More on directional regression

Kyongwon Kim[1,a], Jae Keun Yoo[2,a]

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

Abstract

Directional regression (DR; Li and Wang, 2007) is well-known as an exhaustive sufficient dimension reduction method, and performs well in complex regression models to have linear and nonlinear trends. However, the extension of DR is not well-done upto date, so we will extend DR to accommodate multivariate regression and large p-small n regression. We propose three versions of DR for multivariate regression and discuss how DR is applicable for the latter regression case. Numerical studies confirm that DR is robust to the number of clusters and the choice of hierarchical-clustering or pooled DR.

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

1. Introduction

As the so-called big data era is coming, high-dimensional data analysis is inevitable, and the meaningful reduction of data is essential to data analysis.

In regression of $\mathbf{Y} \in \mathbb{R}^r | \mathbf{X} \in \mathbb{R}^p$, sufficient dimension reduction (SDR) turns out to be effective, and it pursues to reduce the original p-dimensional predictors $\mathbf{X}$ to its lower-dimensional predictor $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ without loss of information on the conditional distribution of $\mathbf{Y} \in \mathbb{R}^r | \mathbf{X} \in \mathbb{R}^p$, where $r \geq 1$, $p \geq 2$ and $\boldsymbol{\eta} \in \mathbb{R}^{p \times d}$ with $d \leq p$. The following conditional independence is equivalent to it,

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

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

Such $\boldsymbol{\eta}$ to satisfy (1.1) can exist multiply, and then we have to choose the minimal one among them. The subspace spanned by the columns of the minimal $\boldsymbol{\eta}$ 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).

For many SDR methodologies, sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook and Weisberg, 1991) and directional regression (DR; Li and Wang, 2007)

[1] Co-Corresponding Author: Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: kimk@ewha.ac.kr
[2] Co-Corresponding Author: Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: peter.yoo@ewha.ac.kr

are popular among the others. The SIR considers the first inverse moments $E(\mathbf{X}|\mathbf{Y})$ and the SAVE does the second inverse moments $\text{cov}(\mathbf{X}|\mathbf{Y})$. So, SIR works well, when the regression has linear trend, while SAVE is better with non-linear trend. The DR, however, combines the first and second moments together, so it works best among the three with complex regression to have periodical and non-linear trends in regression.

Both SIR and SAVE has been extensively extended to cover many types of regression such as multivariate regression and large p-small n regression. So far, DR has not had much attention to this matter. The purpose of this paper is to extend the capability of the directional regression to these cases. For this, three versions of direction regression for multivariate regression will be proposed and we discuss how DR can be extended to the latter regression.

The organization of the paper is as follows. Direction regression is reviewed in Section 2. Section 3 is devoted to extending DR to multivariate regression, and its extension to large p-small n regression is discussed in Section 4. Numerical studies are presented in Section 5. We summarize our work in Section 6.

2. Directional regression

DR (Li and Wang, 2007) generalizes the ideas of inverse regression by considering conditional moments of $\mathbf{X} - \tilde{\mathbf{X}}$, instead of the conditional moment of $\mathbf{X}$, where $\tilde{\mathbf{X}}$ denotes an independent copy of $\mathbf{X}$. Here, without loss of generality, we use a standardized version of predictor variables and denote this as $\mathbf{Z}$. The fundamental idea of DR is to estimate a central subspace based on the empirical direction which can be written as $\{\mathbf{Z}_i - \mathbf{Z}_j | i \neq j, \ i, j = 1, \ldots, n\}$. Here, each empirical direction, $\mathbf{Z}_i - \mathbf{Z}_j$, contributes to describe the patterns of predictor variables and presents the distribution of the direction of $\mathbf{Z}$. Instead of using $\{\mathbf{Z}_i - \mathbf{Z}_j | i \neq j, \ i, j = 1, \ldots, n\}$ directly, DR uses rank 1 matrix

