{22}$$

and

$$\boldsymbol{\nu}^T(2Ip - A(Y, \tilde{Y}))\boldsymbol{\nu} = 0. \tag{23}$$

□

From this theorem, Matrix column space can be as

$$\mathbf{G} = E[2I_p - A(Y, \tilde{Y})]^2, \tag{24}$$

an estimate for $S_{Y|\mathbf{Z}}$. Lee and Wang (2007) computed the discrete estimate for G as follows:

$$\begin{aligned} \hat{\mathbf{G}} = 2\sum E_n^2(\hat{\mathbf{Z}}\hat{\mathbf{Z}}^T - \mathbf{I}_p|Y \in J_h)\hat{p}_h + [\sum E_n(\hat{\mathbf{Z}}|Y \in J_h)E_n(\hat{\mathbf{Z}}^T|Y \in J_h)\hat{p}_h]^2 \\ + 2\sum E_n(\hat{\mathbf{Z}}^T|Y \in J_h)E_n(\hat{\mathbf{Z}}|Y \in J_h)\hat{p}_h \\ \times \sum E_n(\hat{\mathbf{Z}}|Y \in J_h)E_n(\hat{\mathbf{Z}}^T|Y \in J_h)\hat{p}_h, \end{aligned}$$

where

$$E_n(\hat{\mathbf{Z}}|Y \in J_h) = \frac{E_n[\hat{\mathbf{Z}}I(Y \in J_h)]}{E_n I(Y \in J_h)} = \frac{\sum_{i=1}^n \hat{\mathbf{Z}}I(Y_i \in J_h)}{\sum_{i=1}^n I(Y_i \in J_h)}.$$

Now a sufficient tool is provided to estimate $S_{Y|\mathbf{X}}$. Suppose that $\hat{\lambda}_1 \geq ... \geq \hat{\lambda}_p$ and $\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\eta}}_p$ are the eigenvalues and eigenvectors of $\hat{\mathbf{G}}$, respectively. To estimate $S_{Y|X}$ is used from d first eigenvector $\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\eta}}_d$. so that

$$\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_d \tag{25}$$

4. Likelihood acquired directions

The methods for estimating the central subspace that were introduced in the previous section are limited to the normality condition of $Y|X$. In this section, another method is proposed which is based on the likelihood function and also it doesn't necessarily to be normal. Cook and Forzani (2009) introduced the likelihood acquired approach based on the likelihood function. again for this method we divide continuous response to H slices.

Proposition 1
(Cook and Forzani, 2009). Suppose $\mathbf{X}|Y = y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y)$, $y \in S_Y$. Let $\boldsymbol{\beta}$ be a semi-orthogonal basis matrix for $S \subseteq \mathbb{R}^p$ and $(\boldsymbol{\beta}, \boldsymbol{\beta}_0) \in \mathbb{R}^{p\times p}$. Then S is a dimension reduction subspace if and only if the following two conditions are satisfied. For every $y \in S_Y$:

1. $(\boldsymbol{\beta}^T\mathbf{X}|Y = y) \sim N(\boldsymbol{\beta}^T\boldsymbol{\mu} + \boldsymbol{\beta}^T\boldsymbol{\Delta}\boldsymbol{\beta}\boldsymbol{\nu}_y, \boldsymbol{\beta}^T\boldsymbol{\Delta}_y\boldsymbol{\beta})$; for some amounts $\boldsymbol{\nu}_y \in \mathbb{R}^{dim(S)}$.
2. $\boldsymbol{\beta}_0^T\mathbf{X}|(\boldsymbol{\beta}^T\mathbf{X}, Y = y) \sim N(\mathbf{H}\boldsymbol{\beta}^T\mathbf{X} + (\boldsymbol{\beta}_0^T - \mathbf{H}\boldsymbol{\beta}^T)\boldsymbol{\mu}, \mathbf{D})$,

where $\mathbf{D} = (\boldsymbol{\beta}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}_0)^{-1}$ and $\mathbf{H} = (\boldsymbol{\beta}_0^T\boldsymbol{\Delta}\boldsymbol{\beta})(\boldsymbol{\beta}^T\boldsymbol{\Delta}\boldsymbol{\beta})^{-1}$.

Using the above proposition, we can obtain the likelihood function for the LAD model.

Lemma 2
Suppose that $\mathbf{B} \in \mathbb{R}^{p\times p}$ be a symmetric positive definite matrix, and $(\boldsymbol{\beta}, \boldsymbol{\beta}_0) \in \mathbb{R}^{p\times p}$, A full rank matrix with $\boldsymbol{\beta}^T\boldsymbol{\beta}_0 = 0$, then:

$$\boldsymbol{\beta}(\boldsymbol{\beta}^T\mathbf{B}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T + \mathbf{B}^{-1}\boldsymbol{\beta}_0(\boldsymbol{\beta}_0^T\mathbf{B}^{-1}\boldsymbol{\beta}_0)^{-1}\boldsymbol{\beta}_0^T\mathbf{B}^{-1} = \mathbf{B}^{-1}, \tag{26}$$

and consequently:

$$\mathbf{I}_p - \mathbf{P}^T_{\boldsymbol{\beta}(\mathbf{B})} = \mathbf{P}_{\boldsymbol{\beta}_0(\mathbf{B}^{-1})}. \tag{27}$$

In addition, if $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ be orthogonal, then:

$$(\boldsymbol{\beta}_0^T\mathbf{B}^{-1}\boldsymbol{\beta}_0)^{-1} = \boldsymbol{\beta}_0^T\mathbf{B}\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^T\mathbf{B}\boldsymbol{\beta}(\boldsymbol{\beta}^T\mathbf{B}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T\mathbf{B}\boldsymbol{\beta}_0 \tag{28}$$

$$-(\boldsymbol{\beta}_0^T\mathbf{B}^{-1}\boldsymbol{\beta}_0)(\boldsymbol{\beta}_0^T\mathbf{B}^{-1}\boldsymbol{\beta}) = (\boldsymbol{\beta}_0^T\mathbf{B}\boldsymbol{\beta})(\boldsymbol{\beta}^T\mathbf{B}\boldsymbol{\beta})^{-1} \tag{29}$$

$$\boldsymbol{\beta}_0(\boldsymbol{\beta}_0^T\mathbf{B}^{-1}\boldsymbol{\beta}_0)^{-1}\boldsymbol{\beta}_0^T = \mathbf{B} - \mathbf{B}\boldsymbol{\beta}(\boldsymbol{\beta}^T\mathbf{B}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T\mathbf{B} \tag{30}$$

$$|\boldsymbol{\beta}_0^T\mathbf{B}\boldsymbol{\beta}_0| = |\mathbf{B}||\boldsymbol{\beta}^T\mathbf{B}^{-1}\boldsymbol{\beta}| \tag{31}$$

(Rao, 1973)

Theorem 4
Under the LAD model when d is known and the normal assumption is hold, the MLE of $S_{Y|\mathbf{X}}$, maximizes over $S \in \mathcal{G}_{(d,p)}$ the log likelihood function

$$L_d(S) = -\frac{np}{2}(1 + log(2\pi)) + \frac{n}{2}log|\mathbf{P}_S\tilde{\boldsymbol{\Sigma}}\mathbf{P}_S|_0 - \frac{n}{2}log|\tilde{\boldsymbol{\Sigma}}|$$
$$-\frac{1}{2}\sum_{y=1}^{h} n_y log|\mathbf{P}_S\tilde{\boldsymbol{\Delta}}_y\mathbf{P}_S|_0 \tag{32}$$

Maximizes it, in which $|\mathbf{A}|_0$ is product of nonzero eigenvalues of symmetric and semi-definite $\mathbf{A}$ (Cook and Forzani, 2009).

