Communications for Statistical Applications and Methods
2020, Vol. 27, No. 4, 431–443

https://doi.org/10.29220/CSAM.2020.27.4.431
Print ISSN 2287-7843 / Online ISSN 2383-4757

On hierarchical clustering in sufficient dimension reduction

Chaeyeon Yoo[a], Younju Yoo[a], Hye Yeon Um[a], Jae Keun Yoo[1,a]

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

Abstract

The K-means clustering algorithm has had successful application in sufficient dimension reduction. Unfortunately, the algorithm does have reproducibility and nestness, which will be discussed in this paper. These are clear deficits for the K-means clustering algorithm; however, the hierarchical clustering algorithm has both reproducibility and nestness, but intensive comparison between K-means and hierarchical clustering algorithm has not yet been done in a sufficient dimension reduction context. In this paper, we rigorously study the two clustering algorithms for two popular sufficient dimension reduction methodology of inverse mean and clustering mean methods throughout intensive numerical studies. Simulation studies and two real data examples confirm that the use of hierarchical clustering algorithm has a potential advantage over the K-means algorithm.

Keywords: central subspace, hierarchical clustering, informative predictor subspace, K-means clustering, multivariate slicing, sufficient dimension reduction

1. Introduction

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

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

where $\perp\!\!\!\perp$ stands for statistical independence. The minimal subspace spanned by the columns $\boldsymbol{\eta}$ satisfying (1.1) is called *central subspace* $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. Hereafter, $\boldsymbol{\eta}$ and d will represent an orthonormal basis matrix and the structural dimension of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. The d-dimensional predictor $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ is called sufficient predictors.

Naturally, the main stream of SDR is to estimate $\boldsymbol{\eta}$. For $r = 1$, two inverse regression methods of sliced inverse regression (SIR) (Li, 1991) and sliced average variance estimation (SAVE) (Cook and Weisberg, 1991) are often used. The two methods commonly require a condition called *linearity condition* such that $E(\mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X})$ is linear in $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$. In their methodological development and practical implementation, the categorization of the response variable is essential. Its categorization is called *slicing*. The response is sliced for each category or to have equal numbers of observations (although it is not strictly required). Readers are recommended to read Yoo (2016a, 2016b) for further insights about SDR and details on SIR and SAVE.

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

Published 31 July 2020 / journal homepage: http://csam.or.kr

The background theories in SIR and SAVE remain the same when the response is multi-dimensional, but there are practical changes in slicing responses. Then usual slicing scheme with multivariate responses is as follows. Let $\mathbf{Y} = (y_1, \ldots, y_r)^{\mathrm{T}}$. First, slice one of $(y_1, \ldots, y_r)$ into h_1 slices. For simplicity, we consider the first response variable as y_1. Next, pick another response, simply y_2. Then, slice y_2 into h_2 slices within each slice constructed by y_1 so to have total $h_1 \times h_2$ slices. Following this slicing scheme, responses are sliced one-at-a-time, and the resulted slicing has a hierarchical structure with total $\prod_{i=1}^{r} h_i$ slices. For example, if $r = 4$ and $h_i = 2$ for $i = 1, 2, 3, 4$, the total slices are 16, which is the minimum number of slices. This hierarchical slicing scheme is straightforward and can be easily implemented, but the number of slices exponentially increases according to the number of response variables. This brings a smaller number of observations per slice, which results in the poor estimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ through SIR and SAVE.

To avoid this issue, the K-means clustering algorithm (KCA) has been adopted successfully in replacing hierarchical slicing. Setodji and Cook (2004) and Yoo *et al.* (2010) use the KCA to categorize the multi-dimensional responses, and extended the direct applicability SIR and SAVE to multivariate regression. The clusters constructed by the KCA play the same role as slices.

For another reason, the KCA has been used in SDR when extracting further information on the marginal distribution of $\mathbf{X}$. As discussed earlier, SIR and SAVE commonly require the linearity condition. The condition is for the marginal distribution of $\mathbf{X}$, neither for the conditional distribution of $\mathbf{Y}|\mathbf{X}$ or $\mathbf{X}|\mathbf{Y}$ nor for the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$. To relieve the violation of the condition or to capture more information on the structure of $\mathbf{X}$, the use of KCA has been considered in Li *et al.* (2004), Yoo (2016c, 2018) and Lee *et al.* (2019). Within each cluster, the covariance between $\mathbf{Y}$ and $\mathbf{X}$ (Li, *et al.* 2004; Yoo, 2018; Lee *et al.* 2019) and the mean $\mathbf{X}$ (Yoo, 2016c) are computed to restore $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

The use of KCA successfully extends the applicability of SDR methodologies to various data. The KCA, however, has two shortcomings compared to the usual slicing. The first is reproducibility. Under the same number of slices, the usual slicing always yields the same categorization result, while the KCA does not. The other is nestness. For example, a response $\mathbf{Y} \in \mathbb{R}^1$ is sliced twice into three and six slices with letting their results be $H^y_{(3)}$ and $H^y_{(6)}$, respectively. Then, $H^y_{(6)}$ is nested in $H^y_{(3)}$, in sense that two in $H^y_{(6)}$ are perfectly matched with one of $H^y_{(3)}$. However, this does not happen for the KCA. These two properties of slicing are important, because the information of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ are extracted equally-balanced and the methods provide the same result whenever using the same number of slices.

In clustering, the hierarchical clustering algorithm (HCA) satisfies two properties; in addition, there is no absolute reason why the KCA alone is applicable in SDR. However, any intensive comparison between KCA and HCA has not yet been done in sufficient dimension reduction context up to date.

