Communications for Statistical Applications and Methods
2016, Vol. 23, No. 2, 93–103

http://dx.doi.org/10.5351/CSAM.2016.23.2.093
Print ISSN 2287-7843 / Online ISSN 2383-4757

Tutorial: Dimension reduction in regression with a notion of sufficiency

Jae Keun Yoo[1,a]

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

Abstract

In the paper, we discuss dimension reduction of predictors $\mathbf{X} \in \mathbb{R}^p$ in a regression of $Y|\mathbf{X}$ with a notion of sufficiency that is called sufficient dimension reduction. In sufficient dimension reduction, the original predictors $\mathbf{X}$ are replaced by its lower-dimensional linear projection without loss of information on selected aspects of the conditional distribution. Depending on the aspects, the central subspace, the central mean subspace and the central k^{th}-moment subspace are defined and investigated as primary interests. Then the relationships among the three subspaces and the changes in the three subspaces for non-singular transformation of $\mathbf{X}$ are studied. We discuss the two conditions to guarantee the existence of the three subspaces that constrain the marginal distribution of $\mathbf{X}$ and the conditional distribution of $Y|\mathbf{X}$. A general approach to estimate them is also introduced along with an explanation for conditions commonly assumed in most sufficient dimension reduction methodologies.

Keywords: central subspace, central k^{th}-moment subspace, central mean subspace, dimension reduction subspace, regression, sufficient dimension reduction

1. Introduction

High-dimensional data can arise in any place at any moment. It is common that useful information be extracted from such data for important decision making and that necessary statistical models be built to investigate the association between variables and prediction. In these cases, dimension reduction of data is inevitable to avoid obstacles like the curse of dimensionality. This is one reason why various dimension reduction techniques have been developed and remain popular.

To have more intuition of necessity of dimension reduction, the prediction of Y is of main interest in the following regression of $Y|\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$:

$$Y|\mathbf{X} = X_1 + \varepsilon, \tag{1.1}$$

where $(X_1, \ldots, X_p) \overset{iid}{\sim} N(0,1) \perp\!\!\!\perp \varepsilon \sim N(0,1)$ and $\perp\!\!\!\perp$ stands for independence. By construction, the regression in (1.1) depends on $\mathbf{X}$ only through X_1 regardless of the number of predictors, p.

We consider the following two ways to predict Y. First, the response Y is predicted in the usual way of fitting the multiple linear regression $Y|\mathbf{X} \in \mathbb{R}^p$ and then doing prediction. The second way is to predict Y from a simple linear regression of $Y|\hat{\beta}^{\mathrm{T}}\mathbf{X} \in \mathbb{R}^1$, where $\hat{\beta} \in \mathbb{R}^p$ is the ordinary least square coefficient vector estimated from the multiple linear regression. The clear difference between the two is placed on the dimension of predictors used in fitting the regression. The dimension in

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

Published 31 March 2016 / journal homepage: http://csam.or.kr

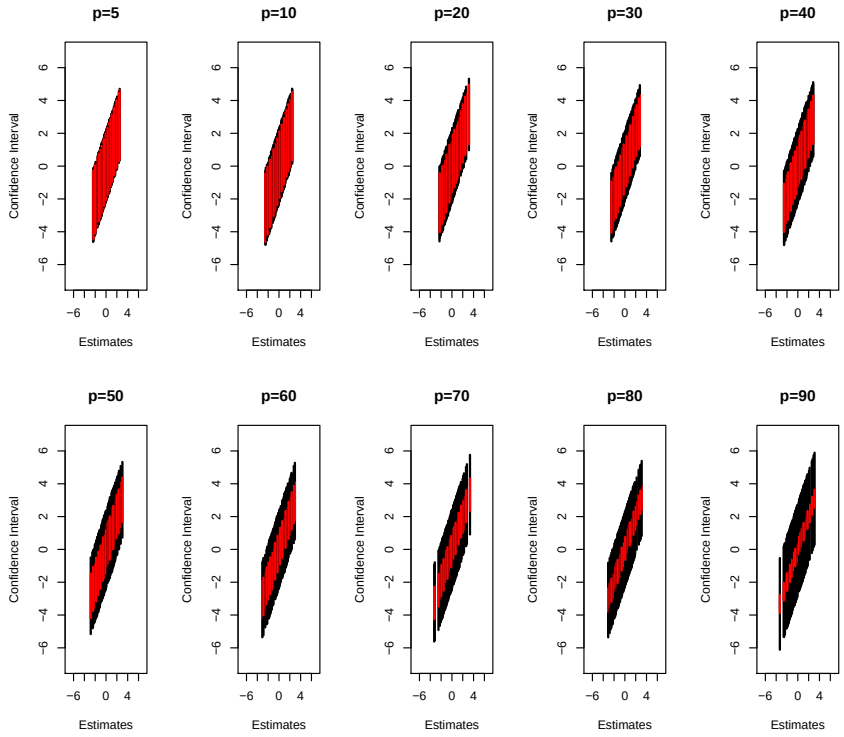

Figure 1: *Prediction confidence interval in toy regression; black, usual multiple linear regression fit; red, dimension reduction linear regression fit.*

the former is p, while it is always 1 in the latter where the dimension of $\mathbf{X}$ is reduced from p to 1. Prediction confidence intervals are computed to compare the performances between the two with varying $p = 5, 10, 20, 30, 40, 50, 60, 70, 80, 90$ (Figure 1). The sample sizes in all cases were $n = 100$.

