Communications for Statistical Applications and Methods
2021, Vol. 28, No. 3, 267–279

https://doi.org/10.29220/CSAM.2021.28.3.267
Print ISSN 2287-7843 / Online ISSN 2383-4757

Fused inverse regression with multi-dimensional responses

Youyoung Cho[a], Hyoseon Han[a], Jae Keun Yoo[1,a]

[a]Department of Statistics, Ewha Womans University, Korea

Abstract

A regression with multi-dimensional responses is quite common nowadays in the so-called big data era. In such regression, to relieve the curse of dimension due to high-dimension of responses, the dimension reduction of predictors is essential in analysis. Sufficient dimension reduction provides effective tools for the reduction, but there are few sufficient dimension reduction methodologies for multivariate regression. To fill this gap, we newly propose two fused slice-based inverse regression methods. The proposed approaches are robust to the numbers of clusters or slices and improve the estimation results over existing methods by fusing many kernel matrices. Numerical studies are presented and are compared with existing methods. Real data analysis confirms practical usefulness of the proposed methods.

Keywords: central subspace, fused sliced inverse regression, multivariate regression, pooled approach, sufficient dimension reduction

1. Introduction

With the recent advances in computing technology, it has become possible to perform calculations and modeling on vast amounts of data that were difficult before. With high-dimensional data modeling, the so-called curse of dimension is often faced, and it is one of main issues in such data analysis.

In regression of $\mathbf{Y} \in \mathbb{R}^u|\mathbf{X} \in \mathbb{R}^p$, sufficient dimension reduction (SDR) seeks to replace the original p-dimensional predictors $\mathbf{X}$ by its lower-dimensional predictor $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ without loss of information on the conditional distribution of $\mathbf{Y} \in \mathbb{R}^u|\mathbf{X} \in \mathbb{R}^p$, where $u \geq 1$, $p \geq 2$ and $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$ with $d \leq p$. It is equivalently stated as the following independent statement:

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}, \tag{1.1}$$

where $\perp\!\!\!\perp$ stands for statistical independence.

For further usage, for $p \times q$ matrix $\mathbf{M}$, we define a notation $\mathcal{S}(\mathbf{M})$ as a subspace spanned by the columns of $\mathbf{M}$. Multiple $\boldsymbol{\eta}$ to satisfy (1.1) can exist, and then it is natural to choose the minimal one among them. The subspace spanned by the minimal one is called the *central subspace* $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. Throughout the rest of the paper, $\boldsymbol{\eta}$ and d will stand for an orthonormal basis matrix and the structural dimension of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. The d-dimensional linearly transformed predictor $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ is called sufficient predictors. For further insights about SDR, readers are recommended to read Yoo (2016a, b).

When the dimension of $\mathbf{Y}$, u is bigger than or equal to 2, the regression is called multivariate regression. The demand of multivariate regression has rapidly grown according to advent of big

[1] Corresponding author: Department of Statistics, Ewha Womans University, 11-1 Daehyun-Dong Seodaemun-Gu, Seoul 120-750, Korea. E-mail: peter.yoo@ewha.ac.kr

Published 31 May 2021 / journal homepage: http://csam.or.kr

data era. Repeated measures, longitudinal data, or curve or time series data often appear in big data, and the analysis of such data is difficult due to high-dimensionality of predictors. For example, the total number of regression coefficients to estimate in a classical multivariate regression of $\mathbf{Y} = (Y_1, \ldots, Y_u)^{\mathrm{T}} | \mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is equal to $p \times u$ and multiply increase with adding more responses. Therefore, to avoid this complexity in the analysis, a proper dimension reduction of $\mathbf{X}$ is important, and SDR provides a good solution to the problem. So far, various SDR methodologies have been developed to estimate $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ in multivariate regression. Various SDR methodologies for multivariate regression have been proposed (Lee *et al.*, 2019; Setodji and Cook, 2004; Yin and Bura, 2006; Yoo, 2008, 2009; Yoo and Cook, 2007; Yoo *et al.*, 2010). In Setodji and Cook (2004) and Yoo *et al.* (2010), an inverse regression approach, called K-means inverse regression (KIR) and K-means average variance estimation, is adopted, while the other methods combine the information from the coordinate regression of $Y_k|\mathbf{X}$, $k = 1, \ldots, u$, where Y_k is the kth coordinate of $\mathbf{Y} = (Y_1, \ldots, Y_u)^{\mathrm{T}}$.

Here, our interest is given in KIR, which is one of the widely used SDR method in multivariate regression. The key-step in KIR is to do K-means clustering $\mathbf{Y}$. However, different numbers of the clusters provide different outcome by KIR, so it often causes a question regarding how many clusters must be used in KIR. So far, there is no thumb rule for it. To overcome similar issue in sliced inverse regression (SIR) (Li, 1991), Cook and Zhang (2014) propose a fused approach to combine all results from various numbers of slices, and they show that it provides robust estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ to the number of slices and improves the estimation accuracy of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. If this fusing idea is employed in KIR, we have potential advantages to have robust results to the number of clusters and to improve the estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ like SIR. This is the main purpose of the paper. For this, we propose two fused approaches. The first one is to fuse the results based on hierarchical clustering algorithm recommended by Yoo *et al.* (2020), not K-means clustering algorithm in KIR. Another one is to fuse all results by the fused SIR application on the coordinate regression of $Y_k|\mathbf{X}$, $k = 1, \ldots, u$.

The organization of the paper is as follows. Sliced inverse regression and hierarchical inverse regression are reviewed in Section 2. Section 3 is devoted to proposing pooled sliced inverse regression for multivariate regression and two fused approaches for multivariate regression. In Section 4, numerical studies and real data examples are presented. We summarize our work in Section 5.

2. Literature review: sliced and hierarchical inverse regressions

2.1. Sliced inverse regression

Understanding sliced inverse regression (SIR) (Li, 1991) is essential for methodological development, because its main methodological development is based on SIR.