This paper investigates how effective the HCA is in sufficient dimension reduction when comparing KCA throughout intensive numerical studies. We will consider four popular options (Single, Complete, Average and Ward's method) in HCA used when measuring the distance between clusters to compare KCA. Based on the studies, we will provide a practical guideline about how well the HCA can compete with KCA and about which option would be better in HCA. This article does not provide any theoretical comparison between KCA and HCA in sufficient dimension reduction context.

The organization of the paper is as follows. The KCA and HCA are introduced and are discussed from the view of reproducibility and nestness in Section 2. Also, Section 2 provides introduction of two existing methods involving KCA and the humble propose of their hierarchical clustering versions. The following section is devoted to numerical studies. Two real data examples are presented in Section 4. We summarize and conclude the work in Section 5.

2. Use of clustering in sufficient dimension reduction

2.1. K-means clustering algorithm

2.1.1. Algorithm

The K-means clustering algorithm (KCA) is a technique that sets the number of clusters in advance and assigns each data to clusters to minimize a measure of dispersion within the cluster. This analysis divides the sample so that data does not overlap in the predetermined number of clusters. K-means clustering is a popular clustering method applied to various statistical learning that has certain advantages such as the efficient convergence to local optimum (Hastie *et al.*, 2008). The algorithm of this is as follows.

Algorithm of K-means clustering

Step 1. Start the initial cluster with user-selective k.

Step 2. At every step, each observation is reassigned to the nearest cluster.

Step 3. Recompute the center of the cluster where observations are missing and added, and repeat Step 2.

Step 4. Stop when there is no further movement of observations.

2.1.2. No reproducibility and no nestness

To show that the KCA does not have any of reproducibility and nestness in practice, we consider the following multivariate regression of $\mathbf{Y} = (y_1, y_2, y_3)^{\mathrm{T}}|\mathbf{X} = (x_1, \ldots, x_5)^{\mathrm{T}}$: $y_1 = x_1^2 + 0.5\varepsilon_1$; $y_2 = x_1 + 0.5\varepsilon_2$; $y_3 = \exp(x_1) + 0.5\varepsilon_3$. All variables of $(x_1, \ldots, x_5, \varepsilon_1, \ldots, \varepsilon_3)$ are randomly generated from $N(0, 1)$, and the sample size is 100. The regression depends on $\mathbf{X}$ only through x_1. Therefore, the one-dimensional column vector $(1, 0, 0, 0, 0)^{\mathrm{T}}$ spans $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

First, we clustered the responses to have 2 clusters three times. Then, the number of observations for each cluster resulted from the three trials were $(93, 7)$, $(90, 10)$, and $(90, 10)$, where (n_1, n_2) stands for the number of observations of the first and second clusters. It is observed that the first two results are not the same. This directly implies that reproducibility is not guaranteed in KCA. Next, the responses were clustered to have 3 clusters three times. The number of observations for each cluster were $(7, 27, 66)$, $(76, 10, 14)$, and $(19, 74, 7)$; consequently, we can see that the three results are different. This implies that, under the same number of clusters, the multivariate inverse regression methods by Setodji and Cook (2004) and Yoo *et al.* (2010) possibly provides different results. However, it is not clear which result should be used. To investigate no nestness of KCA, the same data was clustered with 4 clusters one time, and the cluster sizes were $(18, 7, 8, 67)$, It is not clearly nested in the results of $(90, 10)$, but is nested in $(93, 7)$.

Next, the predictors using the same data were clustered following the same way as the response variables. Then, with 2 clusters, the sizes were $(53, 47)$, $(58, 42)$ and $(56, 44)$; as well as $(38, 33, 29)$, $(21, 28, 51)$ and $(26, 40, 34)$ for 3 clusters. The cluster sizes are all different for each trial and there is no reproducibility. If the predictors are clustered to have 4 clusters, the resulted sizes are $(19, 26, 30, 25)$ and do not fall into any of 2 cluster results.

2.2. Hierarchical clustering algorithm

The hierarchical clustering algorithm (HCA) builds a decisive and flexible algorithm for clustering. The KCA requires a user-defined number of clusters to obtain clustering solutions, and it is difficult to define a good choice of the parameter. But the HCA does not need to specify the number of clusters. The HCA is as follows.

Step 1. Start with n clusters, where n stands for the number of observations.

Step 2. Merged the two nearest observations into one cluster.

Step 3. At every step, the two clusters closest to the distance are merged. This means that single observations are added to existing clusters or that two existing clusters are merged.

To measure the similarity or distance between two clusters in HCA, single linkage, complete linkage, average linkage, and Ward's method are widely used.

- **Single linkage**: The single linkage is a method of constructing a higher-level cluster by calculating the distance between observations belonging to each cluster in two clusters and merging two clusters with the closest of these values. This is also called the nearest neighbor method. Suppose that there two clusters of $U = (x_1, \ldots, u_{n_u})$ and $V = (y_1, \ldots, y_{n_v})$. Then, the distance between U and W is as follows.

$$d_{U,V} = \min\left\{x_i \in U, y_j \in V : d(x_i, y_j)\right\},$$

  where $d(x, y)$ stands for a measure of distance or similarity between x and y. Here, the usual Euclidean distance has been used.