Figure 1 shows that the differences between the two prediction confidence intervals are even larger, as p increases. This simple example shows that the dimension reduction should be inevitable in high-dimensional regression.

We now focus on dimension reduction of predictors $\mathbf{X}$ in regression. Especially, we will seek for dimension reduction of $\mathbf{X}$ without loss of information on selected aspects of the conditional distribution of $Y|\mathbf{X}$. This type of dimension reduction approach is called *sufficient dimension reduction* (SDR) because a notion of reduction without loss of information is directly related to *sufficiency*.

A regression is to study the conditional distribution of $Y|\mathbf{X}$. Let $F_{Y|\mathbf{X}}(\cdot)$ denoted as the distribution function of $Y|\mathbf{X}$. The goal of SDR is to find lower-dimensional function of $\mathbf{X}$, namely $g(\mathbf{X})$ such that

$$F_{Y|\mathbf{X}}(\cdot) = F_{Y|g(\mathbf{X})}(\cdot). \tag{1.2}$$

Statement (1.2) indicates that two regression of $Y|\mathbf{X}$ and $Y|g(\mathbf{X})$ are the same. We can equivalently consider $Y|g(\mathbf{X})$ in order to study $Y|\mathbf{X}$. That is, the lower-dimensional predictor $g(\mathbf{X})$ can replace $\mathbf{X}$ without loss of information on the conditional distribution of $Y|\mathbf{X}$, if statement (1.2) holds for $g(\mathbf{X})$. To stress this more clearly, statement (1.2) is equivalently rephrased to the following independence

statement:

$$Y \perp\!\!\!\perp \mathbf{X} | g(\mathbf{X}). \tag{1.3}$$

Statement (1.3) directly indicates that $g(\mathbf{X})$ has the same amount of information on $Y|\mathbf{X}$ as $\mathbf{X}$ does.

There are many possibilities regarding forms of $g(\cdot)$. In SDR, we consider a lower-dimensional linear transformation $\mathbf{B}^{\mathrm{T}}\mathbf{X}$ of $\mathbf{X}$, where $\mathbf{B} \in \mathbb{R}^{p\times q}$ with $q < p$. Then the goal of SDR is to pursue a lower-dimensional linear projection $\mathbf{B}^{\mathrm{T}}\mathbf{X}$ of the original p-dimensional predictors $\mathbf{X}$ such that

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^{\mathrm{T}}\mathbf{X}. \tag{1.4}$$

In the past two decades SDR methods for regression have been rapidly developed. Most of them can be considered nonparametric because they do not assume a specific form of $Y|\mathbf{X}$; however, mild conditions are required. Unlike many local nonparametric approaches, they can often avoid the curse of dimensionality because their estimates are global and converge at the usual $\sqrt{n}$ rate.

The goal of the paper is to provide a notion of sufficient dimension reduction. For this, the paper is organized as follows. Section 2 introduces dimension reduction subspaces and target dimension reduction subspaces depending on selected aspects of regression. Some existing conditions for them are also described. Section 3 is devoted to providing general approach of inference on the target subspaces. Common conditions required in most sufficient dimension reduction methods are introduced in Section 4. In Section 5, we summarize the work.

2. Dimension reduction subspaces and centrality

Before starting this section, we set up four models for illustration purpose. For each model, the following variable configuration are commonly used: $\mathbf{X} = (X_1, \ldots, X_5)^{\mathrm{T}} \overset{iid}{\sim} N(0,1) \perp\!\!\!\perp \varepsilon \sim N(0,1)$.

Example 1: $Y|\mathbf{X} = \sum_{i=1}^{5} X_i + \varepsilon$;

Example 2: $Y|\mathbf{X} = X_1(X_1 + X_2) + \varepsilon$;

Example 3: $Y|\mathbf{X} = X_1 + \exp(X_2)\varepsilon$;

Example 4: $Y|\mathbf{X} \sim B(m, p)$, where $p = \exp(\sum_{i=1}^{3} X_i)/\{1 + \exp(\sum_{i=1}^{3} X_i)\}$ and $B(m,p)$ stands for a binomial distribution with total m trials and a success probability p.

2.1. Central subspace

For Examples 1–4 above, it can be seen that the conditional distributions of $Y|\mathbf{X}$ depend on $\mathbf{X}$ only through $\sum_{i=1}^{5} X_i$, (X_1, X_2), (X_1, X_2) and $\sum_{i=1}^{3} X_i$, respectively. Therefore, the following relation can be easily established:

Example 1: $F_{Y|\mathbf{X}} = F_{Y|\sum_{i=1}^{5} X_i} = F_{Y|\mathbf{B}_1^{\mathrm{T}}\mathbf{X}}$, where $\mathbf{B}_1 = (1,1,1,1,1)^{\mathrm{T}}$;

Example 2: $F_{Y|\mathbf{X}} = F_{Y|X_1,X_2} = F_{Y|\mathbf{B}_2^{\mathrm{T}}\mathbf{X}}$, where $\mathbf{B}_2 = \{(1,0,0,0,0),(0,1,0,0,0)\}^{\mathrm{T}}$;

Example 3: $F_{Y|\mathbf{X}} = F_{Y|X_1,X_2} = F_{Y|\mathbf{B}_3^{\mathrm{T}}\mathbf{X}}$, where $\mathbf{B}_3 = \{(1,0,0,0,0),(0,1,0,0,0)\}^{\mathrm{T}}$;

Example 4: $F_{Y|\mathbf{X}} = F_{Y|\sum_{i=1}^{3} X_i} = F_{Y|\mathbf{B}_4^{\mathrm{T}}\mathbf{X}}$, where $\mathbf{B}_4 = (1,1,1,0,0)^{\mathrm{T}}$.