Letting $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{X})$, Li (1991) showed that $\mathbf{\Sigma}^{-1}E(\mathbf{X}|\mathbf{Y}) \in \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, if $E\left(\mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\right)$ is linear in $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$. Defining that $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - \bar{\mathbf{X}})$, the relation of $\mathbf{\Sigma}^{-1}E(\mathbf{X}|\mathbf{Y}) \in \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is equivalent to that of $\mathbf{\Sigma}^{-1/2}E(\mathbf{Z}|\mathbf{Y}) \in \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ according to Yoo (2016b). In practice, $E(\mathbf{Z}|\mathbf{Y})$ is restored instead of $E(\mathbf{X}|\mathbf{Y})$. Therefore, nonparametric estimation of $E(\mathbf{Z}|\mathbf{Y})$ is the primary interest in SIR. It can be done in a simple fashion by categorizing $\mathbf{Y}$ called slicing. Once the slicing is done, $E(\mathbf{X}|\mathbf{Y})$ can be easily replaced with sample means of $\mathbf{X}$ within each category. The estimation method of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ via $\mathbf{\Sigma}^{-1/2}E(\mathbf{Z}|\mathbf{Y})$ is called sliced inverse regression. Its sample algorithm is as follows.

Step 1. Slice $\mathbf{Y}$ to have h categories. Let H_j stand for the jth slice for $j = 1, 2, \ldots h$.

Step 2. Standardize the predictors $\mathbf{X}$ such that $\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$, $i = 1, 2, \ldots, n$, where $\hat{\mathbf{\Sigma}}$ is usual moment estimator of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}^{-1/2}\hat{\mathbf{\Sigma}}^{-1/2} = \hat{\mathbf{\Sigma}}^{-1}$.

Step 3. Calculate the sample means of $\bar{\hat{\mathbf{Z}}}_k = (1/n_k)\sum_{i\in H_k}\hat{\mathbf{Z}}_i$ within each slice for $k = 1,\ldots,h$, where n_k stands for the size of the kth slice. Then, form a kernel matrix $\hat{\mathbf{K}}_{\mathrm{SIR}}$:

$$\hat{\mathbf{K}}_{\mathrm{SIR}} = \left(\frac{n_1}{n}\bar{\hat{\mathbf{Z}}}_1, \frac{n_2}{n}\bar{\hat{\mathbf{Z}}}_2, \ldots, \frac{n_h}{n}\bar{\hat{\mathbf{Z}}}_h\right).$$

Step 4. Do the spectral decomposition of $\hat{\mathbf{M}}_{\mathrm{SIR}} = \hat{\mathbf{K}}_{\mathrm{SIR}}\hat{\mathbf{K}}_{\mathrm{SIR}}^{\mathrm{T}}$ such that $\hat{\mathbf{M}}_{\mathrm{SIR}} = \sum_{i=1}^{p}\hat{\lambda}_i\hat{\gamma}_i\hat{\gamma}_i^{\mathrm{T}}$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$.

Step 5. Let $\hat{\Gamma}_d = (\hat{\gamma}_1,\ldots,\hat{\gamma}_d)$ be the eigenvectors corresponding to the first d largest eigenvalues of $\hat{\mathbf{M}}_{\mathrm{SIR}}$. Defining that $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\Gamma}_d$, $\mathcal{S}(\hat{\boldsymbol{\eta}})$ is the estimate of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

2.2. Hierarchical inverse regression

In the SIR algorithm, if following the slicing scheme for multivariate responses, it often faces the curse of dimensionality. For example, if there are five dimensional responses, the least number of slices should be 32(=2^5). If the number of observations in data is 50, some slices must have only one observation. Accordingly, this leads unreliable dimension reduction results. It is noted that grouping the observations based on their similarity of the response is essential in the slicing scheme. When $\mathbf{Y}$ is multi-dimensional, grouping by similarity can be done via clustering algorithms. Setodji and Cook (2004) and Yoo *et al.* (2010) successfully replace the usual slicing scheme with the K-means clustering algorithm for SIR, called K-means inverse regression (KIR), and sliced average variance estimation, respectively. In a perspective of fusing, the K-means algorithm is not be effective according to Yoo *et al.* (2020). The benefit of fusing mainly comes from nestness and reproducibility of slicing, but the K-means algorithm does not have the two properties. For details on nestness and reproducibility, readers refer Yoo *et al.* (2020).

Instead, hierarchical clustering algorithms have nestness and reproducibility, and Yoo *et al.* (2020) showed that the application of SIR via hierarchical clustering algorithm have advantage over the K-means clustering algorithm. So, following the guidance of Yoo *et al.* (2020), Ward's hierarchical clustering algorithm will replace the usual slicing scheme.

For multivariate regression, once the responses are clustered, it replaces Step 1 in the SIR algorithm and follows the other steps in the same fashion. We call this approach *hierarchical inverse regression* (HIR).

3. Pooled sliced inverse regression and fused multivariate inverse regression

3.1. Pooled sliced inverse regression

Although clustering methods are effective and efficient alternatives to the usual slicing scheme for multivariate responses, it is inevitable for some clusters to have small sample sizes.

To overcome this issue, the following relationship between the central subspaces of $\mathbf{Y}|\mathbf{X}$ and the coordinate regression of $Y_k|\mathbf{X}$ should be noted:

$$\oplus_{k=1}^{u}\mathcal{S}_{Y_k|\mathbf{X}} \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}, \tag{3.1}$$

where $\mathcal{S}_{Y_k|\mathbf{X}}$ is the central subspace of $Y_k|\mathbf{X}$ and $\oplus$ denotes the direct sum among subspaces ($\mathcal{S}_1 \oplus \mathcal{S}_2 = v_1 + v_2; v_1 \in \mathcal{S}_1, v_2 \in \mathcal{S}_2$).

This relation was firstly observed and utilized by Yoo *et al.* (2010), which proposed pooled sliced average variance estimation. It directly implies that combining all information on the central subspace of the coordinate regressions contains useful information on $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

Following this pooling idea, we newly introduce the following *pooled sliced inverse regression* (pSIR). Let $\mathbf{M}_{\text{SIR}}^{(k)}$ be the population kernel matrices of SIR for $Y_k|\mathbf{X}$. Define $\mathbf{M}_{\text{pSIR}} = (1/r)\sum_{k=1}^{r}\mathbf{M}_{\text{SIR}}^{(k)}$. The columns of the first d largest eigenvectors of $\mathbf{M}_{\text{pSIR}}$ pre-multiplied by $\mathbf{\Sigma}^{-1/2}$ span $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. Its sample algorithm is as follows.

