Research Article GU J Sci 35(2): 748-763 (2022) DOI: 10.35378/gujs.735503

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

Robust Group Identification and Variable Selection in Sliced Inverse Regression Using Tukey's Biweight Criterion and Ball Covariance

Ali ALKENANI*

University of Al-Qadisiyah, Department of Statistics, Al-Diwaniyah, Iraq

Highlights

- This paper focuses on robust group identification and variable selection (RGIVS).
- The RGIVS method is proposed under sufficient dimension reduction settings.
- Numerically, good results are achieved through the proposed method.

Article Info

Received: 11 May 2020
Accepted: 19 May 2021

Keywords

Robust variable selection
Group identification
PACS
SIR
Ball correlation

Abstract

The SSIR-PACS is a group identification and a model-free variable selection method under sufficient dimension reduction (SDR) settings. It combined the Pairwise Absolute Clustering and Sparsity (PACS) with sliced inverse regression (SIR) methods to produce solutions with sparsity and the ability of group identification. However, the SSIR-PACS depends on classical estimates for dispersion and location, squared loss function, and non-robust weights for outliers. In this paper, a robust version of SSIR-PACS (RSSIR-PACS) is proposed. We replaced the squared loss by the criterion of Tukey's biweight. Also, the non-robust weights to outliers, which depend on Pearson's correlations, are substituted with robust weights based on recently developed ball correlation. Moreover, the estimates of the mean and covariance matrix are substituted by the median and ball covariance, respectively. The RSSIR-PACS is robust to outliers in both the response and covariates. According to the results of simulations, RSSIR-PACS produces very good results. If the outliers are existing, the efficacy of RSSIR-PACS is considerably better than the efficacy of the competitors. In addition, a robust criteria to estimate the structural dimension d is proposed. The RSSIR-PACS makes SSIR-PACS practically feasible. Also, we employed real data to demonstrate the utility of RSSIR-PACS.

1. INTRODUCTION

In regressions problems, a huge interest is gone to SDR in the last years [1-9]. Let Y and $\mathrm{X} = (\mathrm{x}_1, \dots, \mathrm{x}_p)^T$ are the outcome and the vector of covariates, respectively. The aim of SDR is to substitute X with orthogonal projection $P_S\mathrm{X}$ on to S of d-dimension , where $d < p$, without losing any information on $Y|\mathrm{X}$. SDR methods are searching for $\mathrm{S}_{Y|\mathrm{X}}$, where $\mathrm{S}_{Y|\mathrm{X}}$ is the intersection of all subspaces S such as $Y \perp\!\!\!\perp \mathrm{X}|P_S\mathrm{X}$ and $\perp\!\!\!\perp$ is the independency. Thus, $P_\beta\mathrm{X}$ summaries the information of X on Y and β is a basis of $\mathrm{S}_{Y|\mathrm{X}}$ [2].

A lot of methods are proposed to estimate $\mathrm{S}_{Y|\mathrm{X}}$. The SIR method is one of them [1]. It is applied in diverse areas like economics, informatics and finance. However, SIR produces linear combinations of all the original predictors and this makes difficulty in interpreting the results. To obtain better interpretability, the reduction of nonzero coefficients number in the SIR directions is very important.

Under least squares, many methods are proposed for better understanding. For examples, Lasso [10], SCAD [11], Elastic Net [12], group Lasso [13], adaptive Lasso [14], OSCAR [15], MCP [16] and PACS [17].

Under SIR framework, the thoughts of SIR were merged with the concepts of regularisation methods. For example, model-selection method for single-index models was proposed by [18]. Similarly, a method for

*e-mail: ali.alkenani@qu.edu.iq

determining the variables contribution is suggested by [4]. Furthermore, the Lasso is combined with SIR by [5] to obtain shrinkage SIR (SSIR). [6] proposed sparse SIR (SPSIR) via merging Lasso and LARS into SIR. [7] combined some of SDR methods with the idea of shrinkage estimation. To improve SIR ability to work when the covariates are highly correlated and $p > n$ where n is the sample size, a regularised SIR (RSIR) method is proposed by [19]. Lasso-SIR method is proposed by [9] for multiple index model and under $p > n$ settings. The authors have shown that Lasso-SIR estimates achieve optimal consistency rate. [20] proposed SSIR-PACS method. The author showed that the SSIR-PACS has advantages over the existing sparse SIR methods in its ability on group identification and variable selection(GIVS). However, the criterion of squared loss was employed between X and Y in SSIR-PACS. Also, the traditional estimates of the mean (μ) and covariance matrix of X (Σ_x) were used inside the squared loss. Moreover, the weighted penalty contains weights depend on Pearson's correlation (PC). It is known that the squared loss, the traditional estimates for μ and Σ_x, and PC are not robust to outliers.

The limitations of SSIR-PACS motivate us to propose RSSIR-PACS. The squared loss is substituted by the criterion of Tukey's biweight (T.B). Furthermore, the non-robust weights to outliers that depend on Pearson correlations are replaced with robust weights based on a recently developed ball correlation. Moreover, the estimates of μ and Σ_x are subsituted by the median and ball covariance(BCov), respectively. The RSSIR-PACS is robust to outliers in the response and covariates.

The rest of this paper is as follows. In Section 2, we give a summary of SIR and SSIR-PACS. RSSIR-PACS and a modification of [8] criteria for estimating the dimension are proposed in Sections 3. Simulations were carried out in Section 4. In Section 5, real data were analysed through the considered methods. In Section 6, the conclusions are given.

2. SIR AND SSIR-PACS METHODS

For estimating $S_{Y|X}$, [2] proposed SIR method. It requires $Z = \Sigma^{-\frac{1}{2}}\left(X - E(X)\right)$ that satisfies $E(Z|P_s Z) = P_s Z$, where $\Sigma_x = Cov(X)$ and s is a basis for $S_{Y|Z}$. This condition links $S_{Y|Z}$ and the inverse regression of Z on Y. The kernel matrix of SIR is $M = Cov\,[E(Z|Y)]$ and $Span(M) \subseteq S_{Y|Z}$.