$$\left\{ \left(\mathbf{Z}_i - \mathbf{Z}_j\right)\left(\mathbf{Z}_i - \mathbf{Z}_j\right)^{\mathrm{T}} \Big| i \neq j, \ i, j = 1, \ldots, n \right\}, \tag{2.1}$$

to represent each empirical direction. Contour regression (CR; Li *et al.* , 2005) utilizes the empirical direction in (2.1) by regressing it within a tiny contour of $\mathbf{Y}$, that is, $\{|\mathbf{Y}_i - \mathbf{Y}_j| < \epsilon | i \neq j, \ i, j = 1, \ldots, n\}$. Then, the subspaces from this regression form an orthogonal complement to the central subspace. However, this method includes massive computation as we need to calculate $\binom{n}{2}$ empirical contour directions. DR adopt a more computationally efficient way by using

$$E\left[\left(\mathbf{Z} - \tilde{\mathbf{Z}}\right)\left(\mathbf{Z} - \tilde{\mathbf{Z}}\right)^{\mathrm{T}} \Big| \mathbf{Y}, \tilde{\mathbf{Y}}\right], \tag{2.2}$$

which can be explained as regressing $\mathbf{Z} - \tilde{\mathbf{Z}}$ directly to the paired response $(\mathbf{Y}, \tilde{\mathbf{Y}})$, where $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{Y}}$ are independent copies of $\mathbf{Y}$. Li and Wang (2007) present the orthogonal complement of a column space of (2.2) is a subspace of a central subspace under mild assumptions. DR can be formulated as a generalized eigendecomposition problems by using

$$\mathbf{M}_{\mathrm{DR}} = \left(2I_p - E\left[\left(\mathbf{Z} - \tilde{\mathbf{Z}}\right)\left(\mathbf{Z} - \tilde{\mathbf{Z}}\right)^{\mathrm{T}} \Big| \mathbf{Y}, \tilde{\mathbf{Y}}\right]\right)^2, \tag{2.3}$$

as a candidate matrix. Here, similar to CR, the candidate matrix of DR involves $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$ and the sample version of $\mathbf{M}_{\mathrm{DR}}$ can be represented as U or V statistic as,

$$
\begin{cases}
\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \left(2I_p - E\left[\left(\mathbf{Z} - \tilde{\mathbf{Z}} \right) \left(\mathbf{Z} - \tilde{\mathbf{Z}} \right)^{\mathrm{T}} \middle| \mathbf{Y}_i, \mathbf{Y}_j \right] \right)^2 & \text{(V statistic)}, \\
\frac{1}{\binom{n}{2}} \sum_{i \neq j}^{n} \left(2I_p - E\left[\left(\mathbf{Z} - \tilde{\mathbf{Z}} \right) \left(\mathbf{Z} - \tilde{\mathbf{Z}} \right)^{\mathrm{T}} \middle| \mathbf{Y}_i, \mathbf{Y}_j \right] \right)^2 & \text{(U statistic)},
\end{cases}
$$

where both of them require $O(n^2)$ operation. However, the candidate matrix in (2.3) can be reformulated without $(\tilde{\mathbf{Z}}, \tilde{\mathbf{Y}})$ and this is the explicit advantage of DR over CR. The kernel matrix of DR can be rephrased as

$$
\begin{aligned}
\mathbf{M}_{\mathrm{DR}} = {}& 2E\left[E\left[\mathbf{Z}\mathbf{Z}^{\mathrm{T}} \middle| \mathbf{Y} \right] \right]^2 + 2E\left[E\left[\mathbf{Z} \middle| \mathbf{Y} \right] E\left[\mathbf{Z}^{\mathrm{T}} \middle| \mathbf{Y} \right] \right]^2 \\
& + 2E\left[E\left[\mathbf{Z}^{\mathrm{T}} \middle| \mathbf{Y} \right] E\left[\mathbf{Z} \middle| \mathbf{Y} \right] E\left[\mathbf{Z} \middle| \mathbf{Y} \right] E\left[\mathbf{Z}^{\mathrm{T}} \middle| \mathbf{Y} \right] \right] - 2I_p.
\end{aligned}
\tag{2.4}
$$