Therefore, with $\mathbf{B}_i^{\mathrm{T}}\mathbf{X}$, $i = 1, 2, 3, 4$, the goal of SDR is achieved for the four regressions, respectively.

For Example 2, consider the following matrices along with $\mathbf{B}_2$:

$$\begin{aligned}
&\mathbf{B}_{c_1}^{\mathrm{T}}\mathbf{X} = (X_1, X_1 + X_2) \quad \text{with } \mathbf{B}_{c_1} = \{(1,0,0,0,0)(1,1,0,0,0)\}^{\mathrm{T}};\\
&\mathbf{B}_{c_2}^{\mathrm{T}}\mathbf{X} = (X_2, X_1 + X_2) \quad \text{with } \mathbf{B}_{c_2} = \{(0,1,0,0,0)(1,1,0,0,0)\}^{\mathrm{T}};\\
&\mathbf{B}_{c_3}^{\mathrm{T}}\mathbf{X} = \left(\frac{1}{2}X_1, X_1 - X_2\right) \quad \text{with } \mathbf{B}_{c_3} = \left\{\left(\frac{1}{2},0,0,0,0\right)(1,-1,0,0,0)\right\}^{\mathrm{T}};\\
&\mathbf{B}_{c_4}^{\mathrm{T}}\mathbf{X} = (-X_1, X_1 + 2X_2) \quad \text{with } \mathbf{B}_{c_4} = \{(-1,0,0,0,0)(1,2,0,0,0)\}^{\mathrm{T}}.
\end{aligned}$$

Define the two coordinates of $\mathbf{B}_2^{\mathrm{T}}\mathbf{X}$ and $\mathbf{B}_{c_i}^{\mathrm{T}}\mathbf{X}$ as (x_1, x_2) and $(x_{c_i,1}, x_{c_i,2})$, respectively. Then, pairs of (x_1, x_2) have one-to-one mapping with those of $(x_{c_i,1}, x_{c_i,2})$ for $i = 1, \ldots, 4$. That is, any choices of $\mathbf{B}_i$ can be perfectly converted $\mathbf{B}_2$, and hence we have that $F_{Y|\mathbf{X}} = F_{Y|\mathbf{B}_2^{\mathrm{T}}\mathbf{X}} = F_{Y|\mathbf{B}_{c_i}^{\mathrm{T}}\mathbf{X}}$ for $i = 1, \ldots, 4$. Then, which one should be chosen among the five candidates? The answer for the question is that any of the five should be good.

Now we will see this choice problem in another direction. Define that $\mathcal{S}(\mathbf{B})$ represents a subspace spanned by the columns of $\mathbf{B} \in \mathbb{R}^{p\times q}$. It is easily seen that $\mathcal{S}(\mathbf{B}_2) = \mathcal{S}(\mathbf{B}_{c_i})$ for $i = 1, \ldots, 4$. $\mathbf{B}_2$ and $\mathbf{B}_{c_i}$ are different; however, their column subspaces are all the same. The choice problem is then no longer an issue if we consider a subspace spanned by the columns of $\mathbf{B}$ to satisfy (1.4). Any orthonormal basis of the subspace should also be fine because all of them span the same subspace. Based on this discussion, we define a dimension reduction subspace as:

Definition 1. *A subspace spanned by the columns of* $\mathbf{B}$ *such that* $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^{\mathrm{T}}\mathbf{X}$ *is called dimension reduction subspace (DRS) (Cook, 1998).*

A DRS always exists because $\mathbf{B}$ can be the identity matrix. Then, another choice problem arises because a regression has two or more DRSs. In Example 2, the columns of $\{(1,0,0,0,0),(0,1,0,0,0)\}^{\mathrm{T}}$ forms a DRS, but the columns of $\{(1,0,0,0,0),(0,1,0,0,0),(0,0,1,0,0)\}^{\mathrm{T}}$ also does a DRS. To choose the minimal subspace among all possible DRSs should be desirable because the minimal one still have full information of $Y|\mathbf{X}$. How then do we find the minimal one among many DRSs? We will consider the intersection of all possible DRSs defined as follows.

Definition 2. *If the intersection of all possible DRSs is a dimension reduction subspace, the intersection is called central subspace (Cook, 1998),* $\mathcal{S}_{Y|\mathbf{X}}$.

If $\mathcal{S}_{Y|\mathbf{X}}$ exists, it is minimal and unique. Therefore, the recovery of $\mathcal{S}_{Y|\mathbf{X}}$ is naturally the mainstream in SDR. The central subspace does not always exist. We will closely investigate the conditions to guarantee the existence of $\mathcal{S}_{Y|\mathbf{X}}$ in later section.

2.2. Central mean subspace

The central subspace is to provide a complete information of the dependence of $Y|\mathbf{X}$, but certain aspects of $Y|\mathbf{X}$ may be of primary interest, not $Y|\mathbf{X}$ itself. Indeed, a regression problem is often understood by a study of the mean function $E(Y|\mathbf{X})$. If so, investigating $E(Y|\mathbf{X})$ through $\mathcal{S}_{Y|\mathbf{X}}$ should be overwork because the scope of $\mathcal{S}_{Y|\mathbf{X}}$ is usually expected to be larger than necessary. The regressions like Examples 1, 2 and 4 depend on $\mathbf{X}$ only through their mean functions, and hence the consideration of $E(Y|\mathbf{X})$ is adequate to capture all information on $Y|\mathbf{X}$.