- **Complete linkage**: In the compete linkage algorithm, the similarity between two clusters is defined as the farthest distance between any two observations in a different cluster, unlike single linkage.

$$d_{U,V} = \max\{x_i \in U, y_j \in V : d(x_i, y_j)\}.$$

- **Average linkage**: The average linkage algorithm is the method that uses the average distance between all pairs of different clusters for the distance between observations in each cluster.

$$d_{U,V} = \frac{1}{n_u n_v} \sum_{i=1}^{n_u} \sum_{j=n_v}^{b} d(x_i, y_j).$$

- **Ward's method**: In the case of single linkage, complete linkage and average linkage all, a cluster was formed using Euclidean square distances. However, Ward's method measures the similarity between two clusters based on the increase in the sum of squares error (ESS) when two clusters are combined. The reason for using ESS is to consider information loss. When the observations are grouped together, information about individual observation is replaced by information about the clusters in which they belong. Ward's method considers this information loss for combining two clusters. Suppose there are k clusters at a current stage, and let U_i stands for the i^{th} cluster for $i = 1, \ldots, k$. The sum of square error for U_i is defined: $\text{ESS}_i = \sum_{j=1}^{n_i}(x_j^{(i)} - \bar{x}^{(i)})(x_j^{(i)} - \bar{x}^{(i)})^{\text{T}}$, where $x_j^{(i)}$ stands for the j^{th} observation in the i^{th} cluster and $\bar{x}^{(i)}$ represents the sample mean for the i^{th} cluster. Then the total ESS is defined as: $\text{ESS} = \sum_{i=1}^{k} \text{ESS}_i$. With the k clusters existing at the current stage, combine two clusters for all possible cases. Then, the clustering results with the smallest increase in ESS is the next stage, in which there are $k - 1$ clusters.

The result of the HCA is usually reported as dendrogram. If one fixes a measure of distance between two clusters and do hierarchical clustering, it always produces the same results, because no randomness is involved in the clustering procedure. The nestness property is guaranteed since the HCA has a hierarchical structure to combine one cluster at a step that starts from n clusters.

2.3. Two sufficient dimension reduction methods involving clustering

2.3.1. Inverse regression method

According to Li (1991), $\mathbf{\Sigma}^{-1}E(\mathbf{X}|\mathbf{Y}) \in \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ under the linearity condition that $E(\mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X})$ is linear in $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$, where $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{X})$. So, a subspace spanned by $\mathbf{\Sigma}^{-1}E(\mathbf{X}|\mathbf{Y})$ varying the values of $\mathbf{Y}$ is contained in $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

It is essential to restore $E(\mathbf{X}|\mathbf{Y})$ without knowing any parametric condition about $\mathbf{X}|\mathbf{Y}$, $\mathbf{Y}|\mathbf{X}$ or $\mathbf{X}$ and $\mathbf{Y}$. One simple possible route to solve this is a partitioning by $\mathbf{Y}$. Then, $E(\mathbf{X}|\mathbf{Y})$ is nothing but the group mean of $\mathbf{X}$ within each partition. This partitioning by $\mathbf{Y}$ is called slicing. When $\mathbf{Y}$ is multi-dimensional, the slicing is replaced by the KCA, which is called K-means inverse regression (KIR) (Setodji and Cook, 2004). Its sample algorithm is as follows.

Step 1. Partition the data by K-means clustering $\mathbf{Y}$ to have h clusters. Let C_k^y stand for the cluster for $k = 1, 2, \ldots h$.

Step 2. Make $\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$, $i = 1, 2, \ldots, n$.

Step 3. Compute the sample means of $\bar{\hat{\mathbf{Z}}}_k$ within each cluster for $k = 1, \ldots, h$, and construct that $\hat{\mathbf{M}}_{\mathrm{KIR}} = ((n_1/n)\bar{\hat{\mathbf{Z}}}_1, (n_2/n)\bar{\hat{\mathbf{Z}}}_2, \ldots, (n_h/n)\bar{\hat{\mathbf{Z}}}_h)$, where n_k stands for the size of the k^{th} cluster.

Step 4. Find the eigenvectors, saying $\hat{\Gamma}_d = (\hat{\gamma}_1, \ldots, \hat{\gamma}(n_1/n_d))$ of corresponding to the first d largest eigenvalues of $\hat{\mathbf{M}}_{\mathrm{KIR}}\hat{\mathbf{M}}_{\mathrm{KIR}}^{\mathrm{T}}$.

Step 5. Let $\hat{\boldsymbol{\eta}} = \hat{\mathbf{\Sigma}}^{-1/2}\hat{\Gamma}_d$, and $\mathcal{S}(\hat{\boldsymbol{\eta}})$ is the estimate of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

2.3.2. Clustering mean method

The method KIR requires the linearity condition. If the condition fails, then the kernel matrix produced by the KIR is not guaranteed to have proper containment of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. According to Li *et al.* (2004), it is possible to reach misleading results because nonlinearity among the predictors possibly makes the performance of most estimation methods worse. Therefore, it is essential to check if the condition holds for KIR. It is not easy to investigate the existence of unobserved nonlinearity among predictors, because they appear linear through graphical inspection. To overcome this issue, Yoo (2016c) proposes the clustering mean method. The main purpose of this article is placed on the methods and its sample implement algorithm is somewhat similar to KIR; therefore, its estimation algorithm is introduced directly. The usual slicing scheme is used to categorize $\mathbf{Y}$ since the clustering mean method (Yoo, 2016c) considers univariate response. For more details on the clustering mean method, readers are recommended to read Yoo (2016c).

Step 1. Partition the data by K-means clustering algorithm for $\mathbf{X}$ to have h_x clusters.