Proof
Suppose that $\boldsymbol{\beta}$ is semi orthogonal matrix for $S_{Y|\mathbf{X}}$ Then, the likelihood logarithm is based on the distribution $(\boldsymbol{\beta}^T\mathbf{X}, \boldsymbol{\beta}_0^T\mathbf{X}|Y)$ In the form of

$$\begin{aligned}
L_d &= \sum_y log\{f(\boldsymbol{\beta}^T\mathbf{X}|Y)f(\boldsymbol{\beta}_0^T\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}, Y)\} \\
&= -\frac{np}{2}log(2\pi) - \frac{n}{2}log|\mathbf{D}| - \frac{1}{2}\sum_y n_y log|\boldsymbol{\beta}^T\boldsymbol{\Delta}_y\boldsymbol{\beta}| \\
&- \frac{1}{2}\sum_y\sum_i[\boldsymbol{\beta}^T(\mathbf{X}_{yi} - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\beta}\boldsymbol{\nu}_y)]^T(\boldsymbol{\beta}^T\boldsymbol{\Delta}_y\boldsymbol{\beta})^{-1}[\boldsymbol{\beta}^T(\mathbf{X}_{yi} - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\beta}\boldsymbol{\nu}_y)] \\
&- \frac{1}{2}\sum_y\sum_i[(\boldsymbol{\beta}_0^T - \mathbf{H}\boldsymbol{\beta}^T)(\mathbf{X}_{yi} - \boldsymbol{\mu})]^T\mathbf{D}^{-1}[(\boldsymbol{\beta}_0^T - \mathbf{H}\boldsymbol{\beta}^T)(\mathbf{X}_{yi} - \boldsymbol{\mu})]
\end{aligned}$$

$$\begin{aligned}
&= -\frac{np}{2}log(2\pi) - \frac{n}{2}log|\mathbf{D}| - \frac{1}{2}\sum_y n_y log|\boldsymbol{\beta}^T\boldsymbol{\Delta}_y\boldsymbol{\beta}| \\
&- \frac{1}{2}\sum_y n_y[\boldsymbol{\beta}^T(\bar{\mathbf{X}}_y - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\beta}\boldsymbol{\nu}_y)]^T(\boldsymbol{\beta}^T\boldsymbol{\Delta}_y\boldsymbol{\beta})^{-1}[\boldsymbol{\beta}^T(\bar{\mathbf{X}}_y - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\beta}\boldsymbol{\nu}_y)] \\
&- \frac{1}{2}\sum_y n_y(\bar{\mathbf{X}}_y - \boldsymbol{\mu})^T\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T(\bar{\mathbf{X}}_y - \boldsymbol{\mu}) \\
&- \sum_y \frac{n_y}{2}tr\{\boldsymbol{\beta}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\Delta}_y\boldsymbol{\beta})^{-1}\} \\
&- \sum_y \frac{n_y}{2}tr\{\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T\tilde{\boldsymbol{\Delta}}_y\}
\end{aligned}$$

where $\mathbf{K} = (\boldsymbol{\beta}_o - \boldsymbol{\beta}\mathbf{H}^T)$, $\mathbf{H} = (\boldsymbol{\beta}_0^T\boldsymbol{\Delta}\boldsymbol{\beta})(\boldsymbol{\beta}^T\boldsymbol{\Delta}\boldsymbol{\beta})^{-1}$ and $\mathbf{D} = (\boldsymbol{\beta}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}_0)^{-1}$. We show the forth statement of above equation by T_4. We set $\bar{\mathbf{a}} = \sum_y f_y\mathbf{a}_y$ where $f_y = \frac{n_y}{n}$. To minimum

$$\frac{T_4}{n} = \sum_y f_y(\mathbf{Z}_y - \bar{\mathbf{B}}\boldsymbol{\nu}_y)^T\mathbf{B}_y^{-1}(\mathbf{Z}_y - \bar{\mathbf{B}}\boldsymbol{\nu}_y)$$

Relative to the terms $\bar{\boldsymbol{\nu}} = 0$. We use the Lagrange coefficient $\lambda \in \mathbb{R}^d$ in which$\mathbf{Z}_y = \boldsymbol{\beta}^T(\bar{\mathbf{X}} - \boldsymbol{\mu})$ $\mathbf{B}_y = \boldsymbol{\beta}^T\boldsymbol{\Delta}_y\boldsymbol{\beta}$ $\bar{\mathbf{B}} = \boldsymbol{\beta}^T\boldsymbol{\Delta}\boldsymbol{\beta}$ Namely, we must minimize $\frac{T_4}{n} + \lambda^T\bar{\boldsymbol{\nu}}$.
We now have a derivative of $\boldsymbol{\nu}_y$:

$$-2f_y\bar{\mathbf{B}}\mathbf{B}_y^{-1}\mathbf{Z}_y + 2f_y\bar{\mathbf{B}}\mathbf{B}_y^{-1}\bar{\mathbf{B}}\boldsymbol{\nu}_y + f_y\lambda = 0, \tag{33}$$

or equivalently

$$2f_y\mathbf{Z}_y + 2f_y\bar{\mathbf{B}}\boldsymbol{\nu}_y + f_y\mathbf{B}_y\bar{\mathbf{B}}^{-1}\lambda = 0.$$

By adding on y, the second term is zero and the third term be λ. So that $\lambda = 2\bar{\mathbf{Z}}$. Therefore, from the equation (33) we have:

$$\boldsymbol{\nu}_y = \bar{\mathbf{B}}^{-1}(\mathbf{Z}_y - \mathbf{B}_y\bar{\mathbf{B}}^{-1}\bar{\mathbf{Z}}).$$

By substituting $\boldsymbol{\nu}_y$ in T_4:

$$\frac{\tilde{T_4}}{n} = \sum_y f_y \bar{\mathbf{Z}}^T \bar{\mathbf{B}}^{-1} \mathbf{B}_j \mathbf{B}_j^{-1} \mathbf{B}_j \bar{\mathbf{B}}^{-1} \bar{\mathbf{Z}} = \bar{\mathbf{Z}}^T \bar{\mathbf{B}}^{-1} \bar{\mathbf{Z}} = (\boldsymbol{\beta}^T \bar{\mathbf{X}} - \boldsymbol{\beta}^T \boldsymbol{\mu})^T \bar{\mathbf{B}}^{-1} (\boldsymbol{\beta}^T \bar{\mathbf{X}} - \boldsymbol{\beta}^T \boldsymbol{\mu}).$$

