Communications for Statistical Applications and Methods
2016, Vol. 23, No. 2, 105–117

http://dx.doi.org/10.5351/CSAM.2016.23.2.105
Print ISSN 2287-7843 / Online ISSN 2383-4757

Tutorial: Methodologies for sufficient dimension reduction in regression

Jae Keun Yoo[1,a]

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

Abstract

In the paper, as a sequence of the first tutorial, we discuss sufficient dimension reduction methodologies used to estimate central subspace (sliced inverse regression, sliced average variance estimation), central mean subspace (ordinary least square, principal Hessian direction, iterative Hessian transformation), and central k^{th}-moment subspace (covariance method). Large-sample tests to determine the structural dimensions of the three target subspaces are well derived in most of the methodologies; however, a permutation test (which does not require large-sample distributions) is introduced. The test can be applied to the methodologies discussed in the paper. Theoretical relationships among the sufficient dimension reduction methodologies are also investigated and real data analysis is presented for illustration purposes. A seeded dimension reduction approach is then introduced for the methodologies to apply to large p small n regressions.

Keywords: Hessian matrix, inverse regression, least squares, permutation test, seeded dimension reduction, sufficient dimension reduction

1. Introduction

As a sequence of the first tutorial (Yoo, 2016), the paper starts with introducing its key materials. Sufficient dimension reduction (SDR) in regression of $Y|\mathbf{X} \in \mathbb{R}^p$ is to replace the original predictor $\mathbf{X}$ by its lower-dimensional linear transformed predictor $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ without loss of information on selected aspects of conditional distribution of $Y|\mathbf{X}$, where $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$. Depending on the selected aspects, the central subspace ($\mathcal{S}_{Y|\mathbf{X}}$) (Cook, 1998a), the central mean subspace ($\mathcal{S}_{E(Y|\mathbf{X})}$) (Cook and Li, 2002) or the central k^{th}-moment subspace ($\mathcal{S}^{(k)}_{Y|\mathbf{X}}$) (Yin and Cook, 2002) become of primary interest to recover. Let $\boldsymbol{\eta}$ represent an orthonormal basis matrix of $\mathcal{S}_{Y|\mathbf{X}}$, $\mathcal{S}_{E(Y|\mathbf{X})}$ or $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$, and $\mathcal{S}(\mathbf{B})$ denote a subspace spanned by the columns of $\mathbf{B} \in \mathbb{R}^{p\times q}$. The central subspace is the intersection of all possible $\mathcal{S}(\mathbf{B})$ satisfying $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^{\mathrm{T}}\mathbf{X}$, where $\perp\!\!\!\perp$ stands for independence. Then $\mathbf{X}$ can be replaced by $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ without loss of information on $Y|\mathbf{X}$. The central mean subspace is the intersection of all possible $\mathcal{S}(\mathbf{B})$ such that $Y \perp\!\!\!\perp E(Y|\mathbf{X})|\mathbf{B}^{\mathrm{T}}\mathbf{X}$, and the original predictor $\mathbf{X}$ can be replaced by $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ without loss of information on $E(Y|\mathbf{X})$. The central k^{th}-moment subspace is the intersection of $\mathcal{S}(\mathbf{B})$ to satisfy $Y \perp\!\!\!\perp \{E(Y|\mathbf{X}), M^{(2)}(Y|\mathbf{X}), \ldots, M^{(k)}(Y|\mathbf{X})\}|\mathbf{B}^{\mathrm{T}}\mathbf{X}$, where $M^{(k)}(Y|\mathbf{X}) = E[\{Y - E(Y|\mathbf{X})\}^k|\mathbf{X}]$ and $M^{(1)}$ is replaced by $E(Y|\mathbf{X})$. Then $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ can replace $\mathbf{X}$ without loss of information on the first k conditional moments of $Y|\mathbf{X}$. The conditions to guarantee the existence of $\mathcal{S}_{Y|\mathbf{X}}$ are mild and its existence is not problematic in practice. The conditions for $\mathcal{S}_{E(Y|\mathbf{X})}$ and $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ are also the same as those for $\mathcal{S}_{Y|\mathbf{X}}$.

[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

Let $\mathcal{S}_{\mathbf{X}}$ represent one of $\mathcal{S}_{Y|\mathbf{X}}$, $\mathcal{S}_{E(Y|\mathbf{X})}$ or $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$. For a non-singular matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, 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 $\mathcal{S}_{\mathbf{X}} = \mathbf{A}\mathcal{S}_{\mathbf{Z}}$. Define that $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}\{\mathbf{X} - E(\mathbf{X})\}$, where $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X})$ and $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$. Often, SDR methodologies estimate $\mathcal{S}_{\mathbf{Z}}$, and then back-transformed to $\mathcal{S}_{\mathbf{X}}$ using $\mathcal{S}_{\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{\mathbf{Z}}$.

Denote $\boldsymbol{\eta}_{\mathbf{Z}}$ as an $p \times d$ true orthonormal basis matrix of $\mathcal{S}_{\mathbf{Z}}$. Most SDR methodologies require linearity condition: $E(\mathbf{Z}|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z} = \nu)$ is linear in ν. Some methods additionally require constant variance condition: $\mathrm{cov}(\mathbf{Z}|\boldsymbol{\eta}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z}) = \mathbf{Q}_{\boldsymbol{\eta}_{\mathbf{Z}}}$, where $\mathbf{Q}_{\boldsymbol{\eta}_{\mathbf{Z}}}$ is an orthonormal projection operator onto the orthogonal complement of $\mathcal{S}(\boldsymbol{\eta}_{\mathbf{Z}})$. Along with the linearity and constant variance conditions, coverage condition is assumed to hold to guarantee that SDR methods exhaustively estimate $\mathcal{S}_{\mathbf{X}}$.

This paper introduces SDR methodologies, classical but popularly used, to estimate $\mathcal{S}_{Y|\mathbf{X}}$, $\mathcal{S}_{E(Y|\mathbf{X})}$ and $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$. Two methodologies to restore $\mathcal{S}_{Y|\mathbf{X}}$ on $Y|\mathbf{X}$ are sliced inverse regression and sliced average variance estimation, which use conditional moments of inverse regression of $\mathbf{X}|Y$. Three methodologies of ordinary least squares, principal Hessian direction and iterative Hessian transformation estimate $\mathcal{S}_{E(Y|\mathbf{X})}$. The covariance method is used to recover $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$. The method is ordinary least squares of a regression of a set of polynomials $(Y, Y^2, \ldots, Y^k)$ given $\mathbf{X}$. To provide a neat solution to SDR for large p and small n regression, a seeded dimension reduction approach is presented.