Let $\bar{X}$ is the estimated mean of X. Also, let $\hat{Z} = \hat{\Sigma}^{-\frac{1}{2}}(X - \bar{X})$ is the estimate of Z, where $\hat{\Sigma}$ is the estimated Σ_x. Let h and n_y are the numbers of slices and observations in yth slice, respectively. Thus, $\widehat{M} = \sum_{y=1}^{h} \hat{f}_y \hat{Z}_y \hat{Z}_y^T$ is the estimated M, where $\hat{f}_y = n_y/n$ and $\hat{Z}_y$ is the average of Z in slice y. Let $\hat{\delta}_1 > \hat{\delta}_2 > \cdots > \delta_p \geq 0$ are the eigenvalues and $\hat{v}_1, \hat{v}_2, \ldots\ldots, \hat{v}_p$ are the corresponding eigenvectors 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 gives an estimator $span(\hat{\beta})$ of $S_{Y|X}$. Usually, $\hat{\beta} \in \mathbb{R}^{p \times d}$ is vector of nonzero coefficients. If there are a huge number of predictors, only the significant predictors are needed to obtain the 'sufficient predictors'. To this end, we need to combine the SIR with the regularisation techniques to make some coefficients of $\hat{\beta}$ going to 0's.

For the best understanding, SIR is formulated by [4] as a regression problem as

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

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 A and C values that minimise F, respectively. Then $span(\hat{A})$ is the space spanned by d largest eigenvectors of M. [5] rewrite F(A, C) as

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

where $B \in \mathbb{R}^{p\times d}$, $\hat{\beta}$ is B value which minimises (2) and $span(\hat{\beta}) = span\left(\hat{\Sigma}^{-\frac{1}{2}}\hat{A}\right)$ is the estimator of $S_{Y|X}$. After that, SSIR estimator of $S_{Y|X}$ is proposed by [5] as a $span(diag(\tilde{\alpha})\hat{\beta})$, where $\tilde{\alpha} = (\tilde{\alpha}_1, \dots, \tilde{\alpha}_p)^T \in \mathbb{R}^p$ are determined through minimizing

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

where $\hat{B}$ and $\hat{C} = \left(\hat{C}_1, \dots., \hat{C}_h\right)$ minimise (2).

The minimisation of (3) can be done according to algorithm of standard Lasso. Let

$$\tilde{Y} = vec(\hat{f}_1^{1/2}\hat{Z}_1, \dots, \hat{f}_h^{1/2}\hat{Z}_h) \in \mathbb{R}^{ph} \text{ and } \tilde{X} = \left(diag\left(\hat{B}\hat{C}_1\right)\hat{\Sigma}^{\frac{1}{2}}, \dots, diag\left(\hat{B}\hat{C}_h\right)\hat{\Sigma}^{\frac{1}{2}}\right)^T \in \mathbb{R}^{ph\times p},$$

where $vec(.)$ is an operator of matrix that stacks the columns of that matrix to a vector. The vector α is the Lasso estimator for regression $\tilde{Y}$ on $\tilde{X}$.

[17] proposed PACS for GIVS. The authors have explained the concept of "group identification" through the following lines " if the coefficients of two covariates are truly equal in magnitude, we would combine these two columns of the design matrix by their sum and if a coefficient were truly zero, we would exclude the corresponding covariates ".

The failure of the existing shrinkage SIR methods to do group identification, motivates [20] to incorporate PACS penalty into SIR to propose SSIR-PACS method. [20] proposes SSIR-PACS for GIVS under SDR settings. The SSIR-PACS is proposed as a solution of the following minimisation

$$\sum_{y=1}^{h} \left\| \hat{f}_y^{1/2}\hat{Z}_y - \hat{\Sigma}^{\frac{1}{2}} diag(\hat{B}\hat{C}_y)\alpha \right\|^2 + \lambda\{\sum_{i=1}^{p} \omega_i|\alpha_i| + \sum_{1\le i<k\le p} \omega_{jk(-)}|\alpha_k - \alpha_i| + \sum_{1\le i<k\le p} \omega_{ik(+)}|\alpha_k + \alpha_i|\}, \tag{4}$$

where ω_i are non-negative weights.

The minimisation of (4) contains two parts. The first is the SIR loss function. The second is PACS penalty, which consists of $\lambda\{\sum_{i=1}^{p} \omega_i|\alpha_i|\}$ that enables sparseness, $\lambda\{\sum_{1\le i<k\le p} \omega_{jk(-)}|\alpha_k - \alpha_i|\}$ that enables the coefficients with similar signs to be set as equal and $\lambda\{\sum_{1\le i<k\le p} \omega_{ik(+)}|\alpha_k + \alpha_i|\}$ that enables the coefficients of different signs to be set as equal in magnitude.

The optimisation of (4) can be done through a standard PACS algorithm. The vector α is the PACS estimator for the regression of $\tilde{Y}$ on $\tilde{X}$. Optimal λ can be selected via cross-validation (C.V) or AIC or BIC.

In summary, the SSIR-PACS is a two-step procedure. Firstly, SIR can be applied to obtain d, $\tilde{Y}$ and $\tilde{X}$. Secondly, compute α via PACS.

Choosing adaptive weights is an important issue in SSIR-PACS. The suitable weights can help SSIR-PACS to be an efficient procedure. In SSIR-PACS method, [20] used the adaptive weights, which were proposed in [17] as:

$$\omega_i = |\tilde{\alpha}_i|^{-1}, \omega_{ik(-)} = (1 - r_{ik})^{-1}|\tilde{\alpha}_k - \tilde{\alpha}_i|^{-1} \text{ and } \omega_{ik(+)} = (1 + r_{ik})^{-1}|\tilde{\alpha}_k + \tilde{\alpha}_i|^{-1} \text{ for } 1 \le i < k \le p, \tag{5}$$

where $\tilde{\alpha}$ is a $\sqrt{n}$ consistent estimator of α, such as SIR estimates or other shrinkage α estimates, and r_{ik} is PC.

3. THE PROPOSED ROBUST SSIR-PACS (RSSIR-PACS)

3.1. Methodology of RSSIR-PACS

SSIR-PACS method is proposed through combining PACS and SIR by [20]. The SIR method depends on the first and second moments estimators, which are not robust to outliers. A robust versions of SIR were proposed by [21, 22]. Moreover, on one side, the influence function of SIR was studied by [23]. On other side, [24] show that PACS is very sensitive to outliers. Although the nice behavior of SSIR-PACS was demonstrated by [20] under normal errors, the main drawback of SSIR-PACS is its high sensitivity to outliers. This encourages us to introduce RSSIR-PACS in this article.