Step 1. Construct $\hat{\mathbf{M}}_{\text{SIR}}^{(k)}$ for a coordinate regression of $Y_k|\mathbf{X}$, $k = 1, \ldots, u$, from the usual SIR application.

Step 2. Compute $\hat{\mathbf{M}}_{\text{pSIR}} = (1/r)\sum_{k=1}^{r}\hat{\mathbf{M}}_{\text{SIR}}^{(k)}$.

Step 3. Do the spectral decomposition of $\hat{\mathbf{M}}_{\text{pSIR}}$ such that $\hat{\mathbf{M}}_{\text{pSIR}} = \sum_{i=1}^{p}\hat{\lambda}_i\hat{\gamma}_i\hat{\gamma}_i^{\text{T}}$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$.

Step 4. Let $\hat{\Gamma}_d = (\hat{\gamma}_1, \ldots, \hat{\gamma}_d)$ be the eigenvectors corresponding to the first d largest eigenvalues of $\hat{\mathbf{M}}_{\text{pool}}$. Defining that $\hat{\boldsymbol{\eta}} = \hat{\mathbf{\Sigma}}^{-1/2}\hat{\Gamma}_d$, $\mathcal{S}(\hat{\boldsymbol{\eta}})$ is the estimate of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

3.2. Fused hierarchical inverse regression

Let $\mathbf{M}_{\text{HIR}}^{\{g\}}$ indicate the kernel matrix of HIR with g clusters constructed by Ward's hierarchical clustering algorithm. Then, the following relation is easily observed:

$$\mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}\left(\mathbf{M}_{\text{HIR}}^{\{g\}}\right) \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}, \quad g = 2, \ldots, h.$$

The case of $g = 1$ is obviously ruled out, because it yields null matrix. This above relation directly indicates that

$$\oplus_{g=2}^{h}\mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}\left(\mathbf{M}_{\text{HIR}}^{\{g\}}\right) = \mathbf{\Sigma}^{-\frac{1}{2}} \oplus_{g=2}^{h}\mathcal{S}\left(\mathbf{M}_{\text{HIR}}^{\{g\}}\right) \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}.$$

Based on this, we newly define $\mathbf{M}_{\text{FHIR}}^{\{g\}}$ as

$$\mathbf{M}_{\text{FHIR}}^{\{g\}} = \left(\mathbf{M}_{\text{HIR}}^{\{2\}}, \mathbf{M}_{\text{HIR}}^{\{3\}}, \ldots, \mathbf{M}_{\text{HIR}}^{\{g\}}\right), \quad g = 3, \ldots, h. \tag{3.2}$$

In (3.2), the case of $\mathbf{M}_{\text{FHIR}}^{\{2\}}$ is out of consideration, because $\mathbf{M}_{\text{FHIR}}^{\{2\}} = \mathbf{M}_{\text{HIR}}^{\{2\}}$. Theoretically, we can see that

$$\mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}\left(\mathbf{M}_{\text{FHIR}}^{\{3\}}\right) \subseteq \mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}\left(\mathbf{M}_{\text{FHIR}}^{\{4\}}\right) \subseteq \cdots \subseteq \mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}\left(\mathbf{M}_{\text{FHIR}}^{\{h\}}\right) \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}.$$

Therefore, $\mathbf{M}_{\text{FHIR}}^{\{g\}}$ becomes a new kernel matrix to estimate $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. Further, for the exhaustive estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, a condition that $\mathbf{\Sigma}^{-1/2}\mathcal{S}\left(\mathbf{M}_{\text{FHIR}}^{\{g\}}\right) = \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is forced, which is normally assumed in SDR literature. The estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ through $\mathbf{M}_{\text{FHIR}}^{\{g\}}$ will be called *fused hierarchical inverse regression* (FHIR).

The sample version $\hat{\mathbf{M}}_{\text{FHIR}}^{\{g\}}$ is computed by replacing the population quantities with usual sample HIR kernel matrices. Fusing all information of the HIR application upto g clusters would cause potential advantages in more robust estimation results to choices of h and more accurate estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ than KIR.

3.3. Fused pooled sliced inverse regression

Let $\mathbf{M}_{\text{pSIR}}^{\{g\}}$ indicate the kernel matrix constructed by pSIR with g slices for all coordinate regressions of $Y_k|\mathbf{X}$. Like FHIR, the following relation is easily observed:

$$\mathbf{\Sigma}^{-\frac{l}{2}}\mathcal{S}\left(\mathbf{M}_{\text{pSIR}}^{\{g\}}\right) \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}, \quad g = 2, \ldots, h.$$

Accordingly like (3.2), we define that

$$\mathbf{M}_{\text{FpSIR}}^{\{g\}} = \left(\mathbf{M}_{\text{pSIR}}^{\{2\}}, \mathbf{M}_{\text{pSIR}}^{\{3\}}, \ldots, \mathbf{M}_{\text{pSIR}}^{\{g\}}\right), \quad g = 3, \cdots, h, \tag{3.3}$$

and the following relation holds for the non-decreasing sequences of $\mathbf{M}_{\text{FpSIR}}^{\{g\}}$, $g = 3, \ldots, h$:

$$\mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}\left(\mathbf{M}_{\text{FpSIR}}^{\{3\}}\right) \subseteq \mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}\left(\mathbf{M}_{\text{FpSIR}}^{\{4\}}\right) \subseteq \cdots \subseteq \mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}\left(\mathbf{M}_{\text{FpSIR}}^{\{h\}}\right) \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}.$$

By assuming that $\mathbf{\Sigma}^{-1/2}\mathcal{S}\left(\mathbf{M}_{\text{FpSIR}}^{\{g\}}\right) = \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ for the exhaustive estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, the quantity $\mathbf{M}_{\text{FpSIR}}^{\{g\}}$ becomes another kernel matrix fully informative to $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ for multivariate regression. We call this SDR approach *fused pooled sliced inverse regression* (FpSIR).

The sample version $\hat{\mathbf{M}}_{\text{FpSIR}}^{\{g\}}$ is constructed by computing $\hat{\mathbf{M}}_{\text{pSIR}}^{\{g\}}$. Any choice of g in $\hat{\mathbf{M}}_{\text{pSIR}}^{\{g\}}$ will provide the same asymptotic results, but their non-asymptotic behaviors can be easily affected by the choice of g. However, by fusing all the pSIR application results upto g slices for all coordinate regressions, more robust and accurate estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is expected than KIR.