To find maximum of $\boldsymbol{\mu}$:

$$\frac{\partial L_d}{\partial \boldsymbol{\mu}} = n\boldsymbol{\beta}(\boldsymbol{\beta}^T \boldsymbol{\Delta} \boldsymbol{\beta})^{-1} \boldsymbol{\beta}^T (\bar{\mathbf{X}} - \boldsymbol{\mu}) + n\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T (\bar{\mathbf{X}} - \boldsymbol{\mu}). \tag{34}$$

Using (27), definition of $\mathbf{H}$ and $\mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Delta})}$, we have:

$$\begin{aligned} \mathbf{K}^T &= \boldsymbol{\beta}_0^T - \mathbf{H}\boldsymbol{\beta}^T = \boldsymbol{\beta}_0^T - (\boldsymbol{\beta}_0^T \boldsymbol{\Delta} \boldsymbol{\beta})(\boldsymbol{\beta}^T \boldsymbol{\Delta} \boldsymbol{\beta})^{-1} \boldsymbol{\beta}^T \\ &= \boldsymbol{\beta}_0^T (\mathbf{I}_p - \mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Delta})}^T) = \boldsymbol{\beta}_0^T \mathbf{P}_{\boldsymbol{\beta}_0(\boldsymbol{\Delta}^{-1})} \\ &= (\boldsymbol{\beta}_0^T \boldsymbol{\Delta}^{-1} \boldsymbol{\beta}_0)^{-1} \boldsymbol{\beta}_0^T \boldsymbol{\Delta}^{-1}. \end{aligned}$$

Thus

$$\begin{aligned} \mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T &= (\boldsymbol{\beta}_0 - \boldsymbol{\beta}\mathbf{H}^T)^T \mathbf{D}^{-1} (\boldsymbol{\beta}_0^T - \mathbf{H}\boldsymbol{\beta}^T) \\ &= \boldsymbol{\Delta}^{-1} \boldsymbol{\beta}_0 (\boldsymbol{\beta}_0^T \boldsymbol{\Delta}^{-1} \boldsymbol{\beta}_0)^{-1} (\boldsymbol{\beta}_0^T \boldsymbol{\Delta}^{-1} \boldsymbol{\beta}_0)^{-1} (\boldsymbol{\beta}_0^T \boldsymbol{\Delta}^{-1} \boldsymbol{\beta}_0)^{-1} \boldsymbol{\beta}_0^T \boldsymbol{\Delta}^{-1} \\ &= \boldsymbol{\Delta}^{-1} \boldsymbol{\beta}_0 (\boldsymbol{\beta}_0^T \boldsymbol{\Delta}^{-1} \boldsymbol{\beta}_0)^{-1} \boldsymbol{\beta}_0^T \boldsymbol{\Delta}^{-1}. \end{aligned} \tag{35}$$

By substituting (35) in equation (34), we get:

$$\frac{\partial L_d}{\partial \boldsymbol{\mu}} = n\boldsymbol{\beta}(\boldsymbol{\beta}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\beta})^{-1} \boldsymbol{\beta}^T (\bar{\mathbf{X}} - \boldsymbol{\mu}) + n\boldsymbol{\Delta}^{-1} \boldsymbol{\beta}_0 (\boldsymbol{\beta}_0^T \boldsymbol{\Delta}^{-1} \boldsymbol{\beta}_0)^{-1} \boldsymbol{\beta}_0^T \boldsymbol{\Delta}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}).$$

Now follows from the (26) from lemma 2 have the following result:

$$\frac{\partial L_d}{\partial \boldsymbol{\mu}} = n\boldsymbol{\Delta}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}).$$

Thus $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$ and $\hat{\boldsymbol{\nu}} = \mathbf{0}$. If we set $\tilde{\boldsymbol{\Sigma}}_y = \tilde{\boldsymbol{\Delta}}_y + (\bar{\mathbf{X}}_y - \bar{\mathbf{X}})(\bar{\mathbf{X}}_y - \bar{\mathbf{X}})^T$, then

$$\begin{aligned} L_d = &-\frac{np}{2} log(2\pi) - \frac{n}{2} log|\mathbf{D}| - \frac{1}{2} \sum_y log|\boldsymbol{\beta}^T \boldsymbol{\Delta}_y \boldsymbol{\beta}| \\ &- \frac{1}{2} \sum_y n_y [(\boldsymbol{\beta}_0^T - \mathbf{H}\boldsymbol{\beta}^T)(\bar{\mathbf{X}}_y - \bar{\mathbf{X}})]^T \mathbf{D}^{-1} [(\boldsymbol{\beta}_0^T - \mathbf{H}\boldsymbol{\beta}^T)(\bar{\mathbf{X}}_y - \bar{X})] \\ &- \frac{1}{2} \sum_y tr((\boldsymbol{\beta}^T \tilde{\boldsymbol{\Delta}}_y \boldsymbol{\beta})(\boldsymbol{\beta}^T \boldsymbol{\Delta}_y \boldsymbol{\beta})^{-1}) \\ &- \frac{1}{2} \sum_y n_y tr((\boldsymbol{\beta}_0^T - \mathbf{H}\boldsymbol{\beta}^T)^T \mathbf{D}^{-1} (\boldsymbol{\beta}_0^T - \mathcal{H}\boldsymbol{\beta}^T) \tilde{\boldsymbol{\Delta}}_y) \\ = &-\frac{np}{2} log(2\pi) - \frac{n}{2} log|\mathbf{D}| - \frac{1}{2} \sum_y n_y log|\boldsymbol{\beta}^T \boldsymbol{\Delta}_y \boldsymbol{\beta}| \\ &- \sum_y \frac{n_y}{2} tr\{\boldsymbol{\beta}^T \tilde{\boldsymbol{\Delta}}_y \boldsymbol{\beta} (\boldsymbol{\beta}^T \boldsymbol{\Delta}_y \boldsymbol{\beta})^{-1}\} - \sum_y \frac{n_y}{2} tr\{\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T \tilde{\boldsymbol{\Sigma}}_y\} \end{aligned}$$

The maximum likelihood of $\mathbf{B}_y$ obtains by:

$$\hat{\mathbf{B}}_y = \widehat{\boldsymbol{\beta}^T \boldsymbol{\Delta}_y \boldsymbol{\beta}} = \boldsymbol{\beta}^T \tilde{\boldsymbol{\Delta}}_y \boldsymbol{\beta},$$

and thus

$$L_d = -\frac{np}{2}log(2\pi) - \frac{nd}{2} - \frac{n}{2}log|\mathbf{D}| - \frac{1}{2}\sum_y n_y log|\boldsymbol{\beta}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\beta}| - \sum_y \frac{n_y}{2}tr\{\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T\tilde{\boldsymbol{\Sigma}}_y\}.$$