The organization of the paper is as follows. Section 2 is devoted to introducing SDR methodologies to recover $\mathcal{S}_{Y|\mathbf{X}}$. In Section 3, SDR methods to estimate $\mathcal{S}_{E(Y|\mathbf{X})}$ and a permutation test to determine structural dimension is discussed. Section 4 is dedicated to explaining how to estimate $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$. Theoretical relationships are investigated and real data example is presented in Section 5. Seeded dimension reduction is introduced in Section 6, and our work as well as more topics in SDR are discussed in Section 7.

Throughout the rest of the paper, we assume that n iid data observations $\{(\mathbf{X}_i, Y_i, i = 1, \ldots, n)\}$, $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X})$ and $\hat{\boldsymbol{\Sigma}}$ stands for usual moment estimator of $\boldsymbol{\Sigma}$, and define the population and sample standardized predictors of $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ of $\mathbf{X}$ as $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}\{(\mathbf{X}_i - E(\mathbf{X})\}$ and $\hat{\mathbf{Z}} = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$, respectively.

2. Estimation of the central subspace

2.1. Sliced inverse regression

Recall d and $\boldsymbol{\eta}_z$, which are the true dimension and an orthonormal basis matrix for $\mathcal{S}_{Y|\mathbf{Z}}$. To recover $\mathcal{S}_{Y|\mathbf{Z}}$, we consider the inverse regression of $\mathbf{Z}|Y$, especially paying attention on its first moment of $E(\mathbf{Z}|Y)$. Then, $E(\mathbf{Z}|Y)$ has the following equivalences:

$$E(\mathbf{Z}|Y) = E\left\{E\left(\mathbf{Z}|\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}, Y\right)\Big|Y\right\} = E\left\{E\left(\mathbf{Z}|\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}\right)\Big|Y\right\}.$$

Now, linearity condition for $\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}$ is assumed to hold. Then the last equation above is:

$$E(\mathbf{Z}|Y) = E\left\{E\left(\mathbf{Z}|\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}\right)\Big|Y\right\} = E\left(\mathbf{P}_{\boldsymbol{\eta}_z}\mathbf{Z}\Big|Y\right) = \mathbf{P}_{\boldsymbol{\eta}_z}E(\mathbf{Z}|Y).$$

By the last equivalence of $E(\mathbf{Z}|Y) = \mathbf{P}_{\boldsymbol{\eta}_z}E(\mathbf{Z}|Y)$ above, we have $E(\mathbf{Z}|Y) \in \mathcal{S}_{Y|\mathbf{Z}}$, and hence $E(\mathbf{Z}|Y = y)$ with varying y induces a proper subset of $\mathcal{S}_{Y|\mathbf{Z}}$. Summarizing this, if linearity condition holds for $\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}$, the following equivalence can be established

$$\mathcal{S}\{E(\mathbf{Z}|Y)\} \subseteq \mathcal{S}_{Y|\mathbf{Z}} \;\Leftrightarrow\; \boldsymbol{\Sigma}^{-1}\mathcal{S}\{E(\mathbf{X}|Y)\} \subseteq \mathcal{S}_{Y|\mathbf{X}}.$$

Inferring $\mathcal{S}_{Y|\mathbf{X}}$ through constructing $E(\mathbf{X}|Y)$ is called sliced inverse regression (SIR) (Li, 1991).

Now construction of $E(\mathbf{Z}|Y)$ in population matters. It needs to be done without assuming any specific distributional assumption of $Y|\mathbf{Z}$. The construction can be done in this direction by two simple ways. If Y is categorical with h level, the construction of $E(\mathbf{Z}|Y = s), s = 1, \ldots, h$, is straightforward, which is the mean of $\mathbf{Z}$ within each category of Y. If Y is continuous or many-valued, the response is categorize Y with h levels, called *slicing*, to try to have equal numbers of observations, that is, $Y \to \tilde{Y}$. Then compute $E(\mathbf{Z}|\tilde{Y} = s)$ for $s = 1, \ldots, h$. As the kernel matrix to estimate $\mathcal{S}_{Y|\mathbf{Z}}$, the SIR constructs $\mathbf{M}_{SIR} = \text{cov}\{E(\mathbf{Z}|Y)\}$. In sample structure, the algorithm of SIR is:

1. Obtain $\tilde{Y}$ by slicing the range of Y into h non-overlapping intervals. Let n_s be the number of observations for $\tilde{Y} = s$.
2. Compute sample inverse mean with each slice $s = 1, \ldots, h$: $\hat{E}(\mathbf{Z}|\tilde{Y} = s) = (1/n_s) \sum_{\tilde{Y}=s} \hat{\mathbf{Z}}_i$.
3. Construct sample covariance estimator

$$\hat{\mathbf{M}}_{SIR} = \widehat{\text{cov}} \left\{ E\left(\mathbf{Z}|\tilde{Y}\right) \right\} = \sum_{s=1}^{h} \frac{n_s}{n} \hat{E}\left(\mathbf{Z}|\tilde{Y} = s\right) \hat{E}\left(\mathbf{Z}|\tilde{Y} = s\right)^{\mathrm{T}},$$

   where n_s stands for the sample size in the s the slices.
4. Perform spectral decomposition $\hat{\mathbf{M}}_{SIR} = \sum_{i=1}^{p} \hat{\lambda}_j \hat{\gamma}_j \hat{\gamma}_j^{\mathrm{T}}$, where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$.
5. Determine the structural dimension $d = \dim(\mathcal{S}_{Y|\mathbf{Z}})$. Let $\hat{d}$ denote an estimate of d.
6. Form an orthonormal basis estimate $(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})$ for $\mathcal{S}_{Y|\mathbf{Z}}$.
7. Back-transform to obtain a sample basis estimate $\hat{\mathbf{\Sigma}}^{-1/2}(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})$ for $\mathcal{S}_{Y|\mathbf{X}}$.