3.4. Remarks on FpSIR and FHIR

For multivariate regression, one can use KIR, FpSIR and FHIR. The method FpSIR is recommended as default among the three, because FpSIR provides quite good estimation performances in various numerical studies, which are given in the next section. The two methods of KIR and FHIR require clustering application, so it cannot be implemented for some data. Also, it is known that outliers often affect clustering results, which may induce undesirable clustering results. Then, KIR and FHIR possibly produce poor estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. So, one fit FpSIR first, and see the results. If the dimension reduction results are not satisfactory, then it should be compared with those of FHIR and KIR.

4. Numerical studies and data analysis

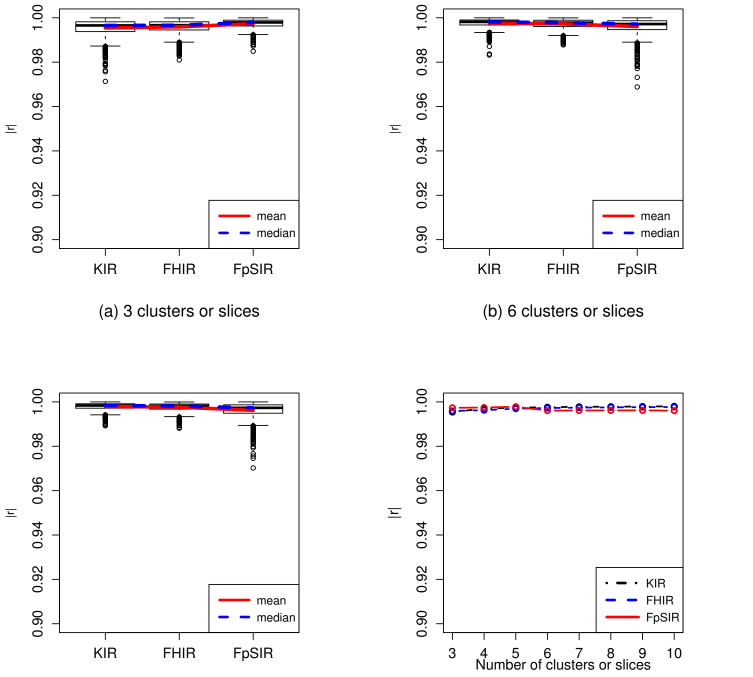
4.1. Numerical studies

For all numerical studies, the sample sizes were 100, and each simulation model was iterated 1,000 times. To measure how the three methods of KIR, FHIR and FpSIR estimate $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ well, absolute value $|r|$ of the square-root of r^2 from a regression of $\boldsymbol{\eta}_i^{\mathrm{T}}\mathbf{X}$ on $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$, $i = 1, \ldots, d$, was computed, where $\hat{\boldsymbol{\eta}}$ stands for the sample estimate of $\boldsymbol{\eta}$. Three to ten numbers of clusters or slices were considered for the three methods of KIR, FHIR and FpSIR.

The numerical studies are summarized by side-by-side boxplots of $|r|$ for 3, 6 and 9 clusters or slices (not all reported) along with a plot of lining mean of $|r|$s against the number of slices, $h = 3, 4, \ldots, 10$.

We considered the following two models, which were investigated in Setodji and Cook (2004) for KIR. In the models, all predictors X_i and random errors ε_i were independently generated from $N(0, 1)$.

- **Model 1** Each coordinate regression of $\mathbf{Y} = (Y_1, \ldots, Y_4)^{\mathrm{T}}|\mathbf{X} = (X_1, \ldots, X_4)^{\mathrm{T}}$ is as follows.



(a) 3 clusters or slices

(b) 6 clusters or slices

(c) 9 clusters or slices

(d) Comparison of means of KIR, FHIR and FpSIR

Figure 1: *Model 1 with homogeneous variance.*

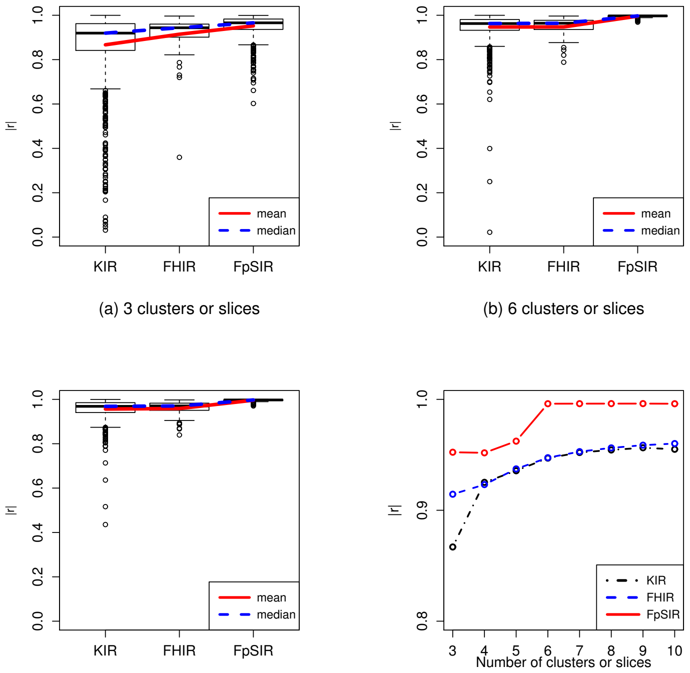
$$
\begin{aligned}
Y_1 &= c_1\mathbf{1}^{\mathrm{T}}\mathbf{X} + c_2\exp\left(c_3\mathbf{1}^{\mathrm{T}}\mathbf{X}\right)\varepsilon_1;\\
Y_2 &= c_1\mathbf{1}^{\mathrm{T}}\mathbf{X} + c_2\exp\left(c_3\left|2-3\mathbf{1}^{\mathrm{T}}\mathbf{X}\right|\right)\varepsilon_2;\\
Y_3 &= c_1\mathbf{1}^{\mathrm{T}}\mathbf{X} + c_2\exp\left(2c_3\mathbf{1}^{\mathrm{T}}\mathbf{X}\right)\varepsilon_3;\\
Y_4 &= c_1\mathbf{1}^{\mathrm{T}}\mathbf{X} + c_2\exp\left(c_3\left|1-\mathbf{1}^{\mathrm{T}}\mathbf{X}\right|\right)\varepsilon_4, \text{ where } \mathbf{1} \text{ is a vector all of which elements consist of 1.}
\end{aligned}
$$