We now move our main interest from $Y|\mathbf{X}$ to $E(Y|\mathbf{X})$. Then, in SDR context, we need to replace the original p-dimensional predictor by a lower-dimensional linearly transformed one without loss of

information on $E(Y|\mathbf{X})$. If so, one should recover a subspace spanned by the columns of $\mathbf{B}$ such that

$$E(Y|\mathbf{X}) = E\left(Y|\mathbf{B}^{\mathrm{T}}\mathbf{X}\right) \Leftrightarrow Y \perp\!\!\!\perp E(Y|\mathbf{X})|\mathbf{B}^{\mathrm{T}}\mathbf{X}, \tag{2.1}$$

where $\mathbf{B}$ is a $p \times q$ matrix with $q < p$.

Based on this discussion, a mean subspace is newly defined.

Definition 3. *A mean subspace (Cook and Li, 2002) is defined as a subspace spanned by the columns of* $\mathbf{B} \in \mathbb{R}^{p\times q}$ *to satisfy* $Y \perp\!\!\!\perp E(Y|\mathbf{X})|\mathbf{B}^{\mathrm{T}}\mathbf{X}$.

When the mean function has all information on $Y|\mathbf{X}$, that is, satisfying

$$Y \perp\!\!\!\perp \mathbf{X}|E(Y|\mathbf{X}),$$

the regression is often called location regression. The location regression forms a large class of regression including many generalized linear models. Their special case is an additive-error regression model (such as usual multiple linear regression or single-index model) in which we have

$$Y - E(Y|\mathbf{X}) \perp\!\!\!\perp \mathbf{X}.$$

In this location regression, a mean subspace should be the primary target over $\mathcal{S}_{Y|\mathbf{X}}$.

Since there can be many mean subspaces in a regression problem, intersecting all possible mean subspaces is required to obtain the unique and minimal one.

Definition 4. *If the intersection of all possible mean subspaces is a mean subspace, the intersection is called central mean subspace (Cook and Li, 2002),* $\mathcal{S}_{E(Y|\mathbf{X})}$.

The conditions to guarantee the existence of $\mathcal{S}_{E(Y|\mathbf{X})}$ are the same as $\mathcal{S}_{Y|\mathbf{X}}$; therefore, they are never a cause of concern in practice.

2.3. Central k^{th}-moment subspace

Recall Example 3. In the example, the regression depends on $\mathbf{X}$ only through X_1 and X_2. Since $\mathcal{S}_{E(Y|\mathbf{X})}$, spanned by $(1, 0, 0, 0, 0)^{\mathrm{T}}$ provides information of X_1 alone, $\mathcal{S}_{E(Y|\mathbf{X})}$ is inadequate to summarize the regression. However, the consideration of the second conditional moment of $\mathrm{var}(Y|\mathbf{X}) = \exp(2X_2)$ along with $E(Y|\mathbf{X})$ can characterize the regression and provide the same information as $\mathcal{S}_{Y|\mathbf{X}}$. That is, in Example 2, while $\mathcal{S}_{E(Y|\mathbf{X})}$ is clearly smaller than necessary, $\mathcal{S}_{Y|\mathbf{X}}$ may be larger than necessary. Expanding the idea of the conditional moments of $Y|\mathbf{X}$ like $\mathcal{S}_{E(Y|\mathbf{X})}$, we newly construct a dimension reduction subspace of $\mathbf{X}$. This construction is similar to $\mathcal{S}_{E(Y|\mathbf{X})}$; however, the goal is to reduce the mean function as well as variance function and up to the k^{th} moment function, leaving the rest of $Y|\mathbf{X}$ as the nuisance parameter.

If so, SDR pursues $\mathbf{B}^{\mathrm{T}}\mathbf{X}$ for $\mathbf{B} \in \mathbb{R}^{p\times q}$ with $q < p$ such that, for $i = 1, \ldots, k$,

$$M^{(i)}(Y|\mathbf{X}) = M^{(i)}\left(Y|\mathbf{B}^{\mathrm{T}}\mathbf{X}\right),$$

where $M^{(k)}(Y|\mathbf{X}) = E[\{Y - E(Y|\mathbf{X})\}^k|\mathbf{X}]$ and $M^{(1)}$ is replaced by $E(Y|\mathbf{X})$. Again, the statement of $M^{(i)}(Y|\mathbf{X}) = M^{(i)}(Y|\mathbf{B}^{\mathrm{T}}\mathbf{X})$, $i = 1, \ldots, k$, can be re-written as the following conditional independence statement:

$$Y \perp\!\!\!\perp \left\{E(Y|\mathbf{X}), \ldots, M^{(k)}(Y|\mathbf{X})\right\} \left|\mathbf{B}^{\mathrm{T}}\mathbf{X}\right. . \tag{2.2}$$

Following the discussion, a k^{th}-moment subspace and the central k^{th}-moment subspace (Yin and Cook, 2002) are newly defined as:

Definition 5. *A k^{th}-moment subspace is defined as a subspace spanned by the columns of $\mathbf{B} \in \mathbb{R}^{p\times q}$ such that $Y \perp\!\!\!\perp \{E(Y|\mathbf{X}), \ldots, M^{(k)}(Y|\mathbf{X})\}|\mathbf{B}^{\mathrm{T}}\mathbf{X}$.*

Definition 6. *If the intersection of all possible k^{th}-moment subspaces is a k^{th}-moment subspace, the intersection is called central k^{th}-moment subspace, $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$.*

If $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ exists, it is unique and minimal. Also, under the conditions to guarantee existence of $\mathcal{S}_{Y|\mathbf{X}}$, $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ exits and is therefore not a cause of concern.