In (4), the squared loss links the covariates with the response. Also, the traditional estimators of μ and Σ_x are used. Moreover, the weighted penalty contains weights that employ PC in their calculations. The criterion of least-squares, the traditional estimators of μ and Σ_x, and PC are very sensitive to outliers [24].

[25] showed that the loss function is robust to outliers in Y and X if its derivative is redescending. T.B function achieves this condition [26]. In this article, the squared loss is substituted by T.B function to obtain the robustness in Y and X and to choose the important predictors in robust way. Also, the estimator of μ is substituted by a robust estimator which is the median. The traditional estimator of Σ_x is substituted by BCov as a robust estimator. Moreover, the non-robust weights are replaced with robust weights that employ robust versions of correlations such as ball correlation. The proposed RSSIR-PACS minimise the following:

$$\sum_{y=1}^{h} \rho\left(\frac{\hat{f}_y^{1/2}\widehat{Ro}Z_y - \widehat{Ro}\Sigma^{\frac{1}{2}}\, diag(\hat{B}\hat{C}_y)\alpha}{\hat{\sigma}}\right) + \lambda\left\{\sum_{i=1}^{p} Ro\omega_i|\alpha_i| + \sum_{1\le i<k\le p} Ro\omega_{ik(-)}|\alpha_k - \alpha_i| + \sum_{1\le i<k\le p} Ro\omega_{ik(+)}|\alpha_k + \alpha_i|\right\}, \quad (6)$$

where $\lambda \geq 0$ is the tunning parameter and $Ro\omega$ is robust version of non-negative weights in (5). The $\widehat{Ro}Z_y$ and $\widehat{Ro}\Sigma^{\frac{1}{2}}$ are non-sensitive versions to outliers of $\hat{Z}_y$ and $\hat{\Sigma}^{\frac{1}{2}}$, respectively. Also, ρ refers to T.B function and $\hat{\sigma}$ is a robust version of σ. In this article, the median absolute deviation (MAD) is employed as an estimate for σ.

The function of T.B is as follows:

$$\rho_c(u) = \begin{cases} \left(\frac{c^2}{6}\right)\left\{1 - \left[1 - \left(\frac{u}{c}\right)^2\right]^3\right\} & if\ |u| \le c \\ \frac{c^2}{6} & if\ |u| \le c \end{cases}, \quad (7)$$

where c controls the robustness.

3.2. Robust Measures for Location and Dispersion

SIR depends on first and second moments estimators, which are not robust to outliers. In this article, non-sensitive versions to outliers of μ and Σ_x were employed inside SIR algorithm. As non-sensitive measures to outliers for location and dispersion, the median and BCov were employed, respectively. As a robust measure for dependency between two random vectors, the BCov was proposed by [27] as follows:

Let $\{U_k, V_k\}_{k=1}^n$ be i.i.d. sample of (U, V). Define $\delta_{ij,k}^U = I\{U_k \in \bar{B}_{\xi U}(U_i, U_j)\}$, where $I(.)$ is an indicator function, $\delta_{ij,kl}^U = \delta_{ij,k}^U \delta_{ij,l}^U$ and $\xi_{ij,klst}^U = (\delta_{ij,kl}^U + \delta_{ij,st}^U - \delta_{ij,ks}^U - \delta_{ij,lt}^U)/2$. $\xi_{ij,klst}^V$ is defined similar to $\xi_{ij,klst}^U$. The empirical BCov is as follows

$$\mathrm{BCov}_n(U,V) = \left(\frac{1}{n^6}\sum_{i,j,k,l,s,t=1}^n \xi_{ij,klst}^U \, \xi_{ij,klst}^V\right)^{1/2}. \tag{8}$$

Then, the ball correlation

$$\mathrm{BCor}(\boldsymbol{X}, Y) = \frac{\mathrm{BCov}(\boldsymbol{X},Y)}{\mathrm{BCov}^{1/2}(\boldsymbol{X},\boldsymbol{X})\times\mathrm{BCov}^{1/2}(Y,Y)} \tag{9}$$

and the sample ball correlation

$$\mathrm{BCor}(\boldsymbol{X}, Y) = \frac{\mathrm{BCov}_n(\boldsymbol{X},Y)}{\mathrm{BCov}_n^{1/2}(\boldsymbol{X},\boldsymbol{X})\times\mathrm{BCov}_n^{1/2}(Y,Y)} \quad . \tag{10}$$

For more details about BCov see [27, 28].

3.3. Choosing the Robust Weights

Choosing the suitable weights is substantial to PACS to be oracle procedure [17]. In their calculations, the weights in (5) employ PC. PC is very sensitive to outliers and thus the weights in (5) give unreliable results if certain types of outliers present. Consequently, PC should be replaced with robust correlation measure to obtain robust weights, which is an important issue.

In this article, the ball correlation is employed instead of PC as a robust correlation to get robust weights as follow:

$$Ro\omega_i = |\tilde{\alpha}_i|^{-1},\ Ro\omega_{ik(-)} = (1 - \mathrm{BCor}_{ik})^{-1}|\tilde{\alpha}_k - \tilde{\alpha}_i|^{-1} \text{ and } Ro\omega_{ik(+)} = (1 + \mathrm{BCor}_{ik})^{-1}|\tilde{\alpha}_k + \tilde{\alpha}_i|^{-1}$$
$$\text{for } 1 \le i < k \le p, \tag{11}$$

where BCor is the ball correlation. $\tilde{\alpha}$ is a robust initial estimate for α. Practically, it can be obtained through robust SIR or other robust shrinkage SDR methods estimates such as robust sparse MAVE (RSMAVE) [29].

3.4. Determination of d

In the estimation procedure of the proposed RSSIR-PACS, $d = \dim(S_{Y|X})$ is assumed as known. In practice, we need to estimate d through data. Many methods are proposed to determine d. See, for example, [1], [30-32] and [8]. [19] adopted a criterion suggested by [8]. [8] proposed to determine d through the nonzero eigenvalues number of $Cov\,[E(X|Y)]$ matrix, or equivalently, number of eigenvalues of the matrix $\Omega = Cov\,[E(X|Y)] + I_p$ that are greater than 1, where I_p refers to a identity matrix of p-dimension. Let $\hat{\Omega}$ is the estimated version of Ω and $\hat{\delta}_1, \ldots, \hat{\delta}_p$ are the eigenvalues of it, k is the number of $\hat{\delta}_i > 1$, and C_n^* is a constant. [8] suggested the following estimator of d,

$$\hat{d} = arg \max_{m\in\{0,1,\ldots,p-1\}} \left\{\frac{n}{2}\sum_{i=1+\min(k,m)}^{p}\left(log(\hat{\delta}_i) + 1 - \hat{\delta}_i\right) - \frac{C_n^* m(2p-m+1)}{2}\right\} \tag{12}$$

Several forms are recommended for C_n^* by the authors. [19] suggested $C_n^* = \log(n)\, h/n$ in their simulations.