- **Model 2** Each coordinate regression of $\mathbf{Y} = (Y_1, Y_2)^{\mathrm{T}}|\mathbf{X} = (X_1, \ldots, X_{10})^{\mathrm{T}}$ is as follows.

$$
\begin{aligned}
Y_1 &= X_1(X_1 + X_2 + 1) + \sigma\varepsilon_1,\\
Y_2 &= \frac{X_1}{0.5 + (X_2 + 1.5)^2} + \sigma\varepsilon_2.
\end{aligned}
$$

All coordinate regressions in Model 1 have the common linear conditional mean of $c_1\mathbf{1}^{\mathrm{T}}\mathbf{X}$. Since $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is spanned by the column vector $\mathbf{1}$, the structural dimension is equal to one. Depending on the choice of the value of c_3, the regression has heteroscedasticity. Two cases of (1, 1, 0) and (0.1, 1, 0.1) for (c_1, c_2, c_3) were considered. In the first case, the model is homoscedastic, while it is heteroscedastic for the second case. Through Model 1, it can be investigated how heteroscedasticity impacts the estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ for the three methods.

Model 2 was designed to compare the estimation performances of the three methods for nonlinear conditional means. Since the central subspace of Model 2 is spanned by the columns of $(1, 0, 0, \ldots, 0)^{\mathrm{T}}$ and $(0, 1, 0, \ldots, 0)^{\mathrm{T}}$, which correspond to X_1 and X_2, respectively, its structural dimension is equal to two. Further, the values of σ were set to 0.5 and 1.



(a) 3 clusters or slices

(b) 6 clusters or slices

(c) 9 clusters or slices

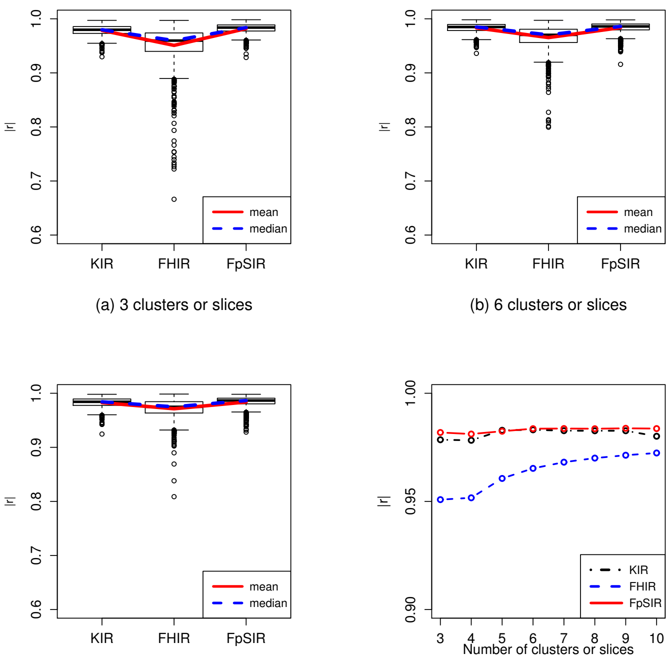
(d) Comparison of means of KIR, FHIR and FpSIR

Figure 2: *Model 1 with heterogeneous variance.*

Numerical studies for Models 1 and 2 are summarized in Figures 1–6. For Model 1 with homogeneous variance, there is no notable difference among all three methods of KIR, FHIR and FpSIR, which yield very reliable estimation results. This is partially because Model 1 is just linear regression. However, according to Figure 2, the proposed FpSIR shows the best and most robust estimation results to the numbers of clusters or slices among the three. Again, KIR and FHIR provide similar estimation performances, although KIR is the worst with 3 clusters. The existence of heteroscedasticy in Model 1 can bring outliers in responses, which affect the clustering results as discussed in Section 3.3. This possibly yield undesirable clustering results of the response variables, and it induces poor estimation results of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. Poor estimation of performances of KIR and FHIR in Model 1 can be partially explained by this aspect. With larger numbers of slices, FpSIR still provide good estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ under heteroscedasticity.

For Model 2, Figures 3 and 5 show that the first sufficient predictor of X_1 are well-estimated by all the three methods for $\sigma = 0.5$ and 1. However, for the second sufficient predictor X_2, KIR yields very sensitive results for small numbers of clusters, while HIR is also quite robust to the numbers of clusters and FpSIR is very robust to the numbers of slices. The estimation performances of the three methods is negatively impacted by larger variability of noise ε. So, it can be concluded that HIR and FpSIR estimate $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ relatively well.

This numerical studies confirm that two proposed fused methods, especially FpSIR, outperform the existing KIR in the estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, so we can expect potential advantages of FHIR and FpSIR over KIR for the dimension reduction of predictors in multivariate regression.