Test statistics for dimension test and their asymptotic distribution (Bura and Cook, 2001) in SIR is:

$$\hat{\Lambda}_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j \overset{d}{\sim} \sum_{k=1}^{(p-m)(h-m)} \omega_k \chi_k^2(1), \quad m = 0, 1, \ldots, \min(p-1, h-1),$$

where $\chi_k^2(1)$s are independent χ^2 with one degree of freedom and h indicates the total number of slices. The weights ω_k can be estimated consistently for use in practice. In practice, the results from SIR are sensitive in the choices of the number of slices, and Cook and Zhang (2014) recently developed a version of SIR, not affected by this.

SIR fails in a regression such that $Y = X_1^2 + \varepsilon$. To overcome this, we need to investigate the second conditional moments of $\mathbf{Z}|Y$.

2.2. Sliced average variance estimation

Under linearity and constant variance conditions for $\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}$, the following relationship for $\text{cov}(\mathbf{Z}|Y)$ (Cook and Weisberg, 1991) has been shown that

$$\mathbf{I}_p - \text{cov}(\mathbf{Z}|Y) = \mathbf{P}_{\boldsymbol{\eta}_\mathbf{Z}} \left\{ \mathbf{I}_p - \text{cov}(\mathbf{Z}|Y) \right\} \mathbf{P}_{\boldsymbol{\eta}_\mathbf{Z}} \in \mathcal{S}_{Y|\mathbf{Z}}.$$

Therefore, a subspace spanned by $\mathbf{I}_p - \text{cov}(\mathbf{Z}|Y = y)$ with varying y results in a proper subset of $\mathcal{S}_{Y|\mathbf{Z}}$, so we have

$$\mathcal{S}\left\{ \mathbf{I}_p - \text{cov}(\mathbf{Z}|Y) \right\} = \mathcal{S}\left[E\left\{ \mathbf{I}_p - \text{cov}(\mathbf{Z}|Y) \right\}^2 \right] \subseteq \mathcal{S}_{Y|\mathbf{Z}}.$$

Restoration of $\mathcal{S}_{Y|\mathbf{Z}}$ through $\mathbf{M}_{SAVE} = E\{\mathbf{I}_p - \text{cov}(\mathbf{Z}|Y)\}^2$ is called sliced average variance estimation (SAVE) (Cook and Weisberg, 1991).

Construction of $\text{cov}(\mathbf{Z}|Y)$ is the same as that of $E(\mathbf{Z}|Y)$ in population. If Y is categorical, the computation of $\text{cov}(\mathbf{Z}|Y)$ is straightforward, otherwise it is computed within each slice after slicing Y. The algorithm of SAVE is similar to that of SIR. First obtain $\tilde{Y}$ by slicing Y with h levels. Then compute sample inverse covariance with each slice:

$$\widehat{\text{cov}}\left(\mathbf{Z}|\tilde{Y}=s\right) = \frac{1}{n_s}\sum_{\tilde{Y}_i \in \text{slice } s}\left(\hat{\mathbf{Z}}_{i\in s} - \bar{\mathbf{Z}}_s\right)\left(\hat{\mathbf{Z}}_{i\in s} - \bar{\mathbf{Z}}_s\right)^{\mathrm{T}},$$

where $\bar{\mathbf{Z}}_s = (1/n_s)\sum_{\tilde{Y}=s}\hat{\mathbf{Z}}_i$. Then construct a sample kernel matrix:

$$\hat{\mathbf{M}}_{SAVE} = \sum_{s=1}^{h}\frac{n_s}{n}\left\{\mathbf{I}_p - \widehat{\text{cov}}\left(\mathbf{Z}|\tilde{Y}=s\right)\right\}\left\{\mathbf{I}_p - \widehat{\text{cov}}\left(\mathbf{Z}|\tilde{Y}=s\right)\right\}.$$

To induce test statistics and their asymptotic distributions, $\mathbf{M}_{SAVE}$ is re-written such that $\mathbf{M}_{SAVE} = \sum_{s=1}^{h}\mathbf{A}_s^2$, where $\mathbf{A}_s = f_s^{1/2}(\mathbf{\Sigma}_s - \mathbf{I}_p)$ and $f_s = n_s/n$. Let $\hat{\theta}_{p-m} = (\hat{\gamma}_{m+1}, \ldots, \hat{\gamma}_p)$, where $(\hat{\gamma}_{m+1}, \ldots, \hat{\gamma}_p)$ are the eigenvectors corresponding to the last $(p-m)$ smallest eigenvalues of $\hat{\mathbf{M}}_{SAVE}$. Test statistics and their asymptotic distributions (Shao *et al.*, 2007) under $\mathrm{H}_0 : d = m$ are then:

$$T_n\left(\hat{\theta}_{p-m}\right) = n\sum_{k=1}^{h}\text{tr}\left\{\left(\hat{\theta}_{p-m}^{\mathrm{T}}\hat{\mathbf{A}}_k\hat{\theta}_{p-m}\right)^2\right\}, \quad m = 0, \ldots, p.$$

Since covariance matrices are computed in SAVE, relatively larger sample sizes are recommended per slice than SIR to have good estimation of $\mathcal{S}_{Y|\mathbf{X}}$.

3. Estimation of the central mean subspace

3.1. Ordinary least squares

General definition of the ordinary least squares of $Y|\mathbf{X}$ and $Y|\mathbf{Z}$ is that $\beta = \mathbf{\Sigma}^{-1}\,\text{cov}(\mathbf{X}, Y)$ and $\beta_z = E(Y\mathbf{Z})$, respectively. According to Cook (1998a), under the linearity condition for $\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}$, it can be shown that

$$\beta_z \in \mathcal{S}_{E(Y|\mathbf{Z})}.$$

Good parts to use the OLS are as follows. First, many properties of the OLS (including its asymptotics) are well derived. As long as the linearity holds, we can find at least one column in basis matrices to span $\mathcal{S}_{E(Y|\mathbf{Z})}$. Any associated model may not be true; however, it provides an adequate fit of the data with respect to $\mathcal{S}_{E(Y|\mathbf{Z})}$. Also, the relationship $\beta_z \in \mathcal{S}_{E(Y|\mathbf{Z})}$ holds even when Y is categorical. However, the OLS is not informative to $\mathcal{S}_{E(Y|\mathbf{Z})}$, for example symmetric regressions like SIR. In addition, the OLS provides only one column, which should be a crucial deficit in the OLS.