In this article, a robust version of $\hat{d}$ in (12) is proposed as the following:
Under Z-scale and without losing of generality because of $\mathrm{S}_{Y|\mathrm{X}} = \Sigma^{-\frac{1}{2}} \mathrm{S}_{Y|\mathrm{Z}}$, we estimates d through the number of eigenvalues of robust matrix $\mathrm{Ro}\Omega = RoM + I_p$ that are greater than 1, where RoM is a robust estimate of M as follows:

$$\widehat{RoM} = \sum_{y=1}^{h} \hat{f}_y \, \widehat{RoZ}_y \widehat{RoZ}_y{}^T, \tag{13}$$

where

$$\widehat{RoZ}_y = \widehat{\mathrm{BCov}}_n{}^{-\frac{1}{2}} \, (X - median(X)) \tag{14}$$

Let $\widehat{\mathrm{Ro}\Omega}$ is a robust version of $\mathrm{Ro}\Omega$ and $\hat{\gamma}_1, \dots, \hat{\gamma}_p$ are the eigenvalues of it, k is the number of $\hat{\gamma}_i > 1$. We suggest the following formula to estimate d,

$$\hat{d} = arg \max_{m \in \{0,1,\dots,p-1\}} \left\{ \frac{n}{2} \sum_{i=1+\min(k,m)}^{p} (log(\hat{\gamma}_i) + 1 - \hat{\gamma}_i) - \frac{C_n^* m(2p-m+1)}{2} \right\} \tag{15}$$

In the simulation section, we used the formula of C_n^* which is proposed by [19] .

4. SIMULATION STUDY

In this section, we compared RSSIR-PACS with SSIR-PACS [20] and RSMAVE [29] through four examples.
For measuring the prediction accuracy, the trace correlation r^*[33] is employed. Let $S(A)$ and $S(B)$ refer to column space spanned by two $p \times d$ full column rank matrices. Let $P_A = A(A^TA)^{-1}A^T$ and $P_B = B(B^TB)^{-1}B^T$ are projection matrices onto $S(A)$ and $S(B)$, respectively. Thus, $r^* = \sqrt{\frac{1}{d} tr(P_A P_B)}$, where, $0 \le r^* \le 1$. If r^* is close to 1, and $S(A)$ is close to $S(B)$.
To evaluate the ability of selection the variables accurately, the true and false positive rates which are denoted by TPR and FPR are used, respectively. TPR is the proportion of predictors which are correctly identified as active to the predictors which are truly active. FPR is the proportion of predictors which are falsely identified as active to inactive predictors. The best method according to variable selection concept is the method that has closer TPR to 1 and closer FPR to 0. In PACS, λ can be chosen via tenfold Cross-validation.

4.1. Direction Estimation and Variable Selection

The data is generated according to the settings in the following examples:

Example 1: 200 datasets contain $n = 100$ and 200 observations are simulated from $Y = 2\mathrm{x}_1 + 2\mathrm{x}_2 + 2\mathrm{x}_3 + \varepsilon$. The $\beta = (2,2,2,0,0,0,0,0,0,0)^T$ and $X \in \mathbb{R}^{10}$ with $d = 1$. The covariates $\mathrm{x}_1, \mathrm{x}_2$ and x_3 are correlated with pairwise correlation $r = 0.75$. The covariates $\mathrm{x}_4, \mathrm{x}_5, \mathrm{x}_6, \mathrm{x}_7, \mathrm{x}_8, \mathrm{x}_9$ and x_{10} are uncorrelated.

Example 2: 200 datasets contain $n = 100$ and 200 observations are simulated from $Y = \exp(\mathrm{x}_1 + \mathrm{x}_2 + \mathrm{x}_3 + 0.5\mathrm{x}_4 + \mathrm{x}_5 + 2\mathrm{x}_6) + \varepsilon$, where $\beta = (1,1,1,0.5,1,2,0,0,0,0)^T$ and $X \in \mathbb{R}^{10}$ with $d = 1$. The covariates $\mathrm{x}_1, \mathrm{x}_2$ and x_3 are pairwise correlated with $r = 0.35$, while the covariates $\mathrm{x}_4, \mathrm{x}_5$ and x_6 are pairwise correlated with $r = 0.75$. The covariates $\mathrm{x}_7, \mathrm{x}_8, \mathrm{x}_9$ and x_{10} are uncorrelated.

Example 3: 200 datasets contain $n = 100$ and 200 observations are simulated from $Y = 5\cos(2\mathrm{x}_1 + 2\mathrm{x}_2 + 2\mathrm{x}_3 + \mathrm{x}_4 + \mathrm{x}_5) + \exp(-(2\mathrm{x}_1 + 2\mathrm{x}_2 + 2\mathrm{x}_3 + \mathrm{x}_4 + \mathrm{x}_5)^2) + \varepsilon$, where $\beta = (2,2,2,1,1,0,0,0,0,0)^T$ and $X \in \mathbb{R}^{10}$ with $d = 1$. The covariates $\mathrm{x}_1, \mathrm{x}_2$ and x_3 are highly pairwise

correlated with $r = 0.75$. Also, the covariates x_4 and x_5 are highly pairwise correlated with $r = 0.75$. The covariates x_6, x_7, x_8, x_9 and x_{10} are uncorrelated.

Example 4: 200 datasets contain $n = 100$ and 200 observations are simulated from the model $Y = \frac{2x_1+2x_2+2x_3}{0.5+(1.5+2x_6+2x_7+2x_8)} + \varepsilon$, where $\beta_1 = (2,2,2,0,0,0,0,0)^T$, $\beta_2 = (0,0,0,0,0,2,2,2)^T$ and $X \in \mathbb{R}^8$ with $d = 2$. For β_1, the covariates x_1, x_2 and x_3 are highly pairwise correlated with $r = 0.75$, while the rest covariates are uncorrelated. For β_2, the first five covariates are uncorrelated, while the covariates x_6, x_7 and x_8 are highly pairwise correlated with $r = 0.75$.