2.4. Relation of $\mathcal{S}_{Y|\mathbf{X}}$, $\mathcal{S}_{E(Y|\mathbf{X})}$, $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ and non-singular transformation of $\mathbf{X}$

If $\mathcal{S}_{E(Y|\mathbf{X})}$, $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ and $\mathcal{S}_{Y|\mathbf{X}}$ exist for a regression of $Y|\mathbf{X}$, we have $\mathcal{S}_{E(Y|\mathbf{X})} \subseteq \mathcal{S}^{(k)}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. Also, the following four relationships naturally hold (Cook, 1998; Yin and Cook, 2002):

$$
\begin{aligned}
&\mathcal{S}^{(1)}_{Y|\mathbf{X}} = \mathcal{S}_{E(Y|\mathbf{X})};\\
&\mathcal{S}_{E(Y|\mathbf{X})} \subseteq \mathcal{S}^{(2)}_{Y|\mathbf{X}} \subseteq \cdots \mathcal{S}^{(k)}_{Y|\mathbf{X}} \cdots \subseteq \mathcal{S}_{Y|\mathbf{X}};\\
&\lim_{k\to\infty} \mathcal{S}^{(k)}_{Y|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}};\\
&\mathcal{S}_{E(Y|\mathbf{X})} = \mathcal{S}^{(k)}_{Y|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}} \text{ under a location regression.}
\end{aligned}
$$

We normally expect that $\mathcal{S}_{E(Y|\mathbf{X})} \subseteq \mathcal{S}^{(2)}_{Y|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$ because the regression models for $Y|\mathbf{X}$ often depend on $\mathbf{X}$ only through the first two conditional moments.

Let $\mathcal{S}_{\mathbf{X}}$ denote one of $\mathcal{S}_{Y|\mathbf{X}}$, $\mathcal{S}_{E(Y|\mathbf{X})}$ or $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ for a regression of $Y|\mathbf{X}$. Then we have the following result for a non-singular transformation of $\mathbf{X}$.

Results 1. *Let $\mathbf{A}$ be a $p\times p$ non-singular matrix and define that $\mathbf{Z} = \mathbf{A}^{\mathrm{T}}\mathbf{X}$. Consider $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{S}_{\mathbf{Z}}$ from two regressions of $Y|\mathbf{X}$ and $Y|\mathbf{Z}$, respectively. Then, we have $\mathcal{S}_{\mathbf{X}} = \mathbf{A}\mathcal{S}_{\mathbf{Z}}$.*

Suppose that $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}\{\mathbf{X} - E(\mathbf{X})\}$, where $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X})$ and $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$. Then $\mathbf{Z}$ is a standardized predictor with mean vector equal to zero and covariance matrix equal to the identity matrix. Result 1 directly implies that $\mathcal{S}_{\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{\mathbf{Z}}$. In most SDR methodologies, $\mathcal{S}_{\mathbf{X}}$ should be the primary target, but $\mathcal{S}_{\mathbf{Z}}$ is often restored first and then back-transformed to $\mathcal{S}_{\mathbf{X}}$ for computational stability. It should be noted that the dimension of $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{S}_{\mathbf{Z}}$ are equal, although bases of $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{S}_{\mathbf{Z}}$ are different.

In SDR, any of $\mathcal{S}_{E(Y|\mathbf{X})}$, $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$, or $\mathcal{S}_{Y|\mathbf{X}}$ should be the primary target subspace. Hereafter, its dimension (often called *structural dimension*) and orthonormal basis will be denoted as d and $\boldsymbol{\eta}$ ($\boldsymbol{\eta}_z$ for $\mathbf{Z}$-scale predictors). The dimension-reduced linearly transformed predictors $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ are also known as *sufficient predictors*.

2.5. Conditions to guarantee the existence of $\mathcal{S}_{Y|\mathbf{X}}$

The central subspace may not always exist. Consider $\mathbf{X} = (X_1, X_2)^{\mathrm{T}}$ with $X_1^2 + X_2^2 = 1$. That is, X_1 and X_2 are uniformly distributed on the unit circle. $Y|\mathbf{X} = X_1^2 + \varepsilon$, where $\varepsilon \sim N(0,1) \perp\!\!\!\perp \mathbf{X}$. The regression of $Y|\mathbf{X}$ above depends on $\mathbf{X}$ only through X_1. Therefore, the column of $\boldsymbol{\eta}_1 = (1,0)^{\mathrm{T}}$ forms a DRS.

By construction of $\mathbf{X}$, $X_1^2 = 1 - X_2^2$. a regression of $Y|\mathbf{X} = 1 - X_2^2 + \varepsilon$ is equivalent to the original one of $Y|\mathbf{X} = X_1^2 + \varepsilon$; therefore the alternative regression depends on $\mathbf{X}$ only through X_2 with $\boldsymbol{\eta}_2 = (0,1)^{\mathrm{T}}$.

The column of $\boldsymbol{\eta}_2$ then also forms a DRS. It is easily seen that $\mathcal{S}(\boldsymbol{\eta}_1) \cap \mathcal{S}(\boldsymbol{\eta}_2) = \mathcal{O}$. The central subspace does not exist; however, two DRSs do.

The existence of $\mathcal{S}_{Y|\mathbf{X}}$ can be guaranteed by constraining the marginal distribution of $\mathbf{X}$ and the conditional distribution of $Y|\mathbf{X}$.