3.2. Principal Hessian direction

Let the columns of $\boldsymbol{\eta}_z$ span $\mathcal{S}_{E(Y|\mathbf{Z})}$. For $\boldsymbol{\eta}_z$, we consider the $p \times p$ Hessian matrix of the regression function

$$H(\mathbf{Z}) = \frac{\partial^2 E(Y|\mathbf{Z})}{\partial\mathbf{Z}\partial\mathbf{Z}^{\mathrm{T}}} = \boldsymbol{\eta}_z^{\mathrm{T}}\frac{\partial E\left(Y|\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}\right)}{\partial\left(\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}\right)\partial\left(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\eta}_z\right)}\boldsymbol{\eta}_z.$$

Therefore, we have $\mathcal{S}[E\{H(\mathbf{Z})\}] \subseteq \mathcal{S}_{E(Y|\mathbf{Z})}$. The problem is that the form of $E(Y|\mathbf{Z})$ is unknown. To overcome this problem, define that $\boldsymbol{\Sigma}_{yzz} = E[\{Y - E(Y)\}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}]$. The following result summarizes the relationship between $\boldsymbol{\Sigma}_{yzz}$ and $H(\mathbf{Z})$.

Result 1. *Stein (1981)'s Lemma under the normality of* $\mathbf{Z}$ *establishes that* $\mathcal{S}(\boldsymbol{\Sigma}_{yzz}) = \mathcal{S}[E\{H(\mathbf{Z})\}]$.

The quantity of $H(\mathbf{Z})$ is not estimable because of unknown $E(Y|\mathbf{Z})$, but Result 1 enables us equivalently to replace $H(\mathbf{Z})$ by $\boldsymbol{\Sigma}_{yzz}$, which can be estimated from data. This approach to recover $\mathcal{S}_{E(Y|\mathbf{Z})}$ through $\boldsymbol{\Sigma}_{yzz}$ is called principal Hessian directions (pHd) (Li, 1992).

It is noted that the original founding of pHd by Li (1992) is two steps. The first step is to show that $H(\mathbf{Z})$ is informative to $\mathcal{S}_{E(Y|\mathbf{Z})}$. The second is the construction of $\boldsymbol{\Sigma}_{yzz}$ to replace $H(\mathbf{Z})$. For the second step, under the normality of $\mathbf{Z}$ To induce the relationship in the second step, Stein's Lemma should be used and is why the normality of $\mathbf{Z}$ is required. Cook (1998b) directly shows the following result without requiring the normality of $\mathbf{Z}$.

Result 2. *Assume that the linearity and constant variance conditions hold for* $\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}$. *Then, we have* $\mathcal{S}(\boldsymbol{\Sigma}_{yzz}) \subseteq \mathcal{S}_{E(Y|\mathbf{Z})}$.

Result 2 gives the direct relationship between $\mathcal{S}(\boldsymbol{\Sigma}_{yzz})$ and $\mathcal{S}_{E(Y|\mathbf{Z})}$ under the linearity and constant variance conditions, which are weaker than the normality of $\mathbf{Z}$. So, Result 2 enables pHd to estimate $\mathcal{S}_{E(Y|\mathbf{Z})}$ under weaker condition, and hence its applicability is in practice enhanced. In addition, according to Cook (1998b), the inference procedure associated with pHd can be greatly simplified, if $\mathrm{cov}(\mathbf{Z}, Y) = 0$. For this, Cook (1998b) suggests to replace Y with the OLS residual $r = Y - E(Y) - \beta_z^{\mathrm{T}}\mathbf{Z}$ so that $\mathrm{cov}(\mathbf{Z}, r) = 0$. Once the kernel matrix $\boldsymbol{\Sigma}_{rzz} = E(r\mathbf{Z}\mathbf{Z}^{\mathrm{T}})$ is constructed, $\boldsymbol{\Sigma}_{rzz}$ estimates $\mathcal{S}_{E(r|\mathbf{Z})}$. Then $\mathcal{S}_{E(Y|\mathbf{Z})}$ is restored through with $\mathcal{S}_{E(r|\mathbf{Z})}$ incorporating β_z. The linearity condition forces that

$$\mathcal{S}_{E(Y|\mathbf{Z})} = \mathcal{S}_{E(r|\mathbf{Z})} + \mathcal{S}(\boldsymbol{\beta}_z),$$

where $\mathcal{S}(\boldsymbol{\beta}_z) \subset \mathcal{S}_{E(r|\mathbf{Z})}$ or $\mathcal{S}(\boldsymbol{\beta}_z) \cap \mathcal{S}_{E(r|\mathbf{Z})} = 0$. Assuming that the linearity and constant variance condition hold, the discussion above can be summarized as:

$$\mathcal{S}(\boldsymbol{\Sigma}_{yzz}) \subseteq \mathcal{S}_{E(Y|\mathbf{Z})}; \quad \mathcal{S}(\boldsymbol{\Sigma}_{rzz}) \subseteq \mathcal{S}_{E(r|\mathbf{Z})}; \quad \mathcal{S}_{E(Y|\mathbf{Z})} = \mathcal{S}_{E(r|\mathbf{Z})} + \mathcal{S}(\boldsymbol{\beta}_z); \quad \mathcal{S}(\boldsymbol{\Sigma}_{rzz}) + \mathcal{S}(\boldsymbol{\beta}_z) \subseteq \mathcal{S}_{E(Y|\mathbf{Z})}.$$

The performance of $\{E(r\mathbf{Z}\mathbf{Z}^{\mathrm{T}}), \beta_Z\}$ often turns out to be better than $E(y\mathbf{Z}\mathbf{Z}^{\mathrm{T}})$.