This alternative version of the candidate matrix implies DR is a combination of SIR and SAVE as $\mathbf{M}_{\mathrm{DR}}$ hinges on $E[\mathbf{Z}|\mathbf{Y}]$ and $E[\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|\mathbf{Y}]$. Therefore, DR is a more comprehensive method than SIR and SAVE because (2.4) also indicates DR contains the subspaces of SIR and SAVE as its subspace. Furthermore, DR enjoys the same computational complexity as SIR and SAVE, $O(n)$. Finally, under mild assumptions and if there is a nonzero basis $\mathbf{B} \in \mathcal{S}_{\mathbf{Y}|\mathbf{Z}}$ with

$$
\mathbf{B}^{\mathrm{T}} \left(2I_p - E\left[\left(\mathbf{Z} - \tilde{\mathbf{Z}} \right) \left(\mathbf{Z} - \tilde{\mathbf{Z}} \right)^{\mathrm{T}} \middle| \mathbf{Y}, \tilde{\mathbf{Y}} \right] \right)^2 \mathbf{B} > 0,
$$

then, DR can exhaustively estimate a central subspace and this can be written as

$$
\mathrm{span}\left(\mathbf{M}_{\mathrm{DR}} \right) = \mathcal{S}_{\mathbf{Y}|\mathbf{Z}}.
$$

The following is the algorithm to calculate the sample version of the candidate matrix of DR.

Step 1. Standardize $(X_1, \ldots, X_n)$ to $\hat{\mathbf{Z}} = (\hat{Z}_1, \ldots, \hat{Z}_n)$.

Step 2. Let $H_1, \ldots, H_s$ be a partition of the range of $\mathbf{Y}$. Compute,

$$
\hat{\Lambda}_{1h} = E_n[\hat{\mathbf{Z}}I(\mathbf{Y} \in H_h)] / E_n[I(\mathbf{Y} \in H_h)], \, \hat{\Lambda}_{2h} = E_n[\hat{\mathbf{Z}}\hat{\mathbf{Z}}^{\mathrm{T}}I(\mathbf{Y} \in H_h)] / E_n[I(\mathbf{Y} \in H_h)].
$$

Step 3. Let $\hat{p}_h = E_n[I(\mathbf{Y} \in J_h)]$. Compute,

$$
\hat{\Gamma}_1 = \sum_{h=1}^{s} \hat{p}_h \hat{\Lambda}_{2h}^2, \quad \hat{\Gamma}_2 = \left(\sum_{h=1}^{s} \hat{p}_h \hat{\Lambda}_{1h} \hat{\Lambda}_{1h}^{\mathrm{T}} \right)^2, \quad \hat{\Gamma}_3 = \left(\sum_{h=1}^{s} \hat{\Lambda}_{1h}^{\mathrm{T}} \hat{\Lambda}_{1h} \right) \left(\sum_{h=1}^{s} \hat{\Lambda}_{1h} \hat{\Lambda}_{1h}^{\mathrm{T}} \right).
$$

Step 4. Finally, compute,

$$
\hat{\mathbf{M}}_{\mathrm{DR}} = 2\hat{\Gamma}_1 + 2\hat{\Gamma}_2 + 2\hat{\Gamma}_3 - 2I_p.
$$

3. Directional regression on multivariate regression

3.1. Hierarchical-slicing directional regression

Directional regression (DR) requires the slicing procedure for implementing it in practice. Therefore, even with multi-dimensional response regression, the theoretical foundation of DR is not changed. Rather, the slicing scheme must be changed in its implementation to afford the multivariate responses.