Results 2. *Let $\mathcal{S}(\alpha)$ and $\mathcal{S}(\phi)$ be DRSs for $Y|\mathbf{X}$. If $\mathbf{X}$ has a density $f(a) > 0$ for $a \in \Omega_x \subset \mathbb{R}^p$ and $f(a) = 0$ otherwise, and if Ω_x is a convex set, then $\mathcal{S}(\alpha) \cap \mathcal{S}(\phi)$ is a DRS.*

Result 2 directly indicates that $\mathcal{S}_{Y|\mathbf{X}}$ exists in regressions, if $\mathbf{X}$ has a density with a convex support.

Results 3. *Let $\mathcal{S}(\alpha)$ and $\mathcal{S}(\phi)$ be DRSs for a location regression of $Y|\mathbf{X}$ such that $Y \perp\!\!\!\perp \mathbf{X}|E(Y|\mathbf{X})$. If $\mathbf{X}$ has a density f on $\Omega_x \subset \mathbb{R}^p$, and if $E(Y|\mathbf{X})$ can be expressed as a convergent power series in the coordinates of $\mathbf{X} = X_k$,*

$$E(Y|\mathbf{X}) = \sum_{k_1,\ldots,k_p}^{\infty} a_{k_1,\ldots,k_p} X_1^{k_1} \cdots X_p^{k_p},$$

then $\mathcal{S}(\alpha) \cap \mathcal{S}(\phi)$ is a DRS.

Under a location regression, we have that $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X} \Leftrightarrow Y \perp\!\!\!\perp \mathbf{X}|E(Y|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X})$. Result 3 requires $E(Y|\mathbf{X})$ to be well-behaved and requires $\mathbf{X}$ to have a density not necessarily positive everywhere. Result 3 can be easily extended to the case with higher-order conditional moments upto k such that

$$Y \perp\!\!\!\perp \mathbf{X} \,\Big|\, \left\{E(Y|\mathbf{X}), \mathbf{M}^{(2)}(Y|\mathbf{X}), \ldots, \mathbf{M}^{(k)}(Y|\mathbf{X})\right\}.$$

Therefore, according to Results 2–3, its existence is not a crucial practical issue along with those of $\mathcal{S}_{E(Y|\mathbf{X})}$ and $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ since the conditions to guarantee the existence of $\mathcal{S}_{Y|\mathbf{X}}$ is mild. Cook (1998, Chapter 6.4) is recommended for proof or more information on Results 2–3.

3. General approach of inference on $\mathcal{S}_{\mathbf{X}}$

Inference on $\mathcal{S}_{\mathbf{X}}$ has two components to estimate the true structural dimension d and an $p \times d$ orthonormal basis matrix $\boldsymbol{\eta}$. Usually, most SDR methodologies under certain conditions (which will be discussed later) construct a kernel matrix $\mathbf{M} \in \mathbb{R}^{p\times p} \geq 0$ such that

$$\mathcal{S}(\mathbf{M}) = \mathcal{S}_{\mathbf{X}}.$$

Next, $\mathbf{M}$ is spectral-decomposed as:

$$\mathbf{M} = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^{\mathrm{T}},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$ and $\gamma_i^{\mathrm{T}}\gamma_i = 1$ and $\gamma_i^{\mathrm{T}}\gamma_j = 0$, $i \neq j$.

Typically, the structural dimension d is determined first. Since $\mathcal{S}(\mathbf{M}) = \mathcal{S}_{\mathbf{X}}$, the rank of $\mathbf{M}$ is the same as d of $\mathcal{S}_{\mathbf{X}}$. The rank of $\mathbf{M}$ is also equal to the number of non-zero eigenvalues among $\lambda_1, \ldots, \lambda_p$. This directly indicates that the structural dimension is equal to the number of non-zero eigenvalues of $\mathbf{M}$, and d is determined via testing a sequence of hypothesis (Li, 1991; Rao, 1965). Beginning with $m = 0$, test the hypothesis of

$$\mathrm{H}_0 : d = m \quad \text{versus} \quad \mathrm{H}_1 : d > m.$$

If $H_0 : d = m$ is rejected, increment m by 1 and redo the test. The test is stopped for the first time $H_0 : d = m$ is not rejected, and set $\hat{d} = m$.

These sequential tests require statistics Λ_m to test under $H_0 : d = m$, which are the sum of the ordered eigenvalues multiplied by n:

$$\Lambda_m = n \sum_{m+1}^{p} \lambda_i, \quad m = 0, 1, \ldots, (p-1).$$

The large sample distribution of Λ_m depends on SDR methodologies.

Once d is determined to $\hat{d}$, $\hat{\boldsymbol{\eta}}$ becomes a set of eigenvectors corresponding to the first $\hat{d}$ largest eigenvalues such that

$$\hat{\boldsymbol{\eta}} = \left(\gamma_1, \ldots, \gamma_{\hat{d}}\right).$$

Then, $\mathcal{S}(\hat{\boldsymbol{\eta}})$ is an estimate of $\mathcal{S}_{\mathbf{X}}$.

4. Common conditions in sufficient dimension reduction methodologies

Here we review crucial conditions in SDR. Using the relation of $\mathcal{S}_{\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{\mathbf{Z}}$, the conditions are discussed with the regression of $Y|\mathbf{Z}$ rather than that of $Y|\mathbf{X}$. Let $\boldsymbol{\eta}_{\mathbf{Z}} \in \mathbb{R}^{p\times d}$ be an orthonormal basis matrix of $\mathcal{S}_{Y|\mathbf{Z}}$.