While pHd is known as a method to estimate $\mathcal{S}_{E(Y|\mathbf{X})}$, Li (1992) and Cook (1998b) did not originally pay attention to that. Li (1992) proposed pHd to capture information on $\mathcal{S}_{Y|\mathbf{X}}$, not $\mathcal{S}_{E(Y|\mathbf{X})}$, and Cook (1998b) extended applicability of pHd by replacing the normality of $\mathbf{Z}$ by linearity and constant variance conditions and developing asymptotics of test statistics. A few year later, Cook and Li (2002) introduced the central mean subspace and showed that the kernel matrix constructed by pHd technically spans a subspace of $\mathcal{S}_{E(Y|\mathbf{X})}$, not $\mathcal{S}_{Y|\mathbf{X}}$.

The kernel matrix of r-based pHd is $\mathbf{M}_{rpHd} = E(y\mathbf{Z}\mathbf{Z}^{\mathrm{T}})E(y\mathbf{Z}\mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$, and the sample version of $E(y\mathbf{Z}\mathbf{Z}^{\mathrm{T}})$ and $\mathbf{M}_{rpHd}$ are constructed as:

$$\hat{\boldsymbol{\Sigma}}_{rzz} = \frac{1}{n}\sum_{i}^{n} r_i\hat{\mathbf{Z}}_i\hat{\mathbf{Z}}_i^{\mathrm{T}} \quad \text{and} \quad \hat{\mathbf{M}}_{rpHd} = \hat{\boldsymbol{\Sigma}}_{rzz}\hat{\boldsymbol{\Sigma}}_{rzz}^{\mathrm{T}},$$

where $\hat{r}_i = Y_i - \bar{Y} - \hat{\beta}_z^{\mathrm{T}}\hat{\mathbf{Z}}_i$. For the dimension determination, $\mathrm{H}_0 : d = m$ is sequentially tested with a statistic $n\sum_{j=m+1}^{p}\hat{\lambda}_j/2\hat{\mathrm{var}}(\hat{r})$:

$$\frac{n\sum_{j=m+1}^{p}\hat{\lambda}_j}{2\hat{\mathrm{var}}(\hat{r})} \sim \frac{1}{2\,\mathrm{var}(r)}\sum_{j=1}^{(p-m)(p-m+1)/2}\omega_j\chi_j^2(1),$$

where $\chi^2_j(1)$s stands for independent χ^2s with 1 degree of freedom.

3.3. Iterative Hessian transformation

The following result directly comes from Cook and Li (2002).

Result 3. *Assume the linearity condition for $\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}$. Suppose that U and V are measurable functions of $\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}$. $E\{(UY+V)\mathbf{Z}\} \in \mathcal{S}_{E(Y|\mathbf{Z})}$, provided that $(UY+V)\mathbf{Z}$ is integrable.*

Letting $Y^* = UY + V$, $E\{(UY+V)\mathbf{Z}\}$ is nothing but OLS from $Y^*|\mathbf{Z}$. This implies that $E\{(UY+V)\mathbf{Z}\} \in \mathcal{S}_{E(Y^*|\mathbf{Z})}$, but we can not always expect that $E\{(UY+V)\mathbf{Z}\} \in \mathcal{S}_{E(Y|\mathbf{Z})}$. The conditions on U and V in Result 3 forces that $E(Y^*\mathbf{Z}) \in \mathcal{S}_{E(Y|\mathbf{Z})}$.

Suppose we can find a vector $\delta_1 \in \mathcal{S}_{E(Y|\mathbf{Z})}$. We can then choose appropriate functions $U : R \mapsto R$ and $V : R \mapsto R$, construct a new variable $Y_1^* = U(\delta_1^{\mathrm{T}}\mathbf{Z})Y + V(\delta_1^{\mathrm{T}}\mathbf{Z})$, and obtain the covariance vector $E(Y_1^*\mathbf{Z})$. By Result 3, we have $E(Y_1^*\mathbf{Z}) \in \mathcal{S}_{E(Y|\mathbf{Z})}$. Let $\delta_2 = E(Y_1^*\mathbf{Z})$ and suppose that $\delta_1 \neq E(Y_1^*\mathbf{Z})$. Then we can apply the same U and V to δ_2, and construct $\delta_3 = E(Y_2^*\mathbf{Z})$, where $Y_2^* = U(\delta_2^{\mathrm{T}}\mathbf{Z})Y + V(\delta_2^{\mathrm{T}}\mathbf{Z})$. If we iterate this successive procedure many times, then many δ_is can be constructed enough to have information on $\mathcal{S}_{E(Y|\mathbf{Z})}$. We can therefore start with δ_1 and apply this successive process in the hope of finding additional predictor vectors in $\mathcal{S}_{E(Y|\mathbf{Z})}$.

We know that $\beta_z \in \mathcal{S}_{E(Y|\mathbf{Z})}$ under linearity condition. Letting $\delta_1 = \beta_z$, $U(t) = t$, and $V(t) = -tE(Y)$, we have $Y_1^* = \beta_z^{\mathrm{T}}\mathbf{Z}Y - \beta_z^{\mathrm{T}}\mathbf{Z}E(\mathbf{Y}) = \beta_z^{\mathrm{T}}\mathbf{Z}\{Y - E(Y)\}$, and we can obtain

$$\begin{aligned}
\delta_2 &= E[\beta_z^{\mathrm{T}}\mathbf{Z}\{Y - E(Y)\}\mathbf{Z}] = E[\{Y - E(Y)\}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\beta_z] = \boldsymbol{\Sigma}_{yzz}\beta_z \\
\delta_3 &= E[\{Y - E(Y)\}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\delta_2] = \boldsymbol{\Sigma}_{yzz}\boldsymbol{\Sigma}_{yzz}\beta_z = \boldsymbol{\Sigma}_{yzz}^2\beta_z \\
&\vdots \\
\delta_p &= \boldsymbol{\Sigma}_{yzz}^{(p-1)}\beta_z.
\end{aligned}$$

Consequently, $\mathcal{S}[\mathbf{M}_{IHTy} = \{\beta_z, \boldsymbol{\Sigma}_{yzz}\beta_z, \ldots, \boldsymbol{\Sigma}_{yzz}^{(p-1)}\beta_z\}] \subseteq \mathcal{S}_{E(Y|\mathbf{Z})}$.