To find the MLE of $\mathbf{K}$, remember that $\mathbf{K} = (\boldsymbol{\beta}_o - \boldsymbol{\beta}\mathbf{H}^T)$ and it suffices to find MLE of $\mathbf{H}$. Therefore,

$$\frac{\partial L_d}{\partial \mathbf{H}} = \sum_y n_y\mathbf{D}^{-1}\boldsymbol{\beta}_0^T\tilde{\boldsymbol{\Sigma}}_y\boldsymbol{\beta} + \sum_y n_y\mathbf{D}^{-1}\mathbf{H}\boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_y\boldsymbol{\beta},$$

Consequently

$$\hat{\mathbf{H}} = (\sum_y n_y\boldsymbol{\beta}_0^T\tilde{\boldsymbol{\Sigma}}_y\boldsymbol{\beta})(\sum_y n_y\boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_y\boldsymbol{\beta})^{-1} = (\boldsymbol{\beta}_0^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\beta})(\boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\beta})^{-1}),$$

where $\tilde{\boldsymbol{\Sigma}} = \sum_y f_y\tilde{\boldsymbol{\Sigma}}_y$. Using (26), the maximum likelihood function relative to $\mathbf{D}$ is:

$$\begin{aligned}\hat{\mathbf{D}} &= (\boldsymbol{\beta}_0^T - \hat{\mathbf{H}}\boldsymbol{\beta}^T)\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\beta}_0^T - \hat{\mathbf{H}}\boldsymbol{\beta}^T)^T\\ &= [(\boldsymbol{\beta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\beta}_0)^{-1}\boldsymbol{\beta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}]\tilde{\boldsymbol{\Sigma}}[(\boldsymbol{\beta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\beta}_0)^{-1}\boldsymbol{\beta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}]\\ &= (\boldsymbol{\beta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\beta}_0)^{-1}.\end{aligned}$$

Using (31), the log likelihood in $\boldsymbol{\beta}$ is as follows:

$$\begin{aligned}L_d &= -\frac{np}{2}(1 + log2\pi) + \frac{n}{2}log|\boldsymbol{\beta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\beta}_0| - \frac{1}{2}(\sum_y n_y log|\boldsymbol{\beta}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\beta}|\\ &= -\frac{np}{2}(1 + log2\pi) + \frac{n}{2}log|\boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\beta}| - \frac{n}{2}log|\tilde{\boldsymbol{\Sigma}}| - \frac{1}{2}\sum_y n_y log|\boldsymbol{\beta}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\beta}|. \qquad (36)\end{aligned}$$

Since $|\mathbf{P}_S\hat{\boldsymbol{\Sigma}}\mathbf{P}_S|_0 = |\boldsymbol{\beta}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta}|$? by substituting in (36), equation (32) is obtained. For the value of $\boldsymbol{\beta}$, $\mathbf{H}$ and $\mathbf{D}$ one can get $\boldsymbol{\Delta}$ one by one. Now, we use the relationships (35) and (26):

$$\boldsymbol{\Delta}^{-1} = \boldsymbol{\beta}\mathbf{A}^{-1}\boldsymbol{\beta}^T + \mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T, \qquad (37)$$

where $\mathbf{A} = \boldsymbol{\beta}^T\boldsymbol{\Delta}\boldsymbol{\beta}$. The estimation of $\mathbf{A}$ is considered as $\hat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Delta}}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that maximize the equation (36). The MLE of $\boldsymbol{\Delta}^{-1}$ obtains by substituting $\boldsymbol{\beta}$, $\mathbf{K}$, $\mathbf{A}$ and $\mathbf{D}$ in (37). Therefore:

$$\begin{aligned}\hat{\boldsymbol{\Delta}}^{-1} &= \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Delta}}\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\beta}}^T + \hat{\mathbf{K}}(\hat{\mathbf{K}}^T\tilde{\boldsymbol{\Sigma}}\hat{\mathbf{K}})^{-1}\hat{\mathbf{K}}^T\\ &= \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Delta}}\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\beta}}^T + \tilde{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Sigma}}\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\beta}}^T.\end{aligned}$$

□

Also

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Delta}} + \mathbf{P}^T_{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\Delta}})}\hat{\mathbf{M}}\mathbf{P}_{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\Delta}})},$$

where $\hat{\mathbf{M}}$ is estimation of $var(\boldsymbol{\mu}_y)$ and all estimators are set to maximize the logarithm of likelihood function.

4.1. Robustness of $\hat{S}_{Y|\mathbf{X}}$ to non-normality

Diaconis and Freedman (1984) showed that under appropriate conditions, almost all projection of high-dimensional data are normal. Therefore, when data is not normal, the LAD method is expected to maintain its desirable performance. Cook and Forzani (2009) showed by simulation that the errors are not normal, the LAD method has good performance. In the other word, the estimation of $S_{Y|X}$ is robustness to nonnormality.

5. Simulation study

In this section, we compare sufficient dimension reduction methods with each other and also with principal component analysis. We use the MATLAB software to do this simulation study. First for $n = 500$, we generate the random vector $\mathbf{X}$ from the Normal distribution $N_p(0, Ip)$ and $\epsilon \sim N(0,1)$. We consider the following models:

$$
\begin{aligned}
&a)\ Y = 4X_1/a + \epsilon, \quad a = 1, ..., 10, \quad p = 8, \quad h = 5\\
&b)\ Y = X_1^2/(20a) + 0.1\epsilon, \quad a = 1, ..., 10, \quad p = 8, \quad h = 5\\
&c)\ Y = X_1/(10a) + aX_1^2/100 + 0.6\epsilon, \quad a = 1, ..., 10, \quad p = 8, \quad h = 5\\
&d)\ Y = 0.4a(\boldsymbol{\beta}_1^T\mathbf{X})^2 + 3\sin(\boldsymbol{\beta}_2^T\mathbf{X}/4) + 0.2\epsilon, \quad a = 1, ..., 10, \quad p = 20, \quad h = 10.
\end{aligned}
$$

For the first three models, $S_{Y|\mathbf{X}} = span((1, ..., 0)^T)$ and for the forth model is spanned by $S_{Y|\mathbf{X}} = span((1, 1, 1, 0, ..., 0)^T)$ and $span(1, 0, ..., 0, 1, 3)^T)$. We suppose the conditional distribution of $\mathbf{X}|Y$ is normal for the first model and for other models nonnormal. The figure 1 compare the angle between $S_{Y|\mathbf{X}}$ and its estimation by methods SIR, SAVE, DR and LAD with 400 replications. Figure 1a is plot of linear model, it shows that for small amount of a all the methods except PCA have a good performance, by increasing a the SAVE, PCA and DR methods have Poor performance but SIR and LAD behave good. In Figure 1b, we have quadratic and PCA again couldn't do well. Also it is known that SIR perform well just for linear models., but other methods perform similarly and LAD do best among them. Model 3 has both linear and quadratic term. Figure 1c shows average angle between $S_{Y|\mathbf{X}}$ and its estimation for this model. For small a, SIR perform well but by increasing a the strength of the linear trend decreases and the strength of the quadratic trend increases, so the SAVE method do better. As it is obvious in figure, LAD has good performance far all value of a. Again for the model PCA is not good method. In Figure 1d, the purposed model has linear tend in $\boldsymbol{\beta}_2^T X_2$ and quadratic trend in $\boldsymbol{\beta}_1^T X_1$. Since SIR couldn't find quadratic trend and SAVE is poor in finding linear trend, SO as in figure shown, the LAD's performance is best among all methods. The DR mothod is good for small value of a. The PCA couldn't find a trend for this model too.