4.1. Linearity condition

The linearity condition is very common in SDR literature, which is:

$$E\left(\mathbf{Z}|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z} = \nu\right) \text{ is linear in } \nu.$$

The main role of the linearity condition has been understood to force subspaces spanned by the columns of kernel matrices $\mathbf{M} \in \mathbb{R}^{p\times p}$ produced by SDR methods to be a proper subspace of $\mathcal{S}_{Y|\mathbf{Z}}$ such that $\mathcal{S}(\mathbf{M}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. Therefore, most SDR methods require the condition in their methodological development.

This condition is guaranteed to hold if the predictors $\mathbf{X}$ are elliptically distributed (Eaton, 1986). According to Hall and Li (1993), it is shown that (with large p) the linearity condition may hold to a reasonable approximation in many regressions. Li *et al.* (2004) discuss that nonlinearity among the predictors can degrade the performance of most estimation methods, in some applications that yield completely misleading results. Therefore, it is required to investigate if the condition is satisfied. One popular way is to inspect a scatterplot matrix of the predictors. The condition then seems to be satisfied if all cells of the plot look quite linear. However, there may be a chance that unobserved nonlinearity among the predictors exists despite appearing quite linear. Transforming or re-weighting of predictors is typically done if this linearity condition does not hold. However, a transformation of high-dimensional predictors may be inconvenient or infeasible. The re-weighting is computationally intensive (especially if p is high) and causes a deletion of observations in the data.

Recently, Shao *et al.* (2006) provides a new view of the condition with respect to an adaptive estimation of $\boldsymbol{\eta}$ in the following single index model:

$$Y|\mathbf{X} = g\left(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\right) + \varepsilon,$$

where $g(\cdot)$ is unknown link functions and ε represents random error.

A regression of $Y|\mathbf{X}$ depends on $\mathbf{X}$ only through at most one linear combination $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X} \in \mathbb{R}^1$ of the predictors. Therefore, the single index model can be considered as a special case of SDR. An adaptive estimator is an efficient estimator for a model that is only partially specified. We do not know the form of $g(\cdot)$ in the single-index model above. Therefore the model is partially specified, and an asymptotically efficient estimator of $\boldsymbol{\eta}$ is an adaptive estimator. For an adaptive estimators, one can read a seminal paper by Bickel (1982).

According to Shao *et al.* (2006), linearity condition guarantees existence of an adaptive estimate of $\boldsymbol{\eta}$ in the single-index model. Therefore, linearity condition let $\boldsymbol{\eta}$ estimated with the same efficiency as the link function is known. Thus, the linearity condition seems to substitute for knowing the exact conditional distribution of $Y|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$, which has the same as that of $Y|\mathbf{X}$.

4.2. Constant variance condition

Some SDR methods require the following constant variance condition along with the linearity condition:

$$\operatorname{cov}\left(\mathbf{Z}|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}\right) = \mathbf{Q}_{\boldsymbol{\eta}_{\mathbf{Z}}},$$

where $\mathbf{Q}_{\boldsymbol{\eta}_{\mathbf{Z}}}$ is an orthonormal projection operator onto the orthogonal complement of $\mathcal{S}(\boldsymbol{\eta}_{\mathbf{Z}})$.

To understand a role of the constant variance condition, the inverse conditional variance function $\operatorname{cov}(\mathbf{Z}|Y)$ under linearity condition for $\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}$ can be investigated as:

$$\begin{aligned}
\operatorname{cov}(\mathbf{Z}|Y) &= E\left\{\operatorname{cov}\left(\mathbf{Z}|Y, \boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}\right)\Big|Y\right\} + \operatorname{cov}\left\{E\left(\mathbf{Z}|Y, \boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}\right)\Big|Y\right\} \\
&= E\left\{\operatorname{cov}\left(\mathbf{Z}|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}\right)\Big|Y\right\} + \operatorname{cov}\left\{E\left(\mathbf{Z}|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}\right)\Big|Y\right\} \\
&= E\left[E\left\{\left(\mathbf{Z} - E\left(\mathbf{Z}|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}\right)\right)^2\Big|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}\right\}\Big|Y\right] + \operatorname{cov}\left(\mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}\mathbf{Z}|Y\right) \\
&= E\left[E\left\{\left(\mathbf{Z} - \mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}\mathbf{Z}\right)^2\Big|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}\right\}\Big|Y\right] + \mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}\operatorname{cov}(\mathbf{Z}|Y)\mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}} \\
&= E\left\{\operatorname{cov}\left(\mathbf{Q}_{\boldsymbol{\eta}_{\mathbf{Z}}}\mathbf{Z}|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}\right)\Big|Y\right\} + \mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}\operatorname{cov}(\mathbf{Z}|Y)\mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}} \\
&= \mathbf{Q}_{\boldsymbol{\eta}_{\mathbf{Z}}}E\left\{\operatorname{cov}\left(\mathbf{Z}|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}\right)\Big|Y\right\}\mathbf{Q}_{\boldsymbol{\eta}_{\mathbf{Z}}} + \mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}\operatorname{cov}(\mathbf{Z}|Y)\mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}.
\end{aligned}$$

Assuming that the constant variance condition holds additionally, it is simplified:

$$\operatorname{cov}(\mathbf{Z}|Y) = \mathbf{Q}_{\boldsymbol{\eta}_{\mathbf{Z}}} + \mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}\operatorname{cov}(\mathbf{Z}|Y)\mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}.$$