Step 2. Within each cluster, $\mathbf{Y}$ is sliced into h_y groups. So, the data is totally partitioned into $h_x \times h_y$ groups. Let $C_{i(j)}$ denote the partition of the data with the j slice of $\mathbf{Y}$ within the i cluster of $\mathbf{X}$ and let $n_{i(j)}$ be the size of $C_{i(j)}$.

Step 3. Make $\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$, $i = 1, 2, \ldots, n$.

Step 4. Compute the sample means of $\bar{\hat{\mathbf{Z}}}_{i(j)}$ within $C_{i(j)}$ for $i = 1, \ldots, h_x$ and $j = 1, \ldots, h_y$. Build that

$$\hat{\mathbf{M}}_{\mathrm{CCM}} = \left[\left(\frac{n_{1(1)}}{n}\bar{\hat{\mathbf{Z}}}_{1(1)}, \frac{n_{1(2)}}{n}\bar{\hat{\mathbf{Z}}}_{1(2)}, \ldots, \frac{n_{1(h_y)}}{n}\bar{\hat{\mathbf{Z}}}_{1(h_y)}\right), \ldots, \left(\frac{n_{h_x(1)}}{n}\bar{\hat{\mathbf{Z}}}_{h_x(1)}, \frac{n_{h_x(2)}}{n}\bar{\hat{\mathbf{Z}}}_{h_x(2)}, \ldots, \frac{n_{h_x(h_y)}}{n}\bar{\hat{\mathbf{Z}}}_{h_x(h_y)}\right)\right].$$

Step 5. Find the eigenvectors, saying $\hat{\Gamma}_d = (\hat{\gamma}_1, \ldots, \hat{\gamma}_d)$ of corresponding to the first d largest eigenvalues of $\hat{\mathbf{M}}_{\mathrm{CCM}}\hat{\mathbf{M}}_{\mathrm{CCM}}^{\mathrm{T}}$.

Step 6. Let $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\Gamma}_d$, and $\mathcal{S}(\hat{\boldsymbol{\eta}})$ is the estimate of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

2.4. Hclust inverse regression and hierarchical clustering mean method

The goal of this paper is to analyze how HCA can be used in sufficient dimension reduction. So, we propose Hclust inverse regression (HIR) to estimate $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ for multivariate regression. The implementation of HIR is the same as KIR, except to hierarchically-cluster the responses in Step 1 of the KIR algorithm. Four measures are used to measure distances between clusters in HCA; therefore, sHIR, cHIR, aHIR and wHIR are acronymically named according to using single, complete and average linkages and Ward's method, respectively.

We also distinguish KCA and HCA in the clustering mean method. The KCA method used to categorize the predictors is called the K-means clustering mean method (KCCM), which is the original version proposed by Yoo (2016c). However, it will be called Hierarchical clustering mean method (HCCM) if the predictors are hierarchically-clustered. The acronyms of sHCCM, cHCCM, aHCCM, and wHCCM are also defined following the same rationale.

Data generation of $\mathbf{Y}$ and $\mathbf{X}$ would be different since regression is a study of the conditional distribution of $\mathbf{Y}|\mathbf{X}$. In the regression context, it would be common to think that regression data is constructed, as if the predictors are sampled first, and then the responses are generated given the predictors. $\mathbf{X}$ is involved for the generation of $\mathbf{Y}$; however, the opposite direction does not hold. This indicates that different measuring options would be preferred for responses and predictors in the HCA.

3. Numerical studies

For all numerical studies, the sample sizes were 100, and each model was iterated 1000 times. For all models, the dimension of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ was one. The correlation coefficient between $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ and $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ was calculated to measure how the central subspace is well-estimated, where $\hat{\boldsymbol{\eta}}$ stands for the sample estimate of $\boldsymbol{\eta}$. As a summary, boxplots for each clustering method were reported along with lining medians. The single linkage results were omitted since it was the worst in most cases.

3.1. In case of clustering responses

Two models in Setodji and Cook (2004) were studied. And, the dimensions of responses and predictors were four and five, respectively. So that we have a multivariate regression of $\mathbf{Y} = (y_1, \ldots, y_4)^{\mathrm{T}}|\mathbf{X} = (x_1, \ldots, x_5)^{\mathrm{T}}$. The variable configurations were then: $(x_1, \ldots, x_5, \varepsilon_1, \ldots, \varepsilon_4)^{\mathrm{T}} \overset{iid}{\sim} N(0, 1)$. The number of clusters considered were 2, 3, …, 8.

- **Model 1.** $y_1 = 1_5^{\mathrm{T}}\mathbf{X} + 0.1\varepsilon_1$; $y_2 = |1_5^{\mathrm{T}}\mathbf{X}| + 0.1\varepsilon_2$; $y_3 = y_1^2 + y_2\varepsilon_3$; $y_4 = \varepsilon_4$, where $1_5 = (1, 1, 1, 1, 1)^{\mathrm{T}}$. In Model 1, the column vector of $\boldsymbol{\eta} = 1_5$ span $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, and $y_4 \perp\!\!\!\perp \mathbf{X}$. The coordinate regression of $y_2|\mathbf{X}$ is symmetric at zero. The coordinate regression of $y_3|\mathbf{X}$ has quadratic mean function and heteroscedasticity. Therefore, this multivariate regression has quite complex means structure and heteroscedasticity.