As another case, letting $U(t) = t$ and $V(t) = -tE(Y) - t^2$ with $\delta_1 = \beta_z$, we have $Y_1^* = \beta_z^{\mathrm{T}}\mathbf{Z}Y - \beta_z^{\mathrm{T}}\mathbf{Z}E(\mathbf{Y}) - (\beta_z^{\mathrm{T}}\mathbf{Z})^2 = \beta_z^{\mathrm{T}}\mathbf{Z}\{Y - E(Y) - \beta_z^{\mathrm{T}}\mathbf{Z}\} = \beta_z^{\mathrm{T}}\mathbf{Z}r$. Then, the following quantities directly come by Result 3:

$$\delta_2 = E\left[r\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\beta_z\right] = \boldsymbol{\Sigma}_{rzz}\beta_z; \quad \delta_3 = E\left[r\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\delta_2\right] = \boldsymbol{\Sigma}_{rzz}\boldsymbol{\Sigma}_{rzz}\beta_z = \boldsymbol{\Sigma}_{rzz}^2\beta_z; \quad \cdots \quad ; \delta_p = \boldsymbol{\Sigma}_{rzz}^{(p-1)}\beta_z.$$

Consequently, $\mathcal{S}[\mathbf{M}_{IHTr} = \{\beta_z, \boldsymbol{\Sigma}_{rzz}\beta_z, \ldots, \boldsymbol{\Sigma}_{rzz}^{(p-1)}\beta_z\}] \subseteq \mathcal{S}_{E(Y|\mathbf{Z})}$. The number of iteration should be enough to $(p-1)$ times, because $\mathbf{M}_{IHTy}$ or $\mathbf{M}_{IHTr}$ will expect have the full rank. The approach to recover $\mathcal{S}_{E(Y|\mathbf{X})}$ through $\mathbf{M}_{IHTy}$ or $\mathbf{M}_{IHTr}$ is called iterative Hessian direction (IHT).

The IHT is a collection of the OLSs acquired from transformed response variables. The linearity condition only is required in IHT, despite using information of $\boldsymbol{\Sigma}_{yzz}$ and $\boldsymbol{\Sigma}_{rzz}$. Here, we start with the OLS, but the IHT can be thought of as a successive approach to project any initial estimate $\delta_1 \in \mathcal{S}_{E(Y|\mathbf{Z})}$ to some proper kernel matrices. Under the linearity condition, it is clear that $\mathbf{M}_{IHTy}$ and $\mathbf{M}_{IHTr}$ have more informative than β_z. For IHT to be practically useful, the regression should have linear trend, because $\beta_z = 0$ results in $\mathbf{M}_{IHTy} = \mathbf{M}_{IHTr} = 0$.

3.4. Permutation tests

For IHT, the sum of eigenvalues $\hat{\Lambda}_m = n\sum_{i=m+1}^{p}\hat{\lambda}_i$ of $\hat{\mathbf{M}}_{IHTr}$ is proposed as statistics for dimension determination like the others. Cook and Li (2004) derive the asymptotics of the test statistics; however,

a permutation test for the dimension determination is employed here.

The permutation test in SDR is easily implemented and does not require asymptotics. The test algorithm is applicable to most SDR methods (such as the methods introduced in this paper) and are quite common in SDR literature. Before developing the asymptotics of SAVE, its dimension determination has been done with the permutation test. However, it takes longer time than tests done using asymptotics. The test also requires an additional condition of

$$\left(Y, \boldsymbol{\Gamma}_1^{\mathrm{T}}\mathbf{Z}\right) \perp\!\!\!\perp \boldsymbol{\Gamma}_2^{\mathrm{T}}\mathbf{Z},$$

where $\boldsymbol{\Gamma}_1 = (\gamma_1, \ldots, \gamma_m)$ is a set of eigenvectors corresponding to the m largest eigenvalues of $\mathbf{M}_{IHTr}$ under $\mathrm{H}_0 : d = m$, and $\boldsymbol{\Gamma}_2 \in \mathbb{R}^{p\times(p-m)}$ is the orthogonal complement of $\boldsymbol{\Gamma}_1$ with $\boldsymbol{\Gamma}_2^{\mathrm{T}}\boldsymbol{\Gamma}_1 = 0$.

The permutation test algorithm is:

1. Compute $\hat{\mathbf{M}}_{IHTr}$. Under $\mathrm{H}_0 : d = m$, obtain $\hat{\Lambda}_m$ and partition eigenvector matrices
$$\hat{\boldsymbol{\Gamma}}_1 = \left(\hat{\gamma}_1, \ldots, \hat{\gamma}_m\right) \quad \text{and} \quad \hat{\boldsymbol{\Gamma}}_2 = \left(\hat{\gamma}_{m+1}, \ldots, \hat{\gamma}_p\right).$$
2. Construct two vectors of $\hat{V}_i \in \mathbb{R}^{m\times 1} = \hat{\boldsymbol{\Gamma}}_1^{\mathrm{T}}\hat{\mathbf{Z}}_i$ and $\hat{U}_i \in \mathbb{R}^{(p-m)\times 1} = \hat{\boldsymbol{\Gamma}}_2^{\mathrm{T}}\hat{\mathbf{Z}}_i$, for $i = 1, \ldots, n$.
3. Randomly permute the indices i of the $\hat{U}_i$ to obtain the permuted set $\hat{U}_i^*$.
4. Construct the test statistic $\hat{\Lambda}_m^*$ based on a regression of $Y_i|(\hat{V}_i\ ,\ \hat{U}_i^*)$.
5. Repeat Steps (3)–(4) N times, where N is the total number of permutations. The p-value of the hypothesis testing is the fraction of $\hat{\Lambda}_m^*$ that exceed $\hat{\Lambda}_m$.

For $m = 0, \ldots, p$, the dimension test is sequentially evaluated, and is determined as the value in H_0 not to be rejected for the first time. In the permutation test, it should be aware the p-values may not increase as m gets larger. The permutation test for the other SDR methods can be simply done by replacing $\hat{\mathbf{M}}_{IHTr}$ by the sample kernel matrices of SIR, SAVE and pHd.

4. Estimation of the central k^{th}-moment subspace

Let the columns of $\boldsymbol{\eta}_z$ span $\mathcal{S}_{Y|\mathbf{Z}}^{(k)}$. Yin and Cook (2002) show the following result.