Under both linearity and constant variance conditions we have the following equivalences for $\mathbf{I}_p - \operatorname{cov}(\mathbf{Z}|Y)$.

$$\mathbf{I}_p - \operatorname{cov}(\mathbf{Z}|Y) = \mathbf{I}_p - \mathbf{Q}_{\boldsymbol{\eta}_{\mathbf{Z}}} - \mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}\operatorname{cov}(\mathbf{Z}|Y)\mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}} = \mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}} - \mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}\operatorname{cov}(\mathbf{Z}|Y)\mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}} = \mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}\left\{\mathbf{I}_p - \operatorname{cov}(\mathbf{Z}|Y)\right\}\mathbf{P}_{\boldsymbol{\eta}_{\mathbf{Z}}}.$$

This implies that $\mathbf{I}_p - \operatorname{cov}(\mathbf{Z}|Y)$ can provide some information on $\mathcal{S}_{Y|\mathbf{Z}}$.

Cook (2000) discusses the constant variance condition as: the condition holds, if $\mathbf{Z}$ is normally distributed, or it approximately holds, if $\mathbf{Z}$ is elliptically contoured. Experience also shows that it is less crucial than the linearity condition; therefore, the condition can be inspected through a scatterplot matrix of the predictors. The predictors are transformed for normality like the linearity condition if the condition is not satisfied.

4.3. Coverage condition

The goal of linearity and constant variance conditions is to induce the relationship of $\mathcal{S}(\mathbf{M}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. This indicates that any of the two or both do not guarantee an exhaustive estimation of $\mathcal{S}_{Y|\mathbf{Z}}$. To have the exhaustive estimation, $\mathcal{S}(\mathbf{M}) = \mathcal{S}_{Y|\mathbf{Z}}$ is assumed to hold, which is called coverage condition. Different from the previous two conditions, it is not possible to investigate that the coverage condition holds in practice. In most SDR methods, either of linearity or constant variance conditions or both are first assumed to hold for guaranteeing that $\mathbf{M}$ spans proper subsets of $\mathcal{S}_{Y|\mathbf{Z}}$. Then the coverage condition is assumed for the exhaustive estimation of $\mathcal{S}_{Y|\mathbf{Z}}$.

5. Discussion

In the paper, we introduce a notion of sufficient dimension reduction in a regression of $Y|\mathbf{X} \in \mathbb{R}^p$. The goal of SDR is to replace the original predictors $\mathbf{X}$ by its lower-dimensional linear projection without loss of information on selected aspects of the conditional distribution $Y|\mathbf{X}$. Depending the aspects, the central subspace, the central mean subspace and the central k^{th}-moment subspace are defined as primary interest. The conditions to guarantee the existence of the three subspaces are discussed since the three subspaces do not always exist. A general estimation approach to estimate them is also introduced, and the conditions commonly assumed in most SDR methodologies are explained.

In a sequence of the second tutorial, SDR methodologies will be introduced to estimate the central subspace, the central mean subspace and the central k^{th}-moment subspace. For the central subspace, methods using the conditional moments of the inverse regression of $\mathbf{X}|Y$ are used such as sliced inverse regression (Li, 1991) and sliced average variance estimation (Cook and Weisberg, 1991). To estimate the central mean subspace, the ordinary least square (Cook, 1998), principal Hessian direction (Li, 1992) and iterative Hessian transformation (Cook and Li, 2002) are popular among many. The central k^{th}-moment subspace is restored by the covariance method proposed by Yin and Cook (2002). Most of the methodologies have the large-sample tests to determine the structural dimensions. However, a permutation test will be studied, and one of its advantages is no requirement large-sample distributions. Real data analysis for the methodologies will be presented to illustrate how to apply the methodologies in practice, and the results will be compared. A seeded dimension reduction approach (Cook *et al.*, 2007) will also be introduced to provide a neat solution of SDR to large p and small n regression.

Acknowledgments

This works was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (NRF-2014R1A2A1A11049389 and 2009-0093827).

The author is specially grateful to Professor Jeong-Soo Park, Editor-in-Chief, Communications for Statistical Applications and Methods, for the invitation of the paper.

References

Bickel P (1982). On adaptive estimation, *Annals of Statistics*, **10**, 647–671.

Cook RD (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*, Wiley, New York.

Cook RD (2000). Save: a method for dimension reduction and graphics in regression, *Communications in Statistics-Theory and Methods*, **29**, 2109–2121.

Cook RD and Li B (2002). Dimension reduction for conditional mean in regression, *Annals of Statistics*, **30**, 455–474.

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

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.

Eaton ML (1986). A characterization of spherical distributions, *Journal of Multivariate Analysis*, **20**, 272–276.

Hall P and Li KC (1993). On almost linearity of low dimensional projections from high dimensional data, *Annals of Statistics*, **21**, 867–889.

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

Li KC (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma, *Journal of the American Statistical Association*, **87**, 1025–1039.

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

Rao CR (1965). *Linear Statistical Inference and Its Application*, Wiley, New York.

Shao Y, Cook RD, and Weisberg S (2006). The linearity condition and adaptive estimation in single-index regressions, Retrieved March 1, 2016, from: http://arxiv.org/pdf/1001.4802v1.pdf

Yin X and Cook RD (2002). Dimension reduction for the conditional kth moment in regression, *Journal of the Royal Statistical Society Series B*, **64**, 159–175.

Received January 22, 2016; Revised February 24, 2016; Accepted February 24, 2016