- **Model 2.** $y_1 = 0.1(1_4^{\mathrm{T}}\mathbf{X})+\exp(0.1(1_4^{\mathrm{T}}\mathbf{X}))\varepsilon_1$; $y_2 = 0.1(1_4^{\mathrm{T}}\mathbf{X})+\exp(0.2-0.3(1_4^{\mathrm{T}}\mathbf{X}))\varepsilon_2$; $y_3 = 0.1(1_4^{\mathrm{T}}\mathbf{X})+\exp(0.2(1_4^{\mathrm{T}}\mathbf{X}))\varepsilon_3$; $y_4 = 0.1(1_4^{\mathrm{T}}\mathbf{X}) + \exp(0.1 - 0.1(1_4^{\mathrm{T}}\mathbf{X}))\varepsilon_4$, where $1_4 = (1, 1, 1, 1, 0)^{\mathrm{T}}$. For Model 2, the mean and variance functions depend on $\mathbf{X}$ only through $1_4^{\mathrm{T}}\mathbf{X}$, so the column of $\boldsymbol{\eta} = 1_4$ spans

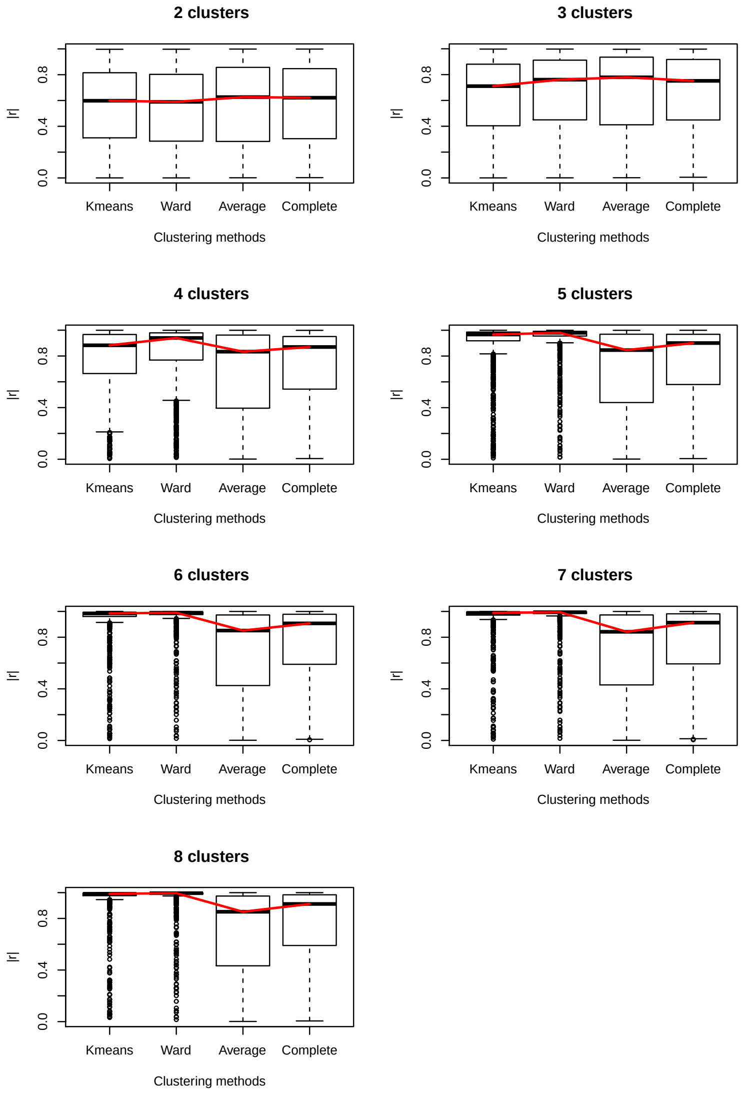


Figure 1: *Side-by-side boxplots for Model 1; red line, the median of* $|r|$*s.*

$\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ with $d = 1$. All coordinate regressions have linear mean and heteroscedasticity as a form of exponential function of $1_4^{\mathrm{T}}\mathbf{X}$.

The side-by-side boxplots of the correlation coefficients $|r|$ between $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ and $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ for the two models are reported in Figures 1 and 2. According to Figures 1 and 2, the three of KIR, wHIR and cHIR result in equally good estimation performances for all numbers of clusters, although the wHIR shows better estimation performances with smaller number of clusters than the other two. The estimation results by aHIR are worse than the former two, especially for Model 2. However, the responses may have unexpected outliers because there exists heteroscedasticity in Model 2 that would affect the