6. Application

For patients with jaw or skeletal and dental problems, such as the lack of proper fitting of the teeth during chewing, lack of proper jaw contact, jawbone opening, mouth opening and other, one of the treatment options is orthognathism surgery. Nowadays the number of people needed to have orthognathism surgery is increasing and the cost of this surgery is high, so it is necessary to have an exact plan for it. There are different surgery for orthognathism surgery. In orthognathism surgery treatment, morphology of craniofacial and position of head and neck and airway change. Sometimes, in orthodontic treatment, the patient may return to the initial condition. By knowing what changes in craniofacial morphology and head and neck and airway conditions occur, this problem can be cured in orthodontic treatment. For this purpose, the patients are selected from the archives in the orthodontic part of department of Dentistry, Tabriz University of Medical Sciences. The dimensions of their jaws and facial expressions are measured. From 47 different angles and depending on the size of the measurements, are placed in two groups. For the first group, lower maxillary surgery was done forward, and for the second group, upper maxillary surgery was done forward. We investigate data in two steps.

6.1. Before surgery

At this stage, that is, before the surgery, for 55 selected patients, the facial dimensions are measured and depending on the size, fall into one of two groups. In fact for response variable $Y = \{1, 2\}$, we have 47 predictor variables. The goal is to provide a model to predict which surgery should be performed after o measuring the dimensions of the patient's face. For this data, 26 patients are in the first group and 29 patients in the second group. Given a total size with sample 55 and the number of predictor variables, it is clear that a suitable logistic model(Because the response variable consists of two groups) can not be fitted to these data. We reduce the dimension of predictors in to two

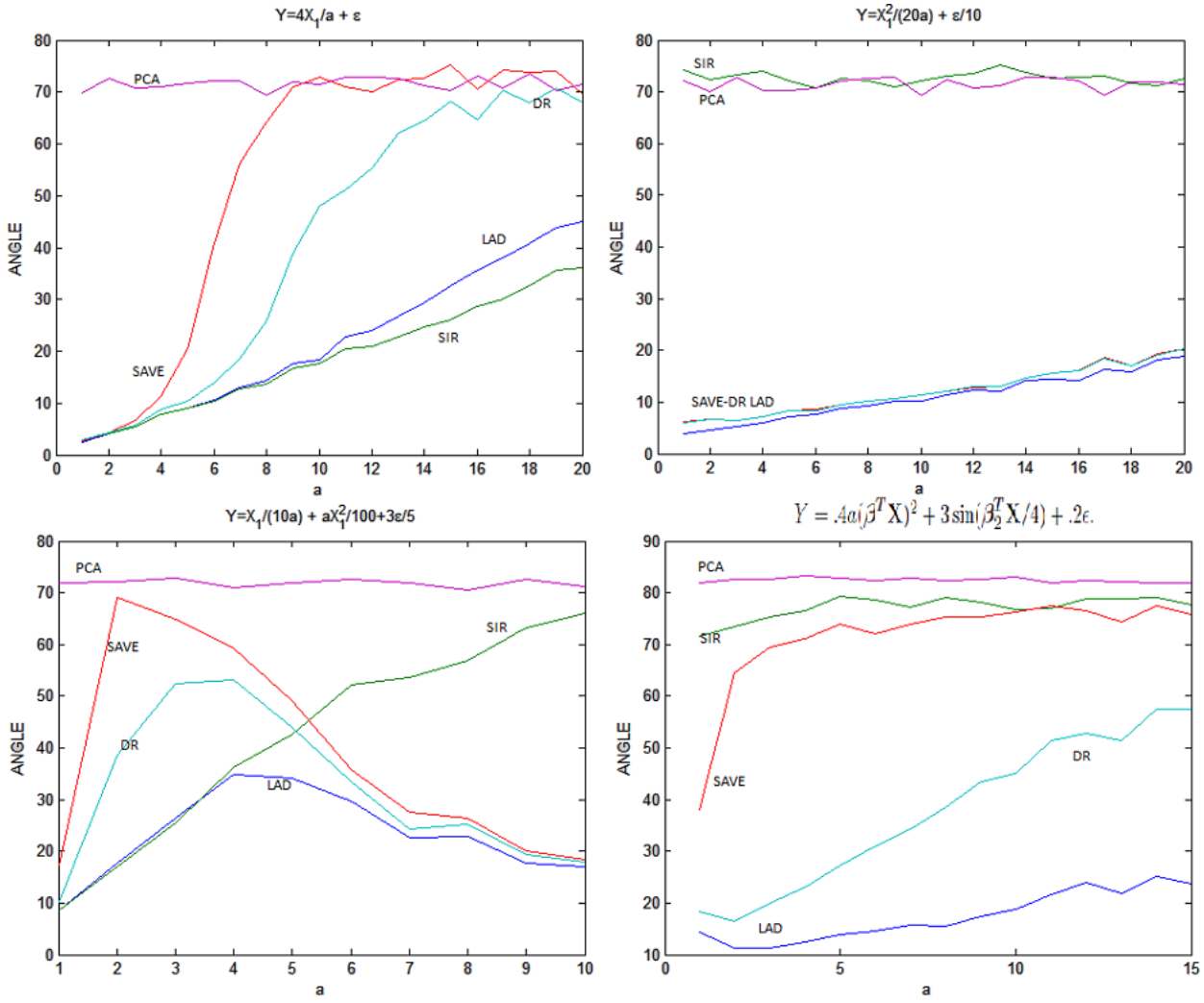

Figure 1. The average angle between $S_{Y|\mathbf{X}}$ and its estimation

using proposed methods. Figure 2 shows the results. As you can see in the figure 2, the DR, $SIR-SAVE$ and LAD methods separate data well. Since the number of sample is relatively lower than the number of predictors, the dimension reduction methods based on inverse regression is not as good as LAD, so we choose the LAD model to offer a model. The following model obtains for data:

$$logit(Y) = -320.35 - 5.73LAD_1 + 8.22LAD_2. \tag{38}$$

Now, according to the model (38), we can quantitatively obtain the correct classification of the data. Table 1 shows the classification values. As you can see in table 1, the SIR and DR methods have not been able to classify the data due to the small number of samples relative to the number of predictors. The values listed in table 1 for the PCA show that the poor performance of this method, because we know that this method is not designed for this purpose.