One simple solution to this is hierarchical-slicing. For example, consider four-dimensional responses $Y = (y_1, y_2, y_3, y_4)^T$. First, slice one of the four responses first into h_1 slices. Then, slice one of the other three responses within each slice. In this way, slice all response variables. It is simple, but there is cost. Even with doing 2 slices per responses, which is the minimal slice size, the total number of slices for the four dimensional responses is $2^4 = 16$. That is, the number of slices increase exponentially. This implies that this hierarchical-slicing is effective to lower-dimensional responses, for example bivariate ones. In this paper, this slicing scheme will be considered only for $r = 2$, and this will be called *hierarchical-slicing DR*.

3.2. Hierarchical-clustering directional regression

In sliced-based approach, it should be noted that the slicing is required for easy computation of the inverse moments. Therefore, the slicing is done under the philosophy that similar values are in slices.

A good option to follow this philosophy for multivariate responses is clustering. The first attempt to cluster the responses for inverse regression is Stedoji and Cook (2004), in which K-means clustering algorithm is adopted. Recently, Yoo *et al.* (2020) suggests that Ward-hierarchical clustering would be better than the former clustering method in many cases. So, by accepting the suggestion by Yoo *et al.* (2020), the responses are replaced with their clusters resulted from Ward-clustering algorithm, and then follow usual DR implementation. This has possible advantages over hierarchical slicing. First, the numbers of groups do not increase exponentially. Second, various numbers of slices can be considered, for example, $h = 2, 3, \ldots$, so small sample size per slice can be avoided. Third, all asymptotic results based on usual slicing scheme still hold. So, the clustering method is recommended for larger values of r than two. This approach will be called *hierarchical-clustering DR*.

3.3. Pooled directional regression

Although double-slicing or clustering methods are effective and efficient for multivariate responses, it is still problematic that some slices have small sample sizes.

One way to overcome this problem is to pool DR's application results from each coordinate regression. This approach initially proposed in Yoo *et al.* (2010) by noting the following relationship between the central subspaces of $\mathbf{Y}|\mathbf{X}$ and of the coordinate regression $Y_k|\mathbf{X}$, $k = 1, \ldots, r$

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

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

It directly indicates that pooling the central subspace of $Y_k|\mathbf{X}$ capture useful information on $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

One deficit of this pooling approach is that the exhaustiveness of DR for the central subspace of $Y_k|\mathbf{X}$ does not guarantee that of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. However, according to Yoo *et al.* (2010), it would not be problematic in practice.

Let $\mathbf{M}_{\text{DR}}^{(k)}$ be the population kernels of DR for $Y_k|\mathbf{X}$. Define $\mathbf{M}_{\text{pDR}} = 1/r \sum_{k=1}^{r} \mathbf{M}_{\text{DR}}^{(k)}$. Its sample algorithm is as follows.

Step 1. Compute $\hat{\mathbf{M}}_{\mathrm{DR}}^{(k)}$ for $Y_k|\mathbf{X}$, $k = 1, \ldots, r$, from the usual DR application.

Step 2. Construct $\hat{\mathbf{M}}_{\mathrm{pDR}} = 1/r \sum_{k=1}^{r} \hat{\mathbf{M}}_{\mathrm{DR}}^{(k)}$.

Step 3. Spectral-decompose $\hat{\mathbf{M}}_{\mathrm{pDR}}$ such that $\hat{\mathbf{M}}_{\mathrm{pDR}} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{\gamma}_i \hat{\gamma}_i^{\mathrm{T}}$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$.

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

This approach to estimate $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is called *pooled DR*.

4. Direction regression in large p–small n data

The so-called large p-small n regression is quite popular these days, and this is maybe why the dimension reduction is necessary. Since DR requires the inversion of the covariance of $\mathbf{X}$, the direct application of DR to such data is not plausible.