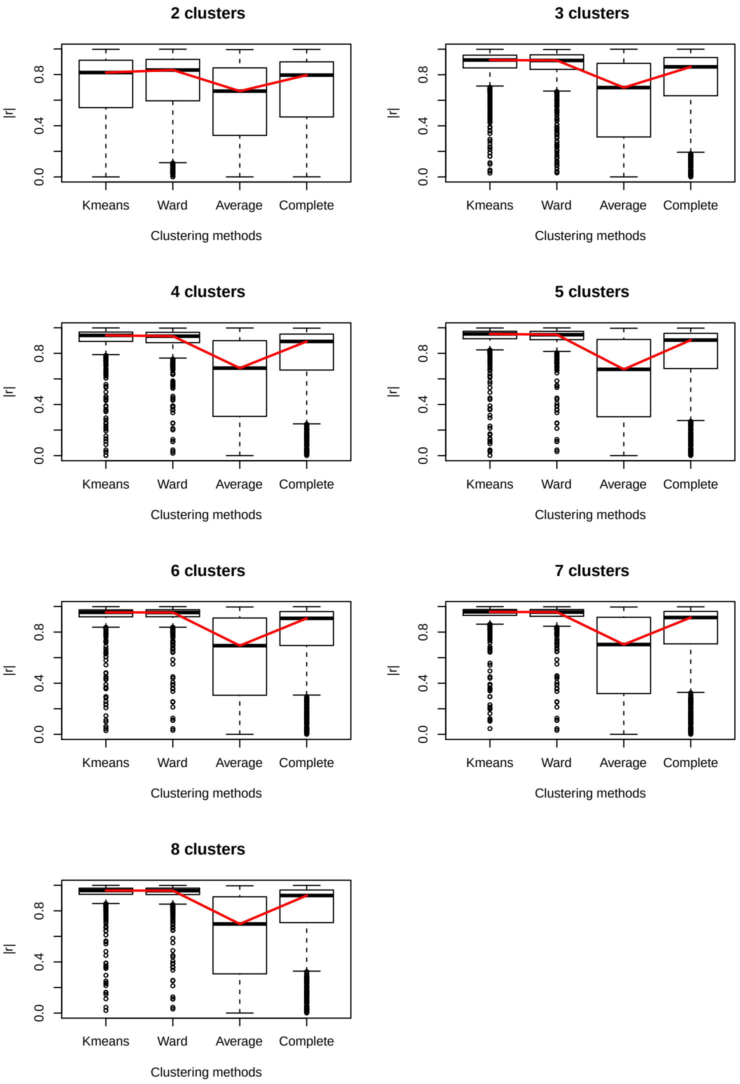


Figure 2: *Side-by-side boxplots for Model 2; red line, the median of* $|r|$*s.*

average and possibly cause bad clustering results.

From the studies, the wHIR is confirmed to compete KIR successfully, which is recommended as a default when hierarchically-cluster responses are required in inverse regression methods.

3.2. In case of clustering predictors

We considered the following two models in Yoo (2018). KCCM and HCCM were fitted to the data in both models to estimate $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ since the linearity did not hold because the predictors were nonlinear.

- **Model 3.** Variables $(u_1, u_3, u_4, u_5, \epsilon, \epsilon, \varepsilon)$ were independently generated as: $u_1 \sim U(0, 1)$, $u_2 =$

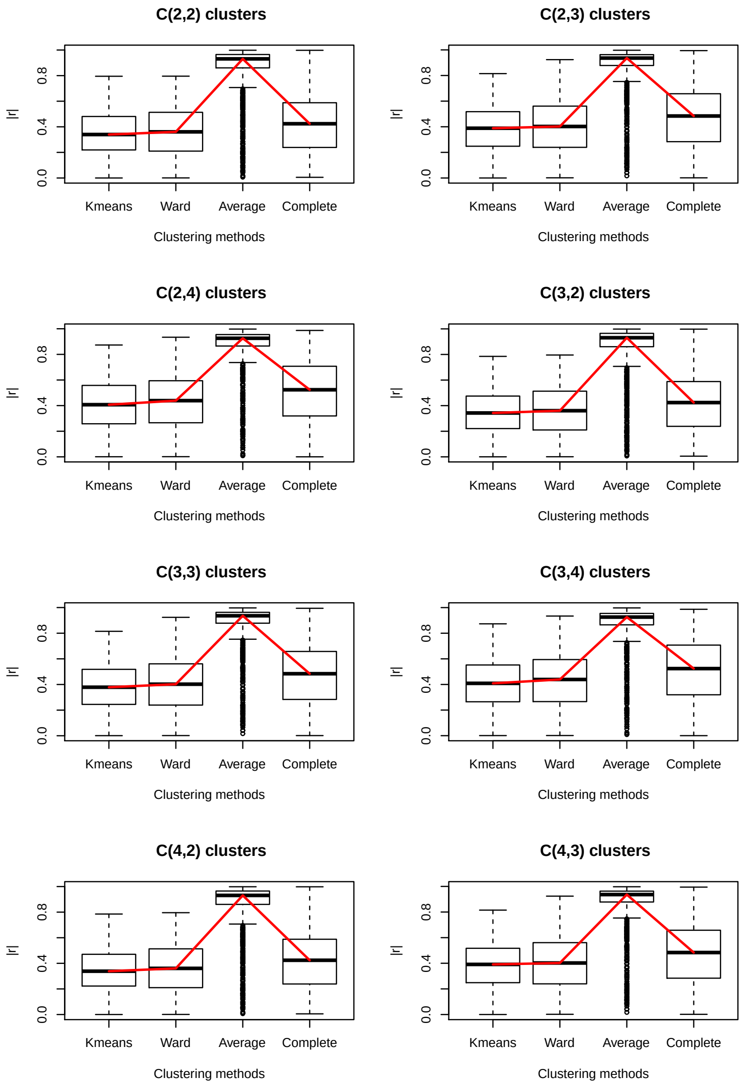


Figure 3: *Side-by-side boxplots for Model 3; red line, the median of* $|r|$*s.*

$\log(u_1) + \epsilon$, $\epsilon \sim U(-0.5, 0.5)$ and $(u_3, u_4, u_5, \varepsilon) \overset{\text{iid}}{\sim} N(0, 1)$. $x_1 = u_1 + u_3$; $x_2 = u_2 + u_4 + u_5$; $x_3 = u_3 - u_4$; $x_4 = u_4$; $x_5 = u_5$. $y|\mathbf{X} = (x_1, x_2, x_3, x_4, x_5)^{\mathrm{T}} = \log(x_1 - x_3 - x_4) + \varepsilon$. In Model 3, a vector of $(1, 0, -1, -1, 0)^{\mathrm{T}}$ spans $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and its structural dimension is one.