To determine if the fitted model correctly predicts the variables of the response variable, each time a sample of data is deleted and the remaining 54 data is subtracted, then appropriate logistic regression model is fitted to them. Then

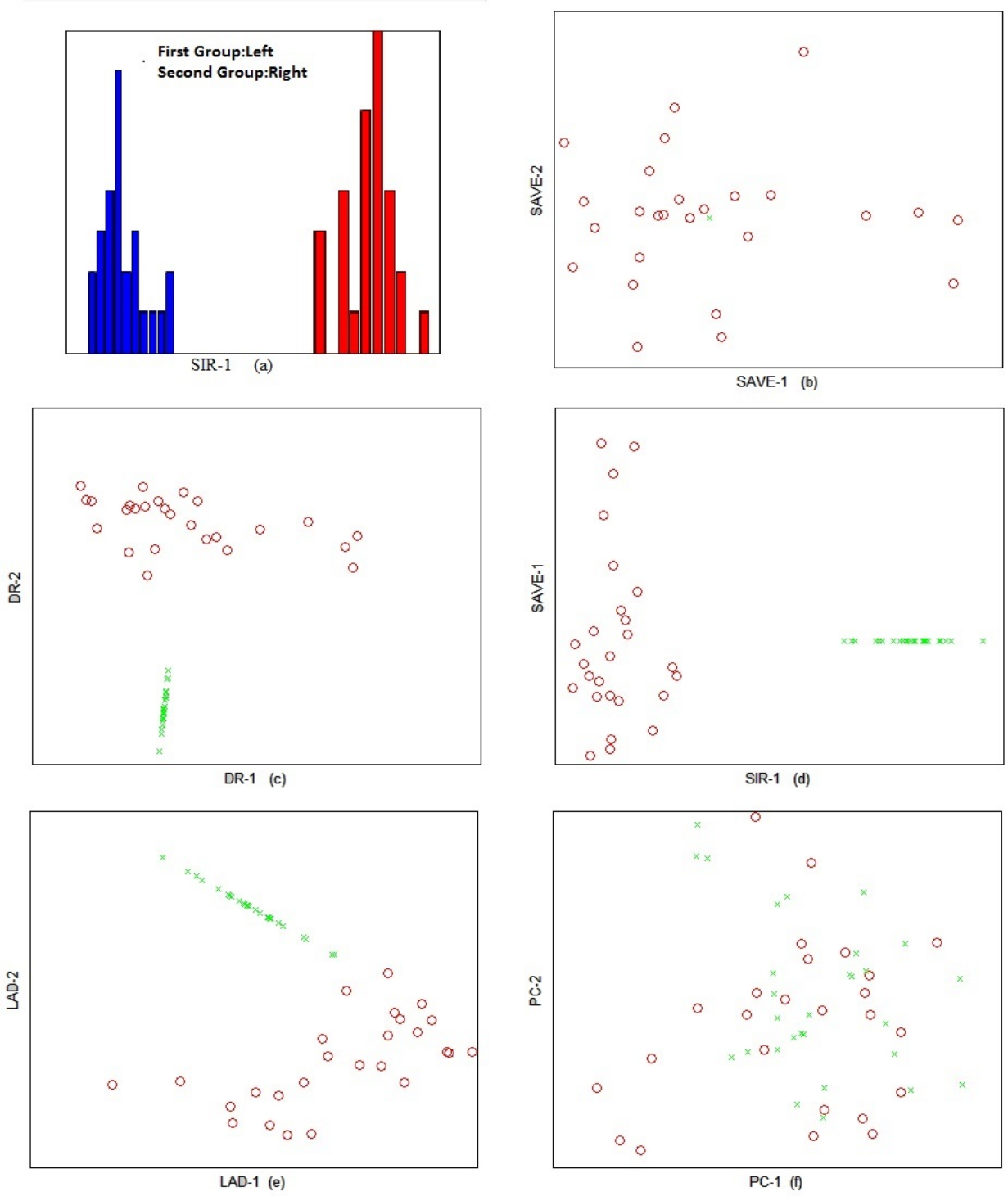

Figure 2. The first two predictors of the various methods for preoperative data (the circle represents the first group and the star represents the second group).

Table 1. Correct percentage of preoperative data classification

Y	$SAVE$	DR	LAD	PCA
0	100	–	100	46.2
1	100	–	100	65.5
Total	100	–	100	56.4

we predict the response variable of the data we removed from fitted model. So this action should be repeated 55 times. The results of 55 repetitions showed that only 5 cases had a prediction error, ie the predictor varies with the actual response variable. In fact, the prediction error is $\frac{2+3}{55}$. The results are reported in the table 2. So, in addition to dimension reduction, the prediction model has good performance.

Table 2. Prediction error classification for first stage data

	predicted value		
observed value	0	1	*total*
0	24	2	26
1	3	26	29
toatal	27	28	55

6.2. The second stage: difference between the data before and after surgery

In this stage, the surgery is done on the patients of two groups. As stated, for the first group, lower maxillary surgery was done forward, and for the second group, upper maxillary surgery was done forward. Then we get the difference of 47 predictor variables before and after surgery. in this section by choosing the appropriate dimension reduction method and offering proper logistic regression model, we predict that which surgery has been performed on the patient? As you can see in figure3, the DR and LAD methods separated data well. But according to the above, the LAD method also performs well for parameter estimation. Therefore, to reduce the dimension of the data, we use the LAD method and the model

$$logit(Y) = -18.06 + 7.89LAD_1 + 2.79LAD_2, \tag{39}$$

is fitted for the differential data. Table 3 also shows the correct percentage for the second stage data, as in the previous table. The LAD method classified the 96.4 the percentage of data correctly. Again, the values in the table 3 for the PCA method show the poor performance of this method. To calculate the prediction error of data, each

Table 3. Correct percentage of difference data classification

Y	$SAVE$	DR	LAD	PCA
0	100	100	96.2	0
1	100	100	96.6	100
total	100	100	96.4	64.4

time, by deleting a sample, we reduce the number of sample size to 54, and fit the appropriate model for it. Then we predict the response variable of the data and remove from fitted model. In this case, the prediction error is $dfrac2 + 455$. The results are presented in the Table 4. Again the proposed model performs well.