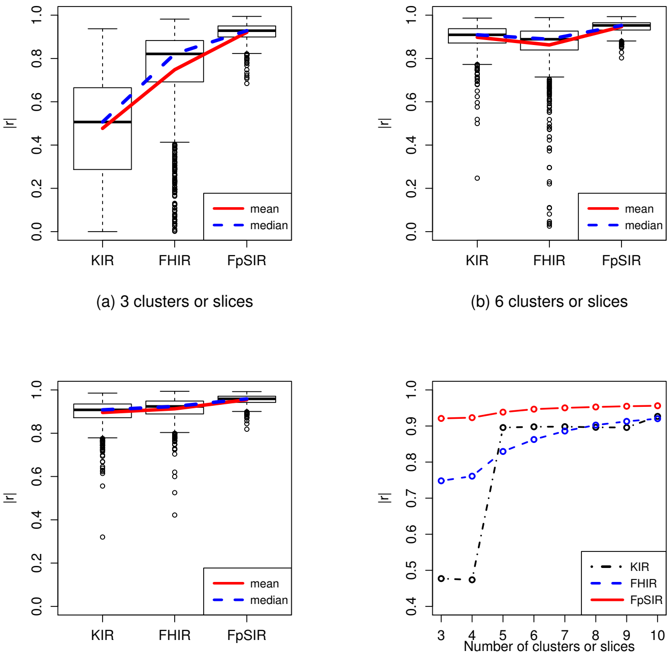
(a) 3 clusters or slices

(b) 6 clusters or slices

(c) 9 clusters or slices

(d) Comparison of means of KIR, FHIR and FpSIR

Figure 3: *Model 2 with $\sigma = 0.5$, the first sufficient predictor of X_1.*



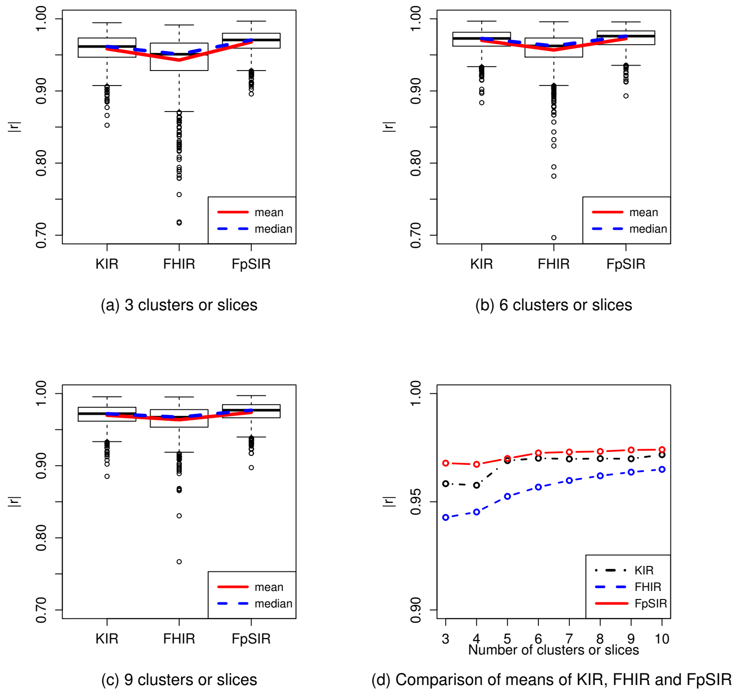
(a) 3 clusters or slices

(b) 6 clusters or slices

(c) 9 clusters or slices

(d) Comparison of means of KIR, FHIR and FpSIR

Figure 4: *Model 2 with $\sigma = 0.5$, the second sufficient predictor of X_2.*



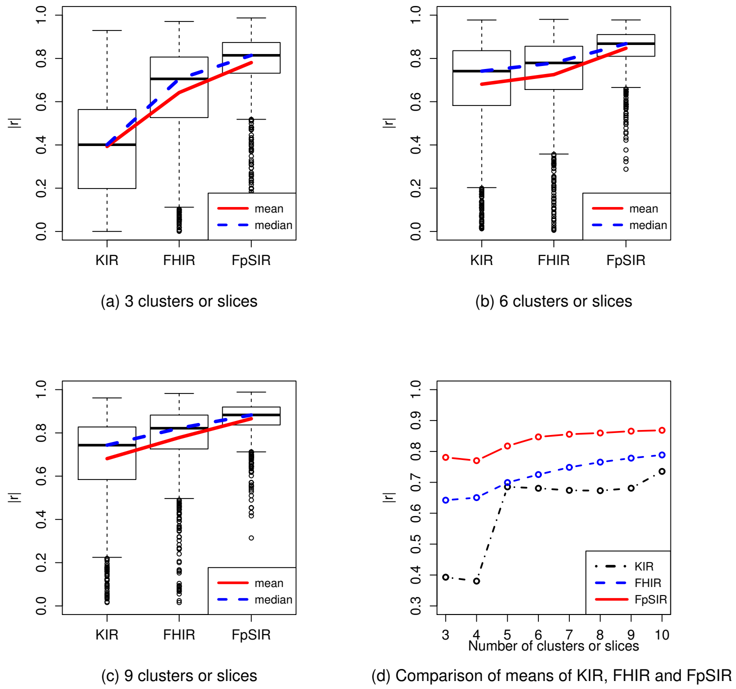
(a) 3 clusters or slices

(b) 6 clusters or slices

(c) 9 clusters or slices

(d) Comparison of means of KIR, FHIR and FpSIR

Figure 5: *Model2 with $\sigma = 1$, the first sufficient predictor of X_1.*



(a) 3 clusters or slices

(b) 6 clusters or slices

(c) 9 clusters or slices

(d) Comparison of means of KIR, FHIR and FpSIR

Figure 6: *Model 2 with $\sigma = 1$, the second sufficient predictor of X_2.*

Table 1: Dimension tests by the application of KIR, FHIR and FpSIR with 3, 6 and 9 clusters or slices: KIR#, KIR with # clusters; FHIR#, FHIR with # clusters; FpSIR#, FpSIR with # slices

	KIR3	KIR6	KIR9	FHIR3	FHIR6	FHIR9	FpSIR3	FpSIR6	FpSIR9
$\text{H}_0 : d = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\text{H}_0 : d = 1$	0.006	**0.154**	**0.078**	0.001	0.005	0.0012	**0.111**	**0.080**	**0.066**
$\text{H}_0 : d = 2$	NA	0.972	0.770	NA	**0.704**	**0.429**	0.189	0.673	0.741

4.2. Minneapolis school data

For the illustration purpose, we considered a multivariate regression analyzed in Yoo (2009). The data is regarding the performance of students in $n = 63$ Minneapolis schools. In the data, there are four dimensional responses $\mathbf{Y}$ of the percents of students in a school scoring above and below average on standardized fourth and sixth grade reading comprehension tests. Among many variables the following five ones were considered as predictors: the pupil teacher ratio, and the square roots of the percentage of children receiving Aid to Families with Dependent Children, the percentage of children not living with both biological parents, the percentage of adults in the school area who completed high school, the percentage of persons in the area below the federal poverty level. The predictors were transformed to satisfy the linearity condition. The square root-scale is necessary to induce the condition required in SIR.