In such case, employing seeded dimension reduction approach to DR would be a possible and realistic route. All quantities in the population DR kernel matrices are computed as $\mathbf{X}$-scale, not $\mathbf{Z}$-scale, and consider it as a seed matrix. Then, the seed matrix is successively projected to Σ, which results a basis of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

However, this is not recommended to do, because the dimension of the seed matrix is too high. Actually, the seed matrix is a $p \times p$ matrix, so it would be not good for seed matrix. Also, the DR kernel matrix involves Σ, so the estimation accuracy is not expected to be good. One way to avoid this is to follow Um *et al.* (2018)'s approach. First, replace the dimension of $\mathbf{X}$ with its u-dimensional principal components. The choice of u must be carefully chosen, considering both the eigenvalues of Σ, so that u becomes relatively smaller than n. Then, with u-dimensional principal components, DR is applied in usual way for the additional reduction under the corporation with responses.

The extension of DR to large p-small n regression is not considered in numerical studies in the following section, because it is heavily affected by initial reduction of principal components, and then all the other steps remain the same as Li and Wang (2007).

Lastly, one can develop regularized DR to embracing L-1 or L-2 penalty, and this is definitely a new load and leave it as an open work.

5. Numerical studies and data analysis

5.1. Numerical studies

For all numerical studies, the sample sizes were 100 with $p = 6$ and 200 with $p = 20$, and each simulation model was iterated 1000 times. We consider hierarchical-slicing, hierarchical-clustering and pooled DR along with the compatible versions of SIR (Setodji and Cook, 2004) and SAVE (Yoo *et al.*, 2010). All dimension reduction methods under consideration is sliced-based, and the same numbers of slices were considered for 4 and 6 slices for $p = 6$ and 6 and 10 slices for $p = 20$.

For all simulation models, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is spanned by the two columns of $\boldsymbol{\eta}_1 = (1, 1, 1, 0, 0, \ldots, 0)^{\mathrm{T}}$ and $\boldsymbol{\eta}_2 = (1, 0, 0, 0, 1, 3, 0, \ldots, 0)^{\mathrm{T}}$. So, the true structural dimension is equal to two.

To measure how the dimension reduction methods under consideration for multivariate regression

a. M2-1: $Y = (y_1, y_2)$ b. M2-2: $Y = (y_1, y_4)$

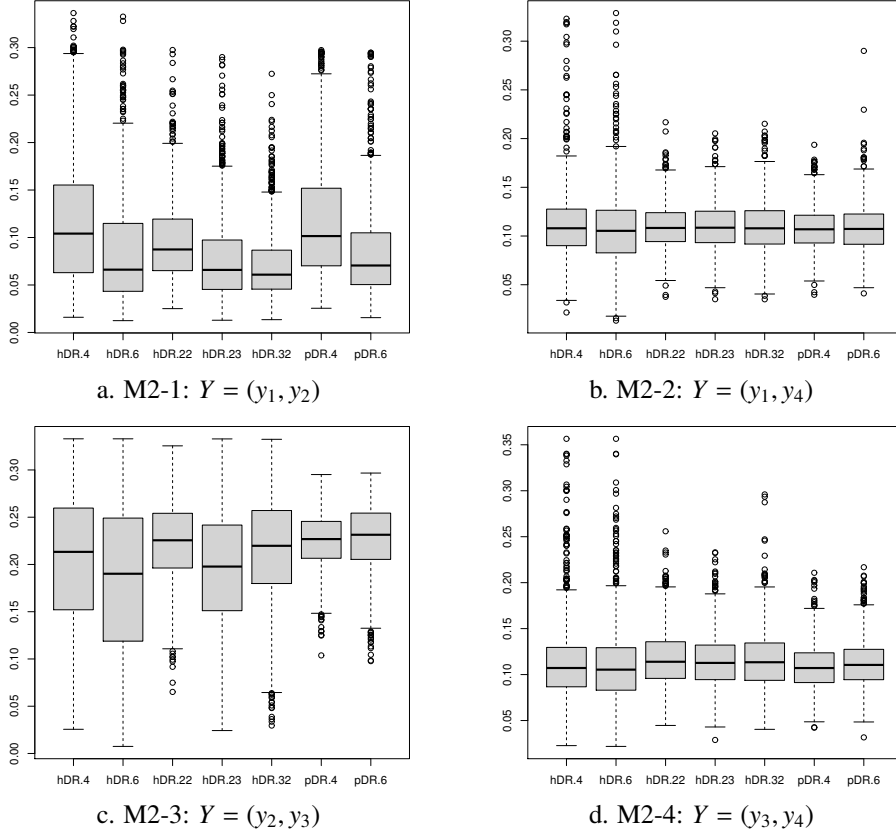c. M2-3: $Y = (y_2, y_3)$ d. M2-4: $Y = (y_3, y_4)$