- **Model 4.** $(x_1, \ldots, x_8)^{\mathrm{T}} \overset{\text{iid}}{\sim} \exp(1) \perp\!\!\!\perp \varepsilon \sim N(0, 1)$. $y|\mathbf{X} = (x_1, \ldots, x_8)^{\mathrm{T}} = \exp(x_1 - x_2) + 0.5\varepsilon$. For Model 4, the central subspace is one-dimensional and is spanned by the vector of $(1, -1, 0, \ldots, 0)^{\mathrm{T}}$.

Figures 3 and 4 provides the summary plots for Models 3 and 4. According to the figures, the aHCCM and cHCCM dominate KCCM and wHCCM, which are a different aspect from the clustering responses. The aHCCM for Model 3 is the best and highly distinguished from the other three. Also,

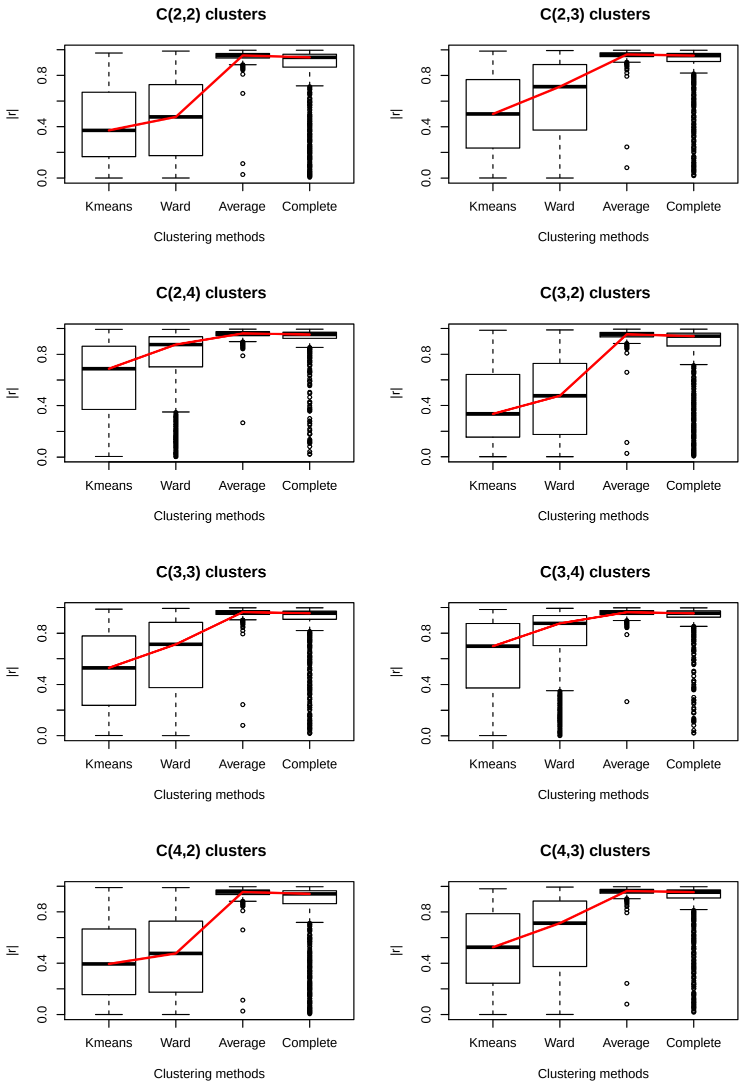
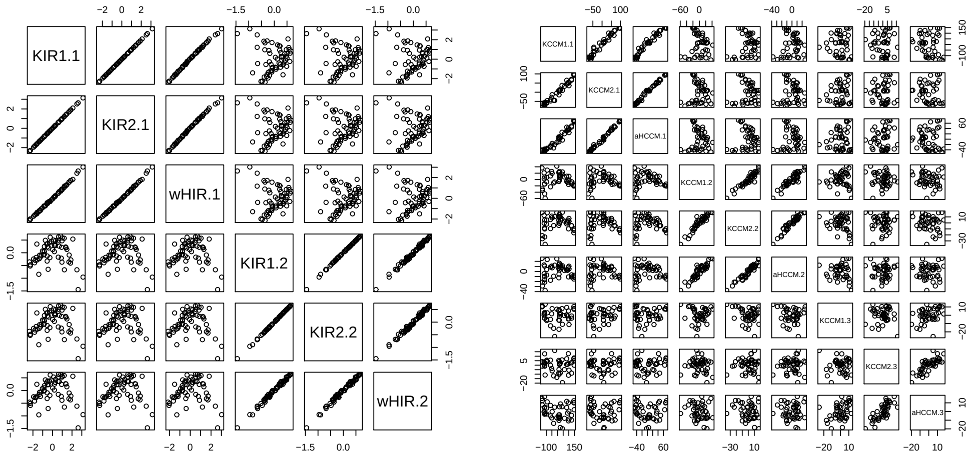


Figure 4: *Side-by-side boxplots for Model 4; red line, the median of* $|r|$*s.*

it is observed that the variability in $|r|$s by aHCCM is smaller than the others. Through the various numerical studies, the aHCCM is recommended for clustering the predictors in the clustering mean method. This confirms that the use of HCA can improve the estimation results and provide a more reliable estimate than KCA.

4. Real data examples: Minneapolis school data and Massachusetts college data

We considered two data sets for illustration purposes. The first example is a multivariate regression using data regarding the performance of students in $n = 63$ Minneapolis schools studied by Yoo (2009). The four dimensional responses $\mathbf{Y}$ are the percentage of students in a school scoring above and



(a) Minneapolis school data (b) Massachusetts college data

Figure 5: *Scatterplot matrices of sufficient predictors.*

below average on standardized fourth and sixth grade reading comprehension tests. The following six predictors were considered: pupil teacher ratio, square roots of percentage of children receiving Aid to Families with Dependent Children, percentage of children not living with both biological parents, percentage of adults in the school area who completed high school, and percentage of persons in the area below the federal poverty level. The predictors were transformed to satisfy the linearity condition. For this multivariate regression, KIR and wHIR were implemented with 5 clusters.