For this regression, KIR, FHIR and FpSIR were applied with 3, 6 and 9 clusters or slices. The estimated first and second sufficient predictors are reported in Figures 7 and 8. As seen in Figure 7, all first sufficient predictors are very close to each other regardless of the numbers of clusters or slices and methods. However, there are some differences in the second sufficient predictors. According to Figure 8, the second sufficient predictors from FpSIR with 3, 6 and 9 slices are not highly correlated. This implies that the second one would be random rather than deterministic, and this induces that it is not informative to $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ according to Yoo (2018). So, the first sufficient predictor should be enough for the regression. On the other hand, in Figure 8, it is observed that the second sufficient predictors from KIR and FHIR with 3, 6 and 9 clusters are highly correlated to each other, so we expect that the structural dimension determination for KIR and FHIR should be, at least, two following the same rationale in Yoo (2018). To investigate this, a permutation dimension test (Yin and Bura, 2006; Yoo, 2016b) were conducted for FHIR and FpSIR, and weighted χ^2 test for KIR (Setoji and Cook, 2004) starting $\text{H}_0 : d = 0$ with nominal level 5%. If $\text{H}_0 : d = 0$ is not rejected, increment d by 1 and redo the test. Then, the structural dimension d is determined as the hypothesized value in H_0 that the first non-rejection occurs. The p-values for the test from KIR, FHIR and FpSIR with 3, 6 and 9 clusters or slices are summarized in Table 1. According to Table 1, as discussed, FHIR and FpSIR determine that $\hat{d} = 1$ and $\hat{d} = 2$, respectively. However, KIR determines that $\hat{d} > 1$ with 3 clusters and $\hat{d} = 1$ with 6 and 9 clusters. This is partially because of sensitiveness of KIR to the number of clusters. To decide $\hat{d} = 1$ or $\hat{d} = 2$, it is necessary for formal theoretical and numerical studies in the dimension estimation of FHIR and FpSIR, but this will be left for further research. Because FpSIR yields the best estimation results in most cases of numerical studies and Yoo (2009) suggests that $\hat{d} = 1$, we tentatively decide that $\hat{d} = 1$. So, one can investigate to find an adequate model for the multivariate regression with the first sufficient predictor, instead of the original five dimensional predictors.

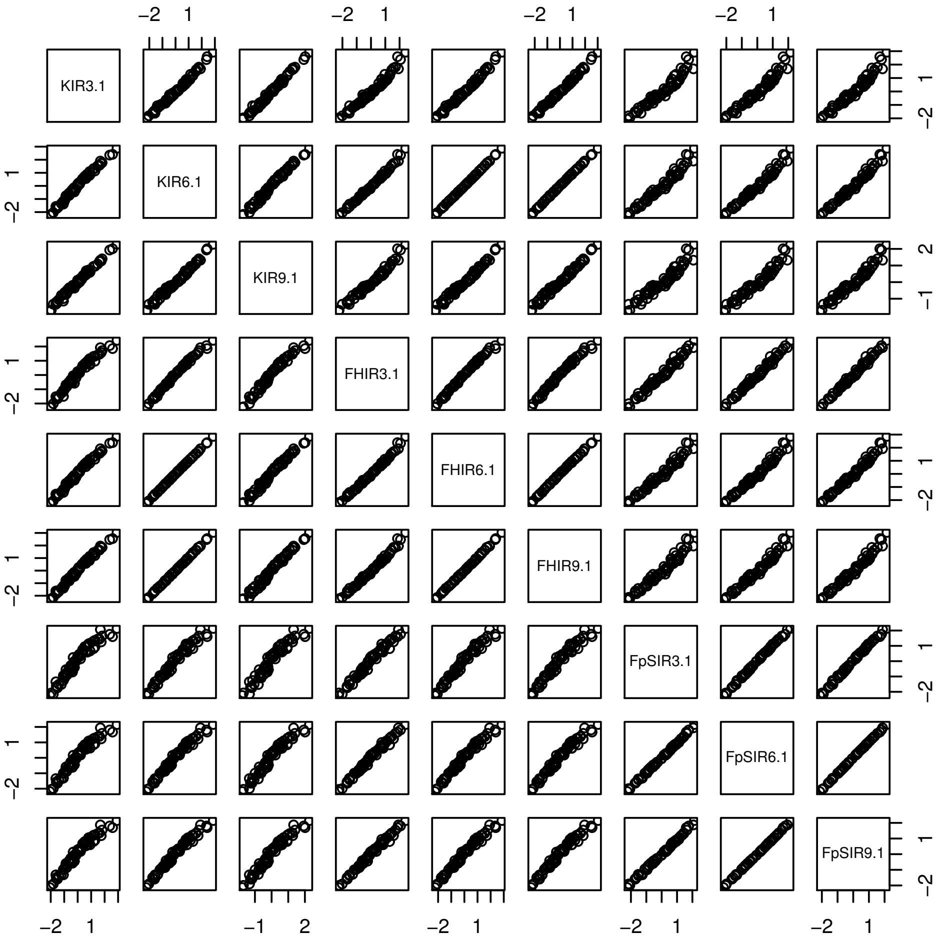


Figure 7: *Scatterplot matrix of the first sufficient predictors obtained from the application of KIR, FHIR and FpSIR with 3, 6 and 9 clusters or slices:* KIR#.1, *KIR with # clusters;* FHIR#.1, *FHIR with # clusters;* FpSIR#.1, *FpSIR with # slices.*

5. Discussion

For multivariate regression, there are few sufficient dimension reduction methods, although multi-dimensional responses become more popular nowadays in the so-called big data era. Existing inverse regression methods are still persuasive in multivariate regression, but they are sensitive to the number of clusters or slices. A fused approach recently developed by Cook and Zhang (2014) shows clear advantage for robustness to the number of slices in slicing-based inverse regression methods. So, in this paper, two fused inverse regression methods for multivariate regression are newly proposed, which are called *fused hierarchical inverse regression* and *fused pooled sliced inverse regression*. Fused hierarchical inverse regression accumulates all kernel matrices from hierarchical inverse regression (Yoo *et al.*, 2020) with various numbers of clusters. In fused hierarchical inverse regression, the multi-dimensional responses are clustered via Ward's hierarchical clustering algorithm. On the other