Figure 1: *two-dimensional response regression models; hDR.#, hierarchical-clustering DR with # clusters; hDR.##, hierarchical-slicing DR with (#, #) slices; pDR.#, pooled DR with # slices.*
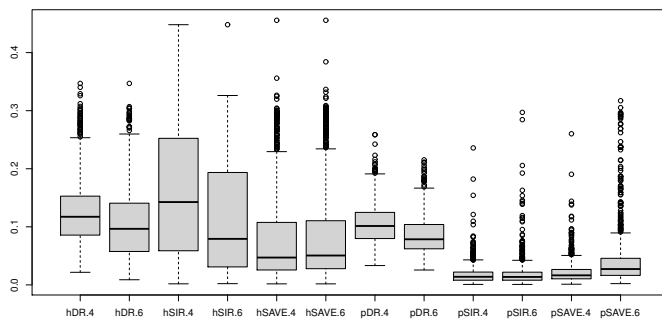
estimate $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ well, we computed the trace correlation distance r_D between $\boldsymbol{\eta}$ on $\hat{\boldsymbol{\eta}}_i$, $i = 1, \ldots, 1000$,

$$r_D = 1 - \sqrt{\frac{1}{2} \sum_{j=1}^{2} \psi_j},$$

where ψ is the eigenvalues of $\boldsymbol{\eta}(\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^{\mathrm{T}}\hat{\boldsymbol{\eta}}(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\hat{\boldsymbol{\eta}})^{-1}\hat{\boldsymbol{\eta}}^{\mathrm{T}}$. Actually, $\sqrt{\frac{1}{2} \sum_{j=1}^{2} \psi_j}$ is called *trace correlation coefficient* (Hooper, 1959), but it is subtracted from one to make smaller values of r_D indicate better estimation.

All predictors X_i and random errors ε_i were independently generated from $N(0, 1)$. And, we consider the following six coordinate regressions:

- $y_1 = 0.4 \left(\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X}\right)^2 + 3 \sin\left(\frac{\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X}}{4}\right) + 0.2\varepsilon_1,$

- $y_2 = 3 \sin\left(\frac{\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X}}{4}\right) + 3 \sin\left(\frac{\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X}}{4}\right) + 0.2\varepsilon_2,$

- $y_3 = 0.4 \left(\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X}\right)^2 + \mathrm{abs}\left(\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X}\right)^{\frac{1}{2}} + 0.2\varepsilon_3,$