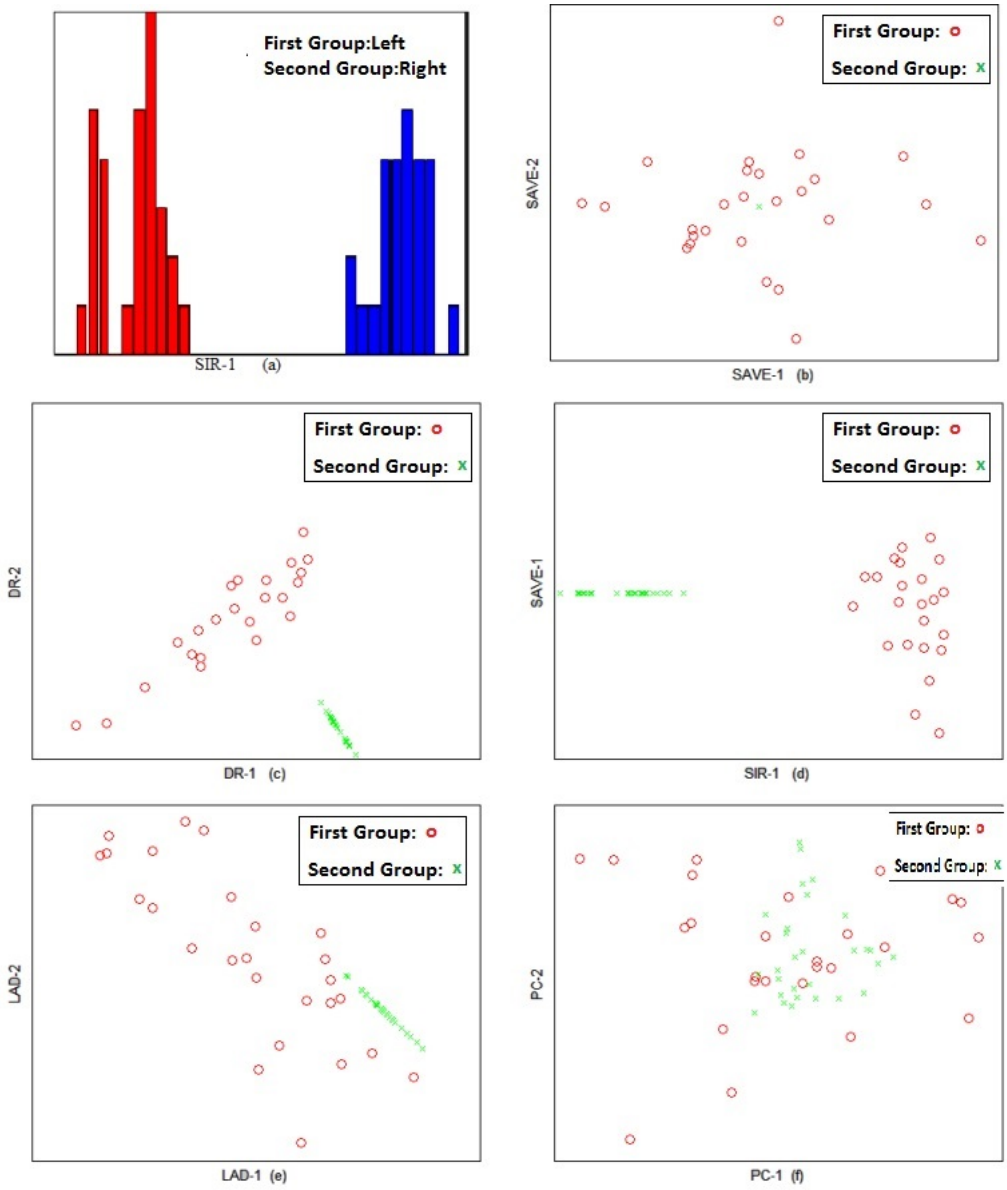

Figure 3. First two predictors of the various methods for the differential data (the circle representing the first group and the star representing the second group).

7. Conclusions

In the present paper, the methods of diminishing the dimension of variables, which include the estimation of central subspace based on the inverse regression, the likelihood acquisition method and principal component analysis considered. By reducing the dimension of variables without losing any information, good results are obtained. The Likelihood Acquired Directions method and methods based on inverse regression have good performance for regression models since they designed for regression purposes. Using a real data associated with the dental problems the Logistic regression is fitted and the correct classification of the data computed. The simulation study

Table 4. Prediction error classification for the second stage data

	predicted value		
observed value	0	1	*total*
0	22	4	26
1	2	27	29
total	24	31	55

is presented to compare the sufficient dimension reduction methods with each other. Among all the dimension reduction methods, the LAD perform the best. The LAD method is robust for nonnormality, too. Also, the Principal Component Analysis has not good performance for regression models.

REFERENCES

1. Afshari, M. (2017) *Nonlinear wavelet shrinkage estimator of nonparametric regularity regression function via cross-validation with simulation study*, International Journal of Wavelets, Multiresolution and Information Processing,15, 1-16.
2. Afshari, M., Lak, F. and Gholizadeh, B. (2017) *A new Bayesian wavelet thresholding estimator of nonparametric regression*, J. Appl. Stat, 44, 649-666.
3. Amato, U., Antoniadis, A. and Feis, I. (2006) *Dimension Reduction in Functional Regression with Applications*, Computational Statistics & Data Analysis, 50, 2422-2446.
4. Bremel, R.D., Homan, E.J. (2010) *An integrated approach to epitope analysis I: Dimensional reduction, visualization and prediction of MHC binding using amino acid principal components and regression approaches*, Immunome Res 6, 7.
5. Bura, E., and Cook, R. D. (2001), *Extending Sliced inverse regression: The weighted Chi-squared test*, Journal of the American Statistical Association 996-103.
6. Cook, R. D. (1994), *Using dimension-reduction subspaces to identify important inputs in models of physical systems*, Proceedings of the Section on Physical and Engineering Sciences Journal of American Statistical Association,18-25.
7. Cook, R. D. (1998), *Regression Graphics*, John Wiley and Sons, New York.
8. Cook, R. D. and Forzani, L. (2009), *Likelihood based sufficient dimension reduction*, Journal of American Statistical Association 104, 197-20.
9. Cook, R. D. and Weisberg, S. (1991), *Discussion of Sliced inverse regression by K. C. Li*, Journal of the American Statistical Association,86, 316-342.
10. Diaconis, P. and Freedman, D. (1984), *Asymptotics of graphical projection pursuit*, The Annals of Statistics 12, 793C815.
11. Duan, N., and Li, K. C. (1991), *Slicing regression: A link-free regression method*, The Annals of Statistics,19(2), 505-530.
12. Fukunaga, F. (1990), *Introduction to Statistical Pattern Recognition*, Academic Press Professional, Inc., San Diego, CA, USA.
13. Hsing, T. and Carroll, R. J. (1992), *An asymptotic theory for sliced inverse regression*, The Annals of Statistics,20(2), 1040-1061.
14. Jimenez, L. O. and Landgrebe, D. A. (1997), *Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data*, IEEE Transactions on Systems, Man and Cybernetics,28, 39-54.
15. Joliffe, I. T. (1986) *Component Analysis*, Springer-Verlag, New York, USA.
16. Li, K. C. (1991), *Sliced inverse regression for dimension reduction (with discussion)*, Journal of the American Statistical Association,86, 316-342.
17. Li, K. C. (1992), *On principal Hessian directions for data visualization and dimension reduction: Another application of Steins lemma*, Journal of American Statistical Association,87, 1025-1039.
18. Li, K. C., Wang, J. L., and Chen, C. H. (1999), *Dimension reduction for censored regression data*, The Annals of Statistics,27, 1-23.
19. Li, K. C., Aragon, Y., Shedden, K., Agnan C. T. (2003), *Dimension reduction for multivariate response data*, Journal of American Statistical Association,98,9-109.
20. Li, B. and Wang, S. (2007), *On directional regression for dimension reduction*, Journal of American Statistical Association,102, 997-1008.
21. Li, L. and Yin, X. (2008), *Sliced inverse regression with regularizations* Biometrics,64, 124-131.
22. Lue, H. H. (2008), *Sliced average variance estimation for censored data*, Communications in Statistics-Theory and Methods,37, 3276-3286.
23. Luo, W., Li, B. and Yin, X. (2014), *On efficient dimension reduction with respect to a statistical functional of interest*, The Annals of Statistics,42(1), 382-412.
24. Ye, Z. and Weiss, R. (2003), *Using the bootstrap to select one of a new class of dimension reduction methods*, Journal of the American Statistical Association,98, 968-979.
25. Yin, X. and Cook, R. D. (2005), *Direction estimation in single-index regressions*, Biometrika .92(2), 371-384.
26. Yin, X., Li, B. and Cook, R.D. (2008), *Successive direction extraction for estimating the central subspace in a multiple-index regression*, Journal of Multivariate Analysis,99, 1733C1757.
27. Yu, Z., Dong, Y., and Huang, M. (2014), *General directional regression*, Journal of Multivariate Analysis,124, 94-104.