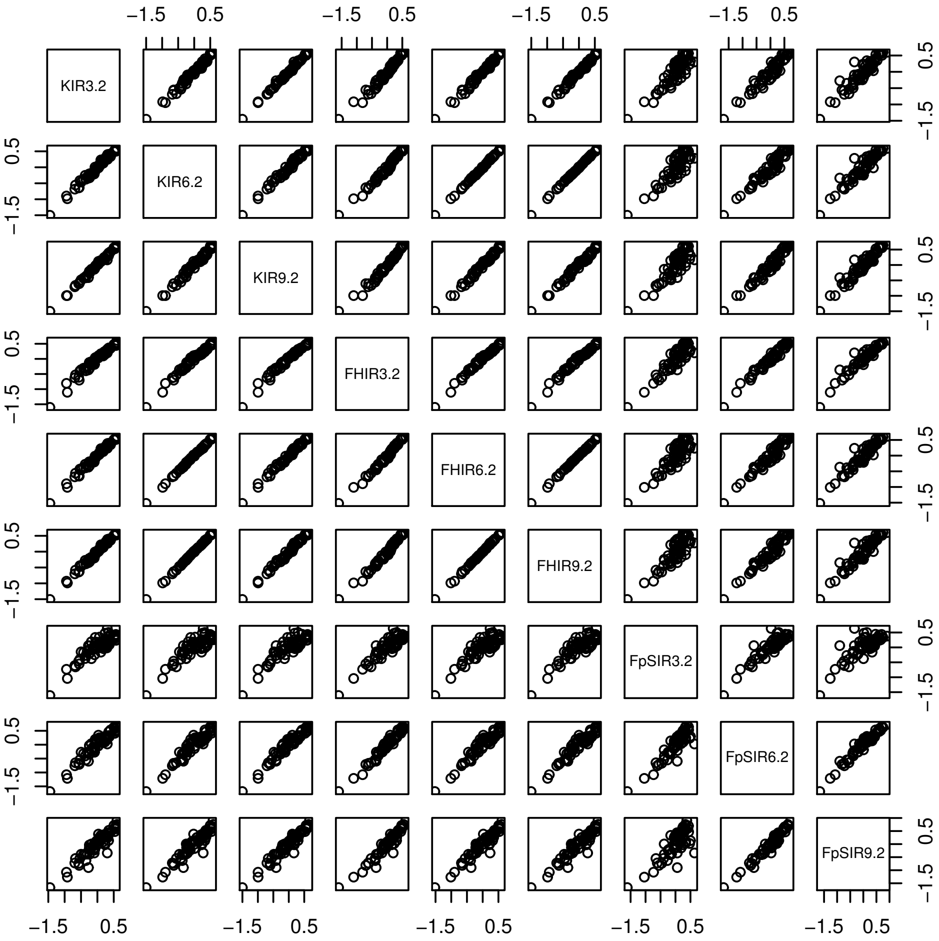


Figure 8: *Scatterplot matrix of the second sufficient predictors obtained from the application of KIR, FHIR and FpSIR with 3, 6 and 9 clusters or slices:* KIR#.2, *KIR with # clusters;* FHIR#.2, *FHIR with # clusters;* FpSIR#.2, *FpSIR with # slices.*

hand, the fused pooled sliced inverse regression has two step procedure. First, fused sliced inverse regression (Cook and Zhang, 2014) is implemented and fused kernel matrices are computed for each coordinate regression. Secondly, collect all kernel matrices from all coordinate regression, and the final fused kernel matrix is constructed. Different from fused hierarchical inverse regression, the clustering algorithm is not used, because the dimension of response in each coordinate regression is equal to one. So, in fused pooled sliced inverse regression, usual slicing scheme is applicable.

Numerical studies confirm that both proposed fused methods provide robustness to choice of clusters or slices and improve the estimation of the central subspace over the existing K-means inverse regression. A real data example shows their practical usefulness in multivariate regression analysis.

Theoretical asymptotics of sample kernel matrices for the dimension determination of fused hierarchical inverse regression and fused pooled sliced inverse regression should be studied and derived.

Since the two proposed fused methods have similar kernel matrices, each kernel matrix is not independent. So, for theoretical development, dependency central limit theorem should be applied. This direction of research is in progress.

6. Acknowledgements

For Jae Keun Yoo, this work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (NRF-2019R1F1A1050715).

References

Cook RD and Zhang X (2014). Fused estimators of the central subspace in sufficient dimension reduction, *Journal of the American Statistical Association*, **109**, 815-–827.

Lee K, Choi Y, Um H, and Yoo JK (2019). On fused dimension reduction in multivariate regression, *Chemometrics and Intelligent Laboratory Systems*, **193**, 103828.

Li KC (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316–327.

Setodji CM and Cook RD (2004). K-means inverse regression, *Technometrics* **46**, 421–429.

Yin X and Bura E (2006). Moment-based dimension reduction for multivariate response regression, *Journal of Statistical Planning and Inference*, **136**, 3675–3688.

Yoo C, Yoo Y, Um HY, and Yoo JK (2020). On hierarchical clustering in sufficient dimension reduction, *Communications for Statistical Applications and Methods*, **27**, 431–443.

Yoo JK (2008). A Novel moment-based dimension reduction approach in multivariate regression, *Computational Statistics and Data Analysis*, **52**, 3843–3851.

Yoo JK (2009). Iterative optimal sufficient dimension reduction for the conditional mean in multivariate regression, *Journal of Data Science*, **7**, 267–276.

Yoo JK (2016a). Tutorial: Dimension reduction in regression with a notion of sufficiency, *Communications for Statistical Applications and Methods*, **23**, 93–103.

Yoo JK (2016b). Tutorial: Methodologies for sufficient dimension reduction in regression, *Communications for Statistical Applications and Methods*, **23**, 95–117.

Yoo JK (2018). Basis-adaptive selection algorithm in dr-package, *The R Journal*, **10**, 124–132.

Yoo JK and Cook RD (2007). Optimal sufficient dimension reduction for the conditional mean in multivariate regression, *Biometrika*, **94**, 231–242.

Yoo JK, Lee K, and Woo S (2010). On the extension of sliced average variance estimation to multivariate regression, *Statistical Methods and Applications*, **19**, 529–540.

Received December 28, 2020; Revised February 3, 2021; Accepted February 5, 2021