a. $p = 6$
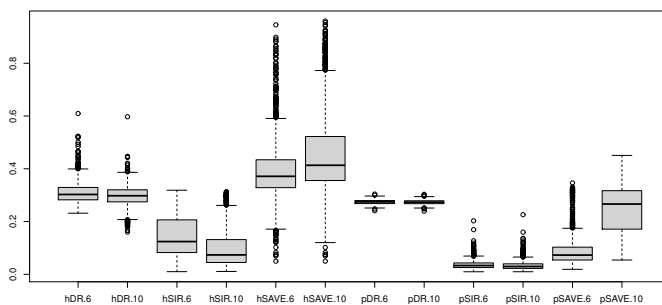
b. $p = 20$

Figure 2: *M4-1:* $Y = (y_1, y_2, y_3, y_4)^{\mathrm{T}}$.
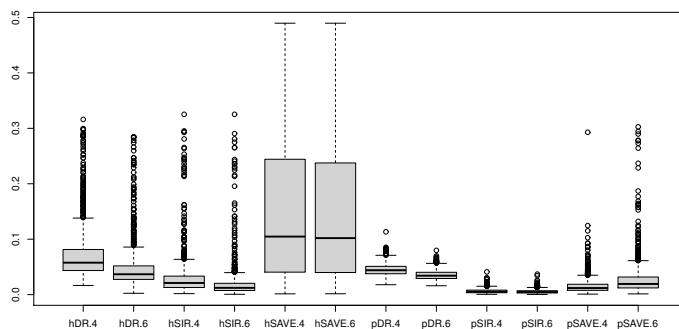
a. $p = 6$
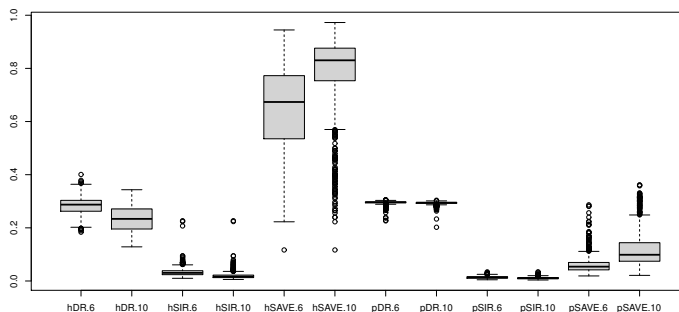
b. $p = 20$

Figure 3: *M4-2:* $Y = (y_1, y_2, y_5, y_6)^{\mathrm{T}}$.

a. $p = 6$

b. $p = 20$
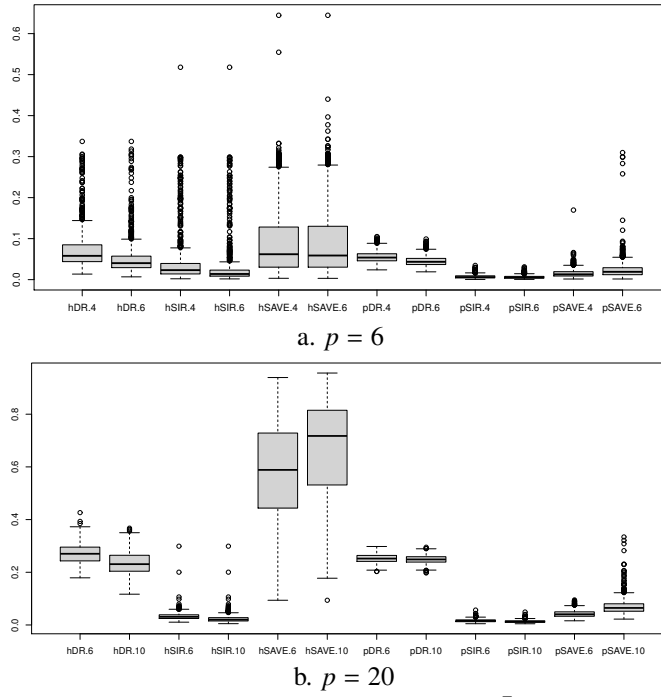
Figure 4: *M4-3:* $Y = (y_3, y_4, y_5, y_6)^{\mathrm{T}}$.

- $y_4 = 3 \sin\left(\frac{\eta_2^{\mathrm{T}} \mathbf{X}}{4}\right) + 0.2\left(1 + \left(\eta_1^{\mathrm{T}} \mathbf{X}\right)^2\right)\varepsilon_4,$

- $y_5 = \eta_1^{\mathrm{T}} \mathbf{X} + \left(\frac{\eta_2^{\mathrm{T}} \mathbf{X}}{4}\right)^2 + 0.2\varepsilon_5,$

- $y_6 = \eta_1^{\mathrm{T}} \mathbf{X}\left(1 + \eta_2^{\mathrm{T}} \mathbf{X}\right) + 0.2\varepsilon_6.$

To compare the estimation performances among hierarchical-slicing, hierarchical-clustering and pooled DR, the following four bivariate regressions were constructed: M2-1: $Y = (y_1, y_2)^{\mathrm{T}}$; M2-2: $Y = (y_1, y_4)^{\mathrm{T}}$; M2-3: $Y = (y_2, y_3)^{\mathrm{T}}$; M2-4: $Y = (y_3, y_4)^{\mathrm{T}}$.