For the above examples, four sampling distributions for x_i and ε are assumed:
1. $N(0,1)$, standard normal.
2. $t_3/\sqrt{3}$, t-distribution with 3 degree of freedom.
3. $0.95\,N(0,1) + 0.05\,N(0, 10^2)$.
4. $0.95\,N(0,1) + 0.05\mathrm{U}(-50, 50)$, 95% from $N(0,1)$ and 5% from uniform distribution.

Table 1. *r^*, TPR and FPR for Example 1*

Dist.	n	Criterion	SSIR-PACS	RSMAVE	RSSIR - PACS
Dist.1	100	r^*Mean (s.e)	0.849(0.148)	0.827(0.168)	0.839(0.160)
		TPR	0.846	0.813	0.829
		FPR	0.129	0.150	0.135
	200	r^* Mean (s.e)	0.948 (0.087)	0.933(0.101)	0.941(0.097)
		TPR	0.957	0.942	0.954
		FPR	0.075	0.095	0.088
Dist.2	100	r^* Mean (s.e)	0.816(0.160)	0.873(0.135)	0.894(0.132)
		TPR	0.822	0.879	0.900
		FPR	0.189	0.173	0.157
	200	r^* Mean (s.e)	0.907(0.095)	0.966(0.039)	0.985(0.031)
		TPR	0.894	0.979	0.998
		FPR	0.114	0.100	0.094
Dist.3	100	r^* Mean (s.e)	0.629(0.241)	0.844(0.148)	0.878(0.140)
		TPR	0.727	0.802	0.856
		FPR	0.413	0.179	0.149
	200	r^* Mean (s.e)	0.656(0.216)	0.932(0.100)	0.969(0.096)
		TPR	0.698	0.939	0.967
		FPR	0.383	0.087	0.070
Dist.4	100	r^* Mean (s.e)	0.421(0.273)	0.821(0.167)	0.859(0.157)
		TPR	0.672	0.791	0.806
		FPR	0.638	0.182	0.155
	200	r^* Mean (s.e)	0.366(0.287)	0.929(0.114)	0.954(0.109)

		TPR	0.590	0.921	0.943
		FPR	0.540	0.099	0.083

Table 2. *r^*, TPR and FPR for Example 2*

Dist.	n	Criterion	SSIR-PACS	RSMAVE	RSSIR - PACS
Dist.1	100	r^* Mean (s.e)	0.829(0.157)	0.780(0.178)	0.818(0.169)
		TPR	0.830	0.795	0.795
		FPR	0.139	0.155	0.142
	200	r^* Mean (s.e)	0.930(0.087)	0.912(0.107)	0.929(0.095)
		TPR	0.949	0.929	0.944
		FPR	0.075	0.099	0.095
Dist.2	100	r^* Mean (s.e)	0.848(0.162)	0.872(0.138)	0.895(0.128)
		TPR	0.855	0.870	0.895
		FPR	0.189	0.180	0.161
	200	r^* Mean (s.e)	0.930(0.109)	0.965(0.048)	0.991(0.038)
		TPR	0.940	0.969	0.994
		FPR	0.130	0.112	0.097
Dist.3	100	r^* Mean (s.e)	0.635(0.246)	0.839(0.150)	0.857(0.145)
		TPR	0.715	0.789	0.820
		FPR	0.425	0.174	0.151
	200	r^* Mean (s.e)	0.659(0.226)	0.931(0.107)	0.954(0.098)
		TPR	0.690	0.951	0.949
		FPR	0.392	0.095	0.077
Dist.4	100	r^* Mean (s.e)	0.427(0.278)	0.814(0.162)	0.844(0.157)
		TPR	0.666	0.780	0.798
		FPR	0.649	0.176	0.160
	200	r^* Mean (s.e)	0.361(0.296)	0.924(0.120)	0.940(0.114)
		TPR	0.570	0.914	0.930
		FPR	0.559	0.110	0.090

Table 3. r*, *TPR and FPR for Example 3*

Dist.	n	Criterion	SSIR-PACS	RSMAVE	RSSIR - PACS
Dist.1	100	r^* Mean (s.e)	0.796(0.190)	0.761(0.204)	0.779(0.193)
		TPR	0.801	0.760	0.786
		FPR	0.160	0.180	0.167
	200	r^* Mean (s.e)	0.896(0.094)	0.871(0.107)	0.904(0.099)
		TPR	0.925	0.903	0.926
		FPR	0.093	0.150	0.100
Dist.2	100	r^* Mean (s.e)	0.807(0.183)	0.832(0.154)	0.866(0.140)
		TPR	0.836	0.847	0.880
		FPR	0.207	0.193	0.172
	200	r^* Mean (s.e)	0.900(0.118)	0.928(0.057)	0.960(0.045)
		TPR	0.922	0.955	0.979
		FPR	0.144	0.127	0.104
Dist.3	100	r^* Mean (s.e)	0.599(0.262)	0.785(0.169)	0.820(0.155)
		TPR	0.698	0.769	0.799
		FPR	0.435	0.188	0.168
	200	r^* Mean (s.e)	0.630(0.240)	0.909(0.121)	0.930(0.108)
		TPR	0.664	0.916	0.934
		FPR	0.405	0.112	0.091
Dist.4	100	r^* Mean (s.e)	0.398(0.285)	0.779(0.168)	0.805(0.168)
		TPR	0.648	0.766	0.786
		FPR	0.661	0.194	0.172
	200	r^* Mean (s.e)	0.330(0.305)	0.889(0.136)	0.910(0.127)
		TPR	0.559	0.890	0.915
		FPR	0.569	0.124	0.103