To see the practical usefulness of hierarchically-clustering predictors Massachusetts college data were considered introduced in Yoo (2016c). Data was collected to investigate how the percentage of students graduating were associated with the measures of quality for incoming students and the features of the colleges. The response represents the percentage of students graduating from Massachusetts 4-year colleges in 1995. The following seven variables were then used as predictors: percentage of freshmen that were among the top 25% percent of their graduating high school class, median mathematics SAT score, median verbal SAT score, percentage of applicants accepted into college, percentage of accepted applicants who enroll, student-to-faculty ratio, out-of-state tuition and whether the college is public or private (with private coded as 1 and 0 otherwise). After eliminating missing values, 46 observations among total 56 were used for analysis. According to Yoo (2016c), the linearity condition fails because there exists non-linearity in predictors. Therefore, KCCM and aHCCM had better be applied with 8 clusters of ($h_x = 4, h_y = 2$).

The two examples were analyzed by the following approach. Both KIR and KCCM were implemented several times to see that different dimension estimation occurred with level 0.05. We then compared the results with those from wHIR and aHCCM. To gain more information on the dimension determination, sufficient predictors under the maximum dimension estimate were compared through scatterplot matrices (Figure 5).

For Minneapolis school data, the KIR determined the dimension of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ as one (KIR1) with p-value 0.067 for $H_0 : d = 1$ or as two (KIR2) with p-values 0.030 for $H_0 : d = 1$ and 0.523 for $H_0 : d = 2$. However, the wHIR concluded that $\hat{d} = 2$ with p-values 0.038 for $H_0 : d = 1$ and 0.677 for $H_0 : d = 2$. Figure 5(a) indicates that the first two sufficient predictors from KIR1-2 and wHIR are essentially the same. It is natural to think that the KCA produce bad clustering by chance, which causes the unexpected underestimation of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$; therefore, it is reasonable to decide that $\hat{d} = 2$.

In case of Massachusetts college data, the KCMM determined that $\hat{d} = 3$ (KCCM1) with p-values 0.000 for $H_0 : d = 1$, 0.024 for $H_0 : d = 2$ and 0.105 for $H_0 : d = 3$ and $\hat{d} = 2$ (KCCM2) with p-values 0.012 for $H_0 : d = 1$ and 0.299 for $H_0 : d = 2$. The aHCCM resulted in $\hat{d} = 2$ with p-values 0.014 for $H_0 : d = 1$ and 0.503 for $H_0 : d = 2$. Figure 5(b) indicates that the correlations between the first two sufficient predictors are very high, while the third sufficient predictors do not have a common relationships like the first two. This implies that the determination of $\hat{d} = 3$ seems to overestimate the true dimension of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

The two real data examples confirms that the use of HCA has advantage over the KCA in practice.

5. Discussion

Sufficient dimension reduction (SDR) pursues to replace the original p-dimensional predictors with its lower-dimensional linear projection without information loss on a regression of $\mathbf{Y} \in \mathbb{R}^r | \mathbf{X} \in \mathbb{R}^p$, where $r \geq 1$ and $p > 2$. Two popular methods of inverse regression (Li, 1991; Cook and Weisberg, 1991) and clustering mean method (Yoo, 2016c) require partitioning data into subgroups. For this, the K-means clustering algorithm (KCA) has been successfully adopted to partition data into subgroups by clustering responses and predictors.

However, two deficits for KCA to have in SDR is the lack of reproducibility and nestness. If reproducibility is not guaranteed, the dimension reduction results will possibly vary, whenever the same KCA procedure is repeatedly applied. This then causes confusion to practitioners. Nestness is also important to magnify information within similar subgroups and to accumulate information between different subgroups.

The hierarchical clustering algorithm (HCA) can overcome deficits; therefore, HCA would be a potentially good or possible better replacement of KCA in SDR. We propose an Hclust inverse regression and Hierarchical clustering mean method using HCA. Throughout numerical studies, the Ward-Hclust inverse regression, which categorize multi-dimensional responses, can compete with or be better than the existing KIR. The various simulation studies also show the Average-hierarchical clustering mean method and categorize the predictors that outperform the K-means clustering mean method.

The use of hierarchical clustering algorithm is therefore confirmed to be beneficiary in SDR literature.

Acknowledgements

For Chaeyeon Yoo, this work was supported by the BK21 Plus Project through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (22A20130011003). 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-2019R1F1A1050715/2019R1A6A1A11051177).

References

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

Hastie T, Tibshirani R, and Friedman J (2008). *The Elements of Statistical Learning* (2nd ed.), Springer, New York.

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 L, Cook RD, and Nachtsheim CJ (2004). Cluster-based estimation for sufficient dimension reduction, *Computational Statistics and Data Analysis*, **47**, 175–193.

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.

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 (2016c). Sufficient dimension reduction through informative predictor subspace. *Statistics : A Journal of Theoretical and Applied Statistics*, **50**, 1086–1099.

Yoo JK (2018). Partial least squares fusing unsupervised learning. *Chemometrics and Intelligent Laboratory Systems*, **175**, 82–86.

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 March 16, 2020; Revised April 17, 2020; Accepted April 17, 2020