Next, to investigate how well the hierarchical-slicing, hierarchical-clustering and pooled DR and their compatible versions of SIR and SAVE estimate $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, the following four dimensional-response regressions were considered : M4-1: $Y = (y_1, y_2, y_3, y_4)^{\mathrm{T}}$; M4-2: $Y = (y_1, y_2, y_5, y_6)^{\mathrm{T}}$; M4-3: $Y = (y_3, y_4, y_5, y_6)^{\mathrm{T}}$.

Numerical studies for M2-1–M2-4 and M4-1–M4-3 are summarized by boxplots of r_Ds in Figures 1–4.

According to Figure 1, three versions of DR for multivariate regression show quite similar performances, although bivariate slicing would be recommended.

In case of four-dimensional regression, the following common aspects are observed. The DR and SIR are robust to hierarchical clustering or pooled approach. The SAVE is very sensitive to the choice of the approach, and the pooled approach is much better than hierarchical-clustering one. Also, with larger p, SAVE worsens dramatically. As expected, SAVE is sensitive to the number of clusters, and so is SIR, especially in M4-1. The multivariate SIR shows the best estimation results for most simulation models. From various numerical studies, it is confirmed that multivariate DR does not outperform the

other SDR methods, but they still produce robust estimation results to the number of clusters and the dimension.

6. Discussion

Sufficient dimension reduction methodologies for multivariate regression and large p-small n regression have not been developed in many ways, even though high-dimensional response regression is getting popular in big data era. An exhaustive estimation method of directional regression (DR; Li and Wang, 2007) is attempted to fill this gap. According to Li and Wang (2007), DR turns out to be an effective dimension reduction tool, when the regression has complex structure to have linear and non-linear trends.

In this paper, we discuss how DR is extended to embracing multivariate regression and large p-small n regression. For multivariate regression, hierarchical clustering replaces usual slicing scheme used in univariate regression and combine all information resulted by DR from each coordinate regression.

Extension of DR to large p-small n regression would be good, as long as one adopts seeded dimension reduction approach (Cook *et al.*, 2007). However, since DR has to compute the covariance of **X** within the slice, the seed matrix constructed from DR would be not useful in practice. So, following Um *et al.* (2018), it is suggested that the predictors are initially reduced by principal component analysis, and then the principal components are re-reduced through DR.

Numerical studies show that multivariate versions of DR do not outperform the same versions of SIR and SAVE. This would be disappointing, but it is confirmed that DR is robust to the number of clusters and the choice of hierarchical-clustering or pooled DR.

The regularized version of DR is left as an open work.

References

Cook RD and Weisberg S (1991). Comment: Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 328--332.

Cook RD, Li B, and Chiaromonte F (2007). Dimension reduction in regression without matrix inversion, *Biometrika*, **94**, 569--584.

Hooper J (1959). Simultaneous equations and canonical correlation theory, *Econometrika*, **27**, 245—256.

Li B and Wang S (2007). On directional regression for dimension reduction, *Journal of the American Statistical Association*, **102**, 997–1008.

Li B, Zha H, and Chiaromonte F (2005). Contour regression: a general approach to dimension reduction, *The Annals of Statistics*, **33**, 1580–1616.

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.

Um HY, Won S, An H, and Yoo JK (2018). Case study: application of fused sliced average variance estimation to near-infrared spectroscopy of biscuit dough data, *The Korean Journal of Applied Statistics*, **31**, 835–842.

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