Table 4. *r^*, TPR and FPR for Example 4*

Dist.	n	Criterion	SSIR-PACS	RSMAVE	RSSIR - PACS
Dist.1	100	r^* Mean (s.e)	0.789(0.144)	0.764(0.150)	0.786(0.146)
		TPR	0.787	0.768	0.783
		FPR	0.173	0.210	0.181
	200	r^* Mean (s.e)	0.900(0.127)	0.859(0.140)	0.883(0.133)
		TPR	0.930	0.879	0.897
		FPR	0.090	0.140	0.120
Dist.2	100	r^* Mean (s.e)	0.800(0.160)	0.827(0.148)	0.849(0.140)
		TPR	0.841	0.873	0.890
		FPR	0.273	0.254	0.230
	200	r^* Mean (s.e)	0.886(0.129)	0.926(0.103)	0.940(0.097)
		TPR	0.900	0.967	0.972
		FPR	0.180	0.165	0.148
Dist.3	100	r^* Mean (s.e)	0.610(0.260)	0.759(0.148)	0.795(0.132)
		TPR	0.677	0.742	0.779
		FPR	0.375	0.210	0.198
	200	r^* Mean (s.e)	0.650(0.179)	0.862(0.130)	0.897(0.123)
		TPR	0.680	0.907	0.925
		FPR	0.339	0.160	0.140
Dist.4	100	r^* Mean (s.e)	0.420(0.287)	0.766(0.147)	0.790(0.140)
		TPR	0.670	0.760	0.775
		FPR	0.621	0.230	0.201
	200	r^* Mean (s.e)	0.395(0.299)	0.856(0.138)	0.875(0.130)
		TPR	0.561	0.890	0.909
		FPR	0.517	0.165	0.148

From Tables 1, 2, 3 and 4, we can notice the following:
1. In case of Dist.1, the performance of SSIR-PACS exceeds the performance of RSSIR-PACS and RSMAVE methods.
2. For the rest cases, the performance of SSIR-PACS is negatively affected while the RSSIR-PACS and RSMAVE methods have good and stable performance. Also, the RSSIR-PACS has the best performance for all the samples size.
3. For RSSIR-PACS estimates and under different settings, the variations in the comparative criteria values are close. While, the variations are big for the SSIR-PACS estimates under different considered settings.

Table 5. *The computing time for different methods (in seconds) for 200 datasets under the settings of example 1 with* $n = 200$

Dist.	Dist.1	Dist.2	Dist.3	Dist.4
SSIR-PACS	39	42	45	46
RSMAVE	64	66	66	67
RSSIR -PACS	39	41	45	46

Table 6. *The computing time for different methods (in seconds) for 200 datasets under the settings of example 2 with* $n = 200$

Dist.	Dist.1	Dist.2	Dist.3	Dist.4
SSIR-PACS	40	42	45	47
RSMAVE	66	67	66	67
RSSIR -PACS	40	42	46	47

Table 7. *The computing time for different methods (in seconds) for 200 datasets under the settings of example 3 with* $n = 200$

Dist.	Dist.1	Dist.2	Dist.3	Dist.4
SSIR-PACS	41	43	45	46
RSMAVE	66	67	67	67
RSSIR -PACS	41	43	45	46

Table 8. *The computing time for different methods (in seconds) for 200 datasets under the settings of example 4 with* $n = 200$

Dist.	Dist.1	Dist.2	Dist.3	Dist.4
SSIR-PACS	42	44	47	47
RSMAVE	67	67	67	67
RSSIR -PACS	41	44	47	47

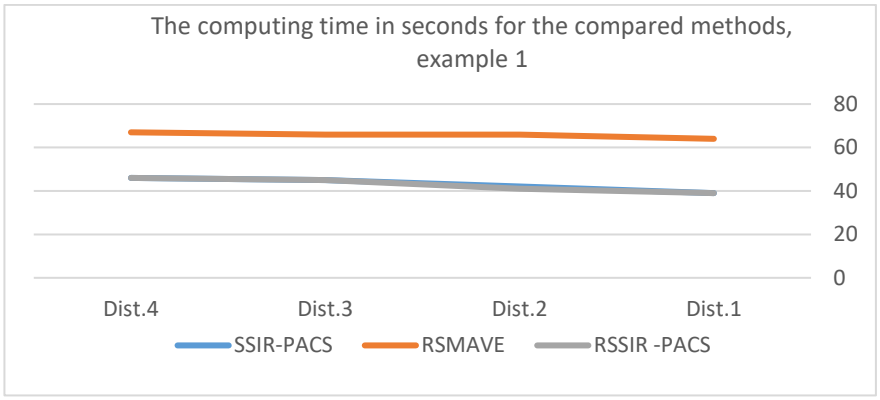

Figure 1. *The computing time for different methods (in seconds) under the settings of example 1*

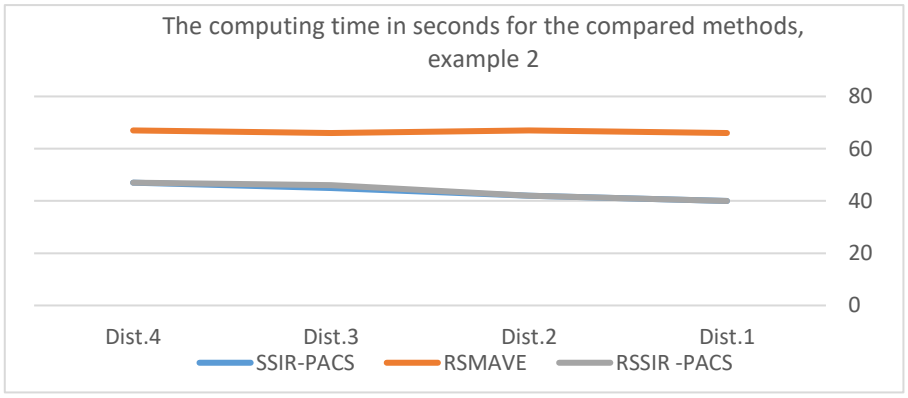

Figure 2. *The computing time for different methods (in seconds) under the settings of example 2*

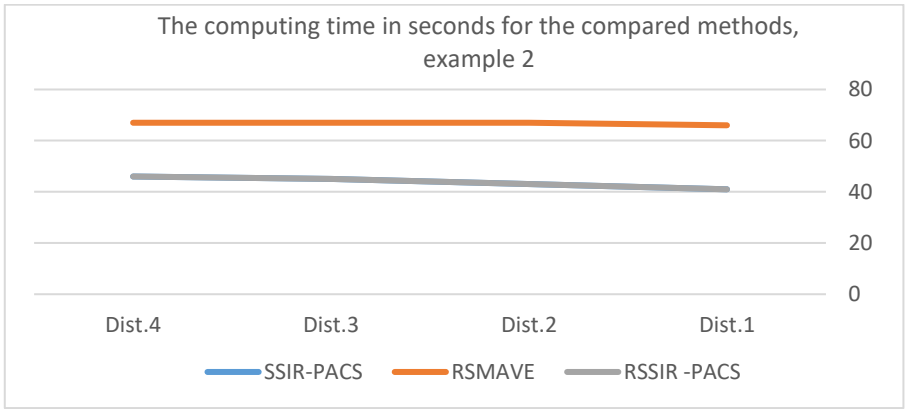

Figure 3. *The computing time for different methods (in seconds)under the settings of example 3*