28. Zhu, L. P. and Zhu, L. X. (2007), *On kernel method for sliced average variance estimation*, of Multivariate Analysis,98, 970-991.
29. Zhu, L. X., Ohtaki, M. and Li, Y. X. (2007), *On hybrid methods of inverse regression based algorithms*, Computational Statistics and Data Analysis,51, 2621-2635.
30. Zhu, L. P., Zhu, L. X. and Feng, Z. H. (2010), *Dimension reduction in regressions through cumulative slicing estimation*, Journal of the American Statistical Association,105, 1455-1466.

Appendix

```
function n = Dmatel(m)

   [p,q]=size(m);

   if mod(q,3)==1
q=q-1;
end
if mod(q,3)==2
q=q-2;
end

   l=m(1,4);
u=m(1,3);

   for i=1:q

   if mod(i,3)==1 & min(m(:,i))¡l
l=min(m(:,i));
end
if mod(i,3)==0 & max(m(:,i))¿u
u=max(m(:,i));
end

   end

   delta=u-l;

   for i=1:p
for j=1:q

   m(i,j)=(m(i,j)-l)/delta;

   end
end
A=m;
A
for i=1:p
for j=1:q

   if mod(j,3)== 1
c(i,j)=m(i,j+1)/(1+m(i,j+1)-m(i,j));
end

   if mod(j,3)== 0
c(i,j)=m(i,j)/(1+m(i,j)-m(i,j-1));
end
```

```
   end
end
B=c;
B

   for i=1:p
for j=1:q

   if mod(j,3)== 1
g(i,j)=(c(i,j)*(1-c(i,j))+c(i,j+2)*c(i,j+2))/(1-c(i,j)+c(i,j+2));
else g(i,j)=0;
end
end
end
D=g;
D

   for i=1:p
for j=1:q
if mod(j,3)== 1
g(i,j)=l+delta*g(i,j);
end
end
end
n=g;
```

```
.....
data = load('ta1255.txt');
Y = data(:,1);
X = data(:,2:end);
figure(1);
WXsir,Wsir = SIR(Y,X,'disc',1);
plotDR(WXsir,Y,'disc','SIR');
```

```
.....
function [Wn,fn,fp,vals] = pfc(Yaux,X,u,morph,parameters)

   if strcmpi(morph,'disc'),
Y = mapdata(Yaux);
parameters.nslices = max(Y);
else
Y = Yaux;
parameters.nslices = length(Y);
end
```

```
data_parameters = setdatapars(Y,X,parameters.nslices);

   Fhandle = F(@F4pfc,data_parameters);

   if strcmpi(morph,'cont')
SIGMAfit = get_fitted_cov(Y,X,parameters.fy);
else
SIGMAfit = get_average_cov(X,data_parameters);
end
SIGMA = data_parameters.sigmag;
SIGMAres = SIGMA - SIGMAfit;

   p = cols(X);
Wn = eye(p);
fp = Fhandle(ones(1,p));
if u == p,

   fn = fp;
else
Wn,vals = firsteigs(inv(SIGMAres)*SIGMA,u);
Wn = orth(Wn);
fn = Fhandle(vals);
vals;
end
```

.....

```
ncols=8; nreps=50; u=1;
nrows=[20 40 60 100 150 200 250 300];
dim_lrt=zeros(nreps,length(nrows));
dim_bic=zeros(nreps,length(nrows));
dim_aic=zeros(nreps,length(nrows));

   for k=1:length(nrows)
disp(['nrows = ' int2str(nrows(k))]);
X = zeros(nrows(k)*3,ncols);
for j=1:nreps
alp = zeros(nrows(k),ncols);
alp(:,ncols) = 1;
mu = [6, 4, 2];
sig = [1, 4, 8];

   t1 = normrnd(0,1,nrows(k)*3, ncols);
t2uno = normrnd( 0,1,nrows(k)*3, 1);
t2 = zeros(nrows(k)*3,ncols);
for i=1:ncols
t2(:,i) = t2uno;
end
```

```
X1 = mu(1)*alp + t1(1:nrows(k),:) + sig(1)* t2(1:nrows(k),:).*alp;
X2 = mu(2)*alp + t1((nrows(k)+1):2*nrows(k),:) + sig(2)*(t2((nrows(k)+1):2*nrows(k),:) ).*alp;
X3 = mu(3)*alp + (t1((2*nrows(k)+1):3*nrows(k),:) ) + sig(3)*(t2((2*nrows(k)+1):3*nrows(k),:) ).*alp;

   X = [X1; X2; X3];
Y = ones(size(X,1),1);
Y(size(X1,1)+1:(size(X1,1)+size(X2,1)),1)=2;
Y(size(X1,1)+size(X2,1)+1:(size(X1,1)+size(X2,1)+size(X3,1)),1)=3;

   [WX,W,fn,d] = ldr(Y,X,'LAD','disc','lrt','alpha',0.05);
dim_lrt(j,k) = d;

   [WX1,W1,fn1,d1]=ldr(Y,X,'LAD','disc','aic');

   [WX2,W2,fn2,d2] = ldr(Y,X,'LAD','disc','bic');
dim_aic(j,k) = d1;
dim_bic(j,k) = d2;
end
end

   mean_eq1 = zeros(3,length(nrows));
mean_eq12 = zeros(3,length(nrows));

   for k=1:length(nrows)
mean_eq1(1,k) = sum(dim_lrt(:,k)==1)/size(dim_lrt,1);
mean_eq12(1,k) = sum(dim_lrt(:,k)¡3)/size(dim_lrt,1);
mean_eq1(2,k) = sum(dim_aic(:,k)==1)/size(dim_aic,1);
mean_eq12(2,k) = sum(dim_aic(:,k)¡3)/size(dim_aic,1);
mean_eq1(3,k) = sum(dim_bic(:,k)==1)/size(dim_bic,1);
mean_eq12(3,k) = sum(dim_bic(:,k)¡3)/size(dim_bic,1);
end

   figure(1);
plot(nrows,mean_eq1);
title('d=1');
xlabel('n_y');
ylabel('F(1)');
ylim([0 1]);

   figure(2);
plot(nrows,mean_eq12);
title('d=1');
xlabel('n_y');
ylabel('F(1,2)');
ylim([0 1]);
```