Result 4. *Under the linearity condition for* $\boldsymbol{\eta}_z^{\mathrm{T}}\mathbf{Z}$, $E(\mathbf{Z}Y^\ell) \in \mathcal{S}_{Y|\mathbf{Z}}^{(\ell)},\ \ell = 1, \ldots, k.$

Letting $\mathbf{M}_{\mathrm{cov}_k,y} = \{E(\mathbf{Z}Y), E(\mathbf{Z}Y^2), \ldots, E(\mathbf{Z}Y^k)\}$, Result 4 directly implies that $\mathcal{S}(\mathbf{M}_{\mathrm{cov}_k,y}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}^{(k)}$. As easily noticed, $E(\mathbf{Z}Y^k)$ is nothing but the OLS of $Y^k|\mathbf{Z}$, which is equal to covariance of $\mathbf{Z}$ and Y^k. Higher orders of Y^k often result in numerical instability. For this, Y is standardized to $W = \{Y - E(Y)\}/\sqrt{\mathrm{var}(Y)}$. Then $\mathbf{M}_{\mathrm{cov}_k} = \{E(\mathbf{Z}W), E(\mathbf{Z}W^2), \ldots, E(\mathbf{Z}W^k)\}$ is constructed in practice as a kernel matrix to estimate $\mathcal{S}_{Y|\mathbf{Z}}^{(k)}$. The approach to recover $\mathcal{S}_{Y|\mathbf{Z}}^{(k)}$ through $\mathbf{M}_{\mathrm{cov}_k}$ is called covariance method (cov_k) (Yin and Cook, 2002).

Yoo (2013b) recently develop a large sample test for the dimension determination; however, the permutation test is popularly used in the case. In Yoo (2013b), a $p \times k$ matrix of $\boldsymbol{\beta} = \boldsymbol{\Theta}\boldsymbol{\Sigma}_{y,k}^{-1}$ is newly considered, and it is shown that $\mathcal{S}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1}\mathcal{S}(\mathbf{M}_{\mathrm{cov}_k,y})$, where $\boldsymbol{\Theta}_k = \mathrm{cov}\{\mathbf{X}, (Y, Y^2, \ldots, Y^k)^{\mathrm{T}}\}$ and $\boldsymbol{\Sigma}_{y,k} = \mathrm{cov}(Y, Y^2, \ldots, Y^k)$. Let $\hat{\boldsymbol{\beta}}$ be the sample version of $\boldsymbol{\beta}$ by replacing the population quantities with their usual moment estimator. Then, Yoo (2013b) shows that $\sqrt{n}\{\mathrm{vec}(\hat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\beta})\}$ converges to normal distribution, where $\mathrm{vec}(\mathbf{A})$ indicates a $pq \times 1$ vector constructed by stacking the columns of a $p \times q$ matrix $\mathbf{A} = (a_1, \ldots, a_q)$: $\mathrm{vec}(\mathbf{A}) = (a_1^{\mathrm{T}}, \ldots, a_q^{\mathrm{T}})^{\mathrm{T}}$.

5. Theoretical relationships between the methodologies and real data example

5.1. Theoretical relationships between the methodologies

According to Ye and Weiss (2003), we have the following relationship of

$$\mathcal{S}(\mathbf{M}_{SIR}) \subseteq \mathcal{S}(\mathbf{M}_{SAVE}).$$

According to Cook and Critchley (2000), especially when Y is categorical,

$$\mathcal{S}(\mathbf{M}_{SAVE}) = \mathcal{S}(\mathbf{M}_{SIR}) \oplus \mathcal{S}_{\Delta_{\mathbf{Z}|Y}},$$

where $\mathcal{S}_{\Delta_{\mathbf{Z}|Y}} = \mathcal{S}\{\mathrm{cov}(\mathbf{Z}|Y = s+1) - \mathrm{cov}(\mathbf{Z}|Y = s)\}$, $s = 1, \ldots, h-1$.

Because $E(\mathbf{Z}Y)$ can be expressed as an average of vectors in $E(\mathbf{Z}|Y)$, we have

$$\beta_z \in \mathcal{S}(\mathbf{M}_{SIR}).$$

According to Yin and Cook (2002), (1) if Y has finite support $R(Y) = \{a_0, a_1, \ldots, a_k\}$, then

$$\mathcal{S}(\mathbf{M}_{\mathrm{cov}_k}) = \mathcal{S}(\mathbf{M}_{SIR}),$$

and (2) if Y is continuous and $\mu_Y = E(\mathbf{Z}|Y)$ is continuous on $R(Y)$, then

$$\lim_{k \to \mathrm{inf}} \mathcal{S}(\mathbf{M}_{\mathrm{cov}_k}) = \mathcal{S}(\mathbf{M}_{SIR}).$$

This may indicate the role of k. For any value of k, $\mathcal{S}(\mathbf{M}_{\mathrm{cov}_k}) \subseteq \mathcal{S}_{E(\mathbf{Z}|Y)}$ and thus cov_k provide lower bounds on $\mathcal{S}_{E(\mathbf{Z}|Y)}$. We may wish to conduct analyses with different numbers of slices, first focusing on low order covariances (few slices) and then on high order covariances (many slices) to gain a more detailed understanding of $Y|\mathbf{Z}$. And, define $\theta = \mathbf{\Sigma}^{-1}E\{t(Y)\mathbf{X}\}$, where $t(Y)$ denotes a function of Y with $E\{t(Y)\} = 0$. Clearly, the linearity condition forces that $\theta \in \mathcal{S}_{Y|\mathbf{X}}$. The SIR takes $t(Y) = J_s(Y)$, where $J_s(Y)$ stands for slice indicators, while cov_k takes $t(Y) = Y^s$, $s = 1, \ldots, k$.

Often SIR and OLS work better than SAVE and pHd, when there exist linear trend in regression. However, SAVE and pHd outperform SIR and OLS with non-linear trend in regression.

Ye and Weiss (2003) show that

$$\mathcal{S}(\omega\mathbf{M}_\bullet + (1-\omega)\mathbf{M}_\dagger) \subseteq \mathcal{S}_{Y|\mathbf{X}},$$

where $0 \geq \omega \geq 1$ and $\mathbf{M}_\bullet$ and $\mathbf{M}_\dagger$ are two choices among $\mathbf{M}_{SIR}$, $\mathbf{M}_{SAVE}$, β_z and $\mathbf{M}_{rpHd}$. This implies that one can improve the estimation of $\mathcal{S}_{Y|\mathbf{X}}$ by combining two SDR methods to compensate the deficit of each other such as $\omega\mathbf{M}_{SIR} + (1-\omega)\mathbf{M}_{SAVE}$ or $\omega\mathbf{M}_{SIR} + (1-\omega)\mathbf{M}_{rpHd}$.