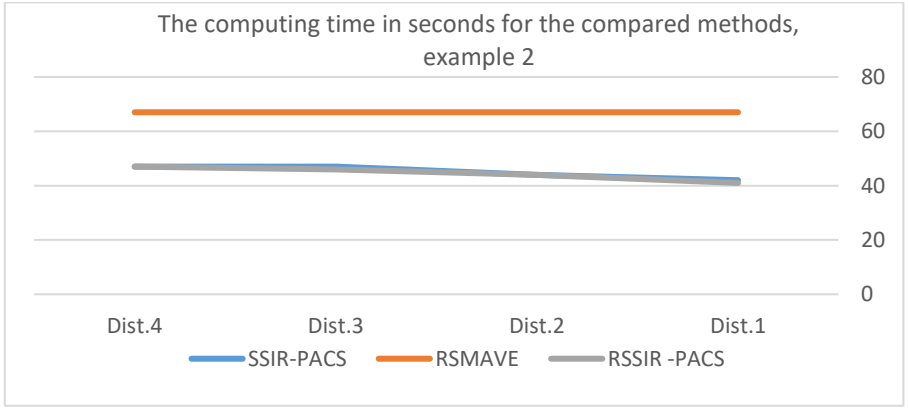

Figure 4. *The computing time for different methods (in seconds) under the settings of example 4*

Later, the computation time was taken into account. Tables 5, 6, 7, 8 and Figures 1, 2, 3 and 4 show the computation time for different methods (in second) under the settings of examples 1, 2, 3 and 4, respectively. From these tables and figures, the computing time for RSSIR -PACS and SSIR -PACS methods is significantly lower than that of RSMAVE method. Moreover, it is obvious that the RSMAVE is time consuming.

4.2. Estimation of d

In this section, the ability of proposed robust formula in (15) to estimate d is checked. We generate the data as in Example 4 Settings according to the above mentioned four sampling distributions with $n = 100\ and\ 200$, where $d = 2$. For each sample size, 200 datasets are generated. Table 9 reports the frequency of $\hat{d}$ out of 200 datasets. For the sake of comparison, the results according to [8]'s formula are also reported. It is clear that our proposed robust method gives very consistent estimation for all settings. It does well under Dist2, Dist3 and Dist4 settings although a bit worse than those under Dist1. The proposed robust method in (15) significantly exceeds the method of [8] for Dist3 and Dist4 according to the frequency of $\hat{d}$.

Table 9. *Frequency of $\hat{d}$ out of 200 datasets*

Dist.	n	Frequency of $\hat{d}$ according to proposed robust method					Frequency of $\hat{d}$ according to [8] method				
		$d=1$	$d=2$	$d=3$	$d=4$	$d\geq 5$	$d=1$	$d=2$	$d=3$	$d=4$	$d\geq 5$
Dist.1	100	9	**160**	30	1	0	12	**160**	28	0	0
	200	1	**185**	13	1	0	2	**185**	12	1	0
Dist.2	100	12	**153**	33	2	0	7	**144**	45	4	0
	200	1	**179**	19	1	0	7	**175**	17	1	0
Dist.3	100	29	**105**	45	15	6	45	**89**	44	18	4
	200	2	**140**	47	10	1	14	**120**	50	11	5
Dist.4	100	41	**104**	30	8	17	49	**89**	45	11	6
	200	15	**138**	25	8	14	21	**120**	31	21	7

5. REAL DATA

The compared methods are applied to pollution data(PD). The PD [34] is analysed through the compared methods.

We centered y and standardised the predictors. The performance of RSSIR-PACS is checked through analysing the PD after including some outliers in Y and X. The data are contaminated with (5%, 10%, 15%, and 20%) of the observations that come from multivariate t_3.

To check the estimation precision of RSSIR-PACS, the correlation between each estimated direction is computed through the considered methods and the estimated directions of SSIR-PACS without outliers.

We refer to it as $Corr(\widehat{\beta}, \hat{\beta}_{\text{SSIR-PACS},0})$. Also, the effective model size (EMS) after accounting for equality of absolute coefficient estimates is reported.

Pollution data (P.D)

The data is collected by [34] to study the effects of the weather, socioeconomic and pollution indicators on mortality rate. The P.D are available at (http://www4.stat.ncsu.edu/~boos/var.select/pollution.html). The P.D consists of $n = 60$ observations and $p = 15$. The y is the mortality rate. The covariates are ($x_1 =$ Avar. annual precipitation), ($x_2 =$ Avar. temperature $-$ January), ($x_3 =$ Avar. temperature $-$ July), ($x_4 =$ Percent of age $\geq$ 65 years), ($x_5 = Ratio\ of$ population / household), ($x_6 =$ school years), ($x_7 =$ Percent of facilated housing), ($x_8 =$ Ratio of population / mile), ($x_9 =$ Percent of non $-$ white), ($x_{10} =$ % of employment), ($x_{11} =$ % of families with income ≤ 3000), ($x_{12} =$ % of hydrocarbons), ($x_{13} =$ % of oxides of nitrogen), ($x_{14} =$ % of sulfur dioxide) and ($x_{15} =$ % of humidity).

Table 10. *The $Corr(\widehat{\beta}, \hat{\beta}_{SSIR-PACS,0})$ and EMS based on the pollution data*

Methods		Outliers %				
		0	5	10	15	20
$Corr(\widehat{\beta}, \hat{\beta}_{\text{SSIR-PACS},0})$	SSIR-PACS	1	0.9065	0.8079	0.6811	0.5743
	RSMAVE	0.9687	0.9676	0.9167	0.8057	0.7137
	RSSIR-PACS	0.9739	0.9722	0.9615	0.9395	0.9081
EMS	SSIR-PACS	5	6	7	9	9
	RSMAVE	5	5	6	7	7
	RSSIR-PRMVN	5	5	5	5	5

From Table 10 and according to the results of the $Corr(\widehat{\beta}, \hat{\beta}_{\text{SSIR-PACS},0})$ and EMS, the following findings are noted:

1. In case of no outliers, the RSSIR-PACS's performance is close to SSIR-PACS's performance. In addition, the performance of RSMAVE is worse than the performance of RSSIR-PACS according to the comparative criteria.
2. In case of there are outliers, SSIR-PACS's performance is negatively affected. The high sensitivity of SSIR-PACS to outliers is obvious, and Table 6 confirms this fact. From other side, RSSIR-PACS produces consistent and stable results, even with 20% of contamination. The performance of RSMAVE is less efficient than the performance of RSSIR-PACS for all the contamination percentages. The robustness of RSMAVE is less than the robustness of RSSIR-PACS because it is robust to outliers in Y only. The performance of RSMAVE worsens as the percentage of contamination increases beyond 0.10 while the performance of RSSIR-PACS is still the best for all the percentages of contamination.

6. CONCLUSION

In this article, we propose RSSIR-PACS method. Under SDR settings, it is a robust group identification and model-free variable selection method. Numerically, the preference of RSSIR-PACS has confirmed through the results of simulations when the outliers are exist in both Y and X. Also, Also, RSSIR-PACS is good competitor to SSIR-PACS in case of no contamination. Simulations and PD analysis show that RSSIR-PACS has high predictive accuracy and high ability for identifying relevant groups**.** In addition, a robust modification of [8] criteria to estimate the structural dimension d is proposed. The RSSIR-PACS idea can be extended to another SDR methods such as MAVE [3]. Also, we can extend the idea to models where y takes discrete values.

CONFLICTS OF INTEREST

No conflict of interest was declared by the author.

REFERENCES

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

[2] Cook, R., "Regression graphics: ideas for studying the regression through graphics", New York, Wily, (1998).

[3] Xia, Y., Tong, H., Li, W., Zhu, L., "An adaptive estimation of dimension reduction space", Journal of the Royal Statistical Society: Series B, 64: 363–410, (2002).

[4] Cook, R., "Testing predictor contributions in sufficient dimension reduction", Annals of Statistics, 32: 1061–92, (2004).

[5] Ni, L., Cook, R. D., Tsai, C. L., "A note on shrinkage sliced inverse regression", Biometrika, 92: 242–247, (2005).

[6] Li, L., Nachtsheim, C. J., "Sparse sliced inverse regression", Technometrics, 48: 503–510, (2006).

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

[8] Zhu, L., Miao, B., Peng, H., "On sliced inverse regression with large dimensional covariates", Journal of American Statistical Association, 101: 630–643, (2006).

[9] Lin, Q., Zhao, Z., Liu, J., "Sparse sliced inverse regression via lasso", Journal of the American Statistical Association, 114: 1726–1739, (2019).

[10] Tibshirani, R., "Regression shrinkage and selection via the Lasso", Journal of the Royal Statistical Society: Series B, 58: 267–288, (1996).

[11] Fan, J., Li, R. Z., "Variable selection via non-concave penalized likelihood and its oracle properties", Journal of the American Statistical Association, 96: 1348–1360, (2001).

[12] Zou, H., Hastie, T., "Regularization and variable selection via the elastic net", Journal of the Royal Statistical Society: Series B, 67: 301–320, (2005).

[13] Yuan, M., Lin, Y., "Model selection and estimation in regression with grouped variables", Journal of the Royal Statistical Society: Series B, 68: 49–67, (2006).

[14] Zou, H., "The adaptive Lasso and its oracle properties", Journal of the American Statistical Association, 101: 1418–1429, (2006).

[15] Bondell, H. D., Reich, B. J., "Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR", Biometrics, 64: 115–123, (2008).

[16] Zhang, C. H., "Nearly unbiased variable selection under minimax concave penalty", Annals of Statistics, 38: 894–942, (2010).

[17] Sharma, D. B., Bondell, H. D., Zhang, H. H., "Consistent group identification and variable selection in regression with correlated predictors", Journal of Computational and Graphical Statistics, 22: 319–340, (2013).

[18] Naik, P. A., Tsai, C. L., "Single-index model selections", Biometrika, 88: 821–832, (2001).

[19] Li, L., Yin, X., "Sliced Inverse Regression with regularizations", Biometrics, 64: 124–131, (2008).

[20] Alkenani, A., "Group identification and variable selection in sliced inverse regression with correlated predictors", Journal of Modern Applied Statistical Methods, (2020).

[21] Gather, U., Hilker, T., Becker, C., "A note on outlier sensitivity of sliced inverse regression", Statistics, 36: 271–281, (2002).

[22] Yohai, V., Sertter, M., "A robust proposal for sliced inverse regression", International conference on robust statistics, abstract, (2005).

[23] Prendergast, L., "Influence functions for sliced inverse regression", Scandinavian Journal of Statistics, 32: 385–404, (2005).

[24] Alkenani, A., Dikheel, T., "Robust Group Identification and Variable Selection in Regression", Journal of Probability and Statistics, 2017, Paper ID 2170816, 8, (2017).

[25] Rousseeuw, P., Yohai, V., "Robust regression by means of s-estimators", Robust and Nonlinear Time Series Analysis, 256-272, (1984).

[26] Tukey, J. W., "A survey of sampling from contaminated distributions", Contributions to Probability and Statistics, 2: 448–485, (1960).

[27] Pan, W., Wang, X., Xiao, W., Zhu, H., "A generic sure independence screening procedure", Journal of American Statistical Association 1–29, (2018). (just-accepted).

[28] Zhang, J., Chen, X., "Robust sufficient dimension reduction via ball covariance", Computational Statistics and Data Analysis, 140: 144–154, (2019).

[29] Yao, W., Wang, Q., "Robust variable selection through MAVE", Computational Statistics and Data Analysis, 63: 42–49, (2013).

[30] Schott, J. R., "Determining the dimensionality in sliced inverse regression", Journal of the American Statistical Association, 89: 141–148, (1994).

[31] Bura, E., Cook, R. D., "Extending sliced inverse regression: The weighted chi-squared test", Journal of the American Statistical Association, 96: 996–1003, (2001).

[32] Cook, R. D., Yin, X., "Dimension reduction and visualization in discriminant analysis", Australian and New Zealand Journal of Statistics, 43: 147–177, (2001).

[33] Zhu, Y., Zeng, P., "Fourier methods for estimating the central subspace and the central mean subspace in regression", Journal of the American Statistical Association, 101: 1638–1651, (2006).

[34] McDonald, G. C., Schwing, R. C., "Instabilities of regression estimates relating air pollution to mortality", Technometrics, 15: 463–481, (1973).