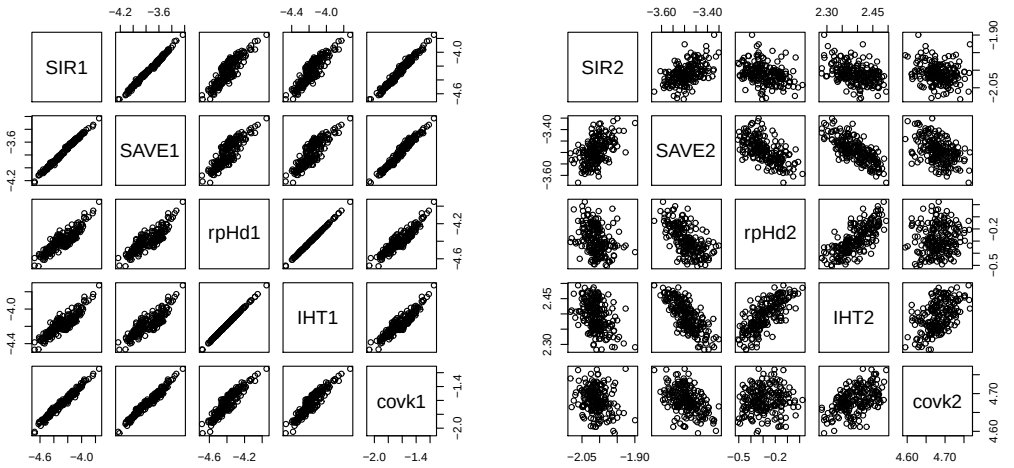
5.2. Real data example

For illustration purposes, Australian Institute of Sport (AIS) data is considered. Data was collected for investigating the relationship between body fat and various predictors with the goal of identifying overweight individuals and understanding factors that may be associated with this condition. We considered lean body mass (LBM) as a response variable and the following eight variables as predictors: (1) height (cm, Ht); (2) weight (kg, Wt); (3) red cell count (RCC); (4) white cell count (WCC); (5) hematocrit (Hc); (6) hemoglobin (Hg); (7) plasma ferritin concentration (Ferr); (8) sum of skinfolds (SSF).

Table 1: Dimension test results for AIS data; rpHd, residual-based pHd; IHT(2), IHT with two iterations

	SIR	SAVE	rpHd	IHT(2)	cov_2
$H_0 : d = 0$	0.00	0.00	0.01	0.00	0.00
$H_0 : d = 1$	0.45	0.04	0.01	0.11	0.21
$H_0 : d = 2$	N/A	0.12	0.23	0.96	N/A

AIS = Australian Institute of Sport, SIR = sliced inverse regression, SAVE = sliced average variance estimation, rpHd = residual-based principal Hessian directions, IHT(2) = iterative Hessian direction with two iterations.



(a) First sufficient predictors (b) Second sufficient predictors

Figure 1: *Scatterplot matrices among the sufficient predictors.*

To reduce the dimension of the predictors, the methods of SIR, SAVE, and pHd are implemented with dr-package in R. AIS data can also be obtained from the same package. For IHT and cov_k with permutation tests, readers can find codes at http://home.ewha.ac.kr/~yjkstat/IHT.txt and http://home.ewha.ac.kr/~yjkstat/covk.txt, respectively.

All predictors were first transformed to log-scale to guarantee that linearity and constant variance conditions hold. The dimension test results are summarized in the following table. For SIR and SAVE, 3 slices were used. According to Table 1, with level 5%, SIR, IHT and the covariance method determine that the structural dimension is equal to 1, while SAVE and rpHd do to 2. To obtain more information before the decision, the scatterplot matrices among the first two sufficient predictors from each method are constructed. Figure 1 shows that all the first sufficient predictors have strong linear relationships with each other, while the second predictors do not. Especially, the second sufficient predictor from cov_k does not have strong relationship with all the other second sufficient predictors. This indicates that the second one is spurious; therefore, it is best concluded that the structural dimension should be one. Marginally standardizing each of the remaining predictors to have a sample standard deviation of 1, the analysis might now be continued by plotting LBM against the estimated sufficient predictor from SIR

$$\begin{aligned}\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X} = {} & -0.07\log(\text{SSF}) + 0.15\log(\text{Wt}) + 0.03\log(\text{Hg}) + 0.02\log(\text{Ht}) \\ & + 0.01\log(\text{WCC}) + 0.02\log(\text{RCC}) - 0.03\log(\text{Hc}) + 0.01\log(\text{Ferr}).\end{aligned}$$

6. Seeded dimension reduction

The SDR methods introduced in the paper commonly require the inversion of $\boldsymbol{\Sigma}$. For a regression with $n < p$, so called large p small n regression, the SDR methods cannot be applied in practice, because $\hat{\boldsymbol{\Sigma}}$ is not invertible. To overcome this issue, Cook *et al.* (2007) proposed a paradigm of sufficient dimension reduction without requiring the inversion of $\hat{\boldsymbol{\Sigma}}$. For this, a seed matrix $\boldsymbol{\nu} \in \mathbb{R}^{p\times q}$ is defined such that

$$\boldsymbol{\Sigma}^{-1}\mathcal{S}(\boldsymbol{\nu}) \subseteq \mathcal{S}_{Y|\mathbf{X}} \Leftrightarrow \mathcal{S}(\boldsymbol{\nu}) \subseteq \boldsymbol{\Sigma}\mathcal{S}_{Y|\mathbf{X}}.$$

The seed matrix $\boldsymbol{\nu}$ is required to be constructed without inverting $\boldsymbol{\Sigma}$. For SIR, OLS, and cov_k, the kernel matrices are:

$$\begin{aligned}
&\text{SIR} : \boldsymbol{\Sigma}^{-1}\{E(\mathbf{X}|Y) - E(\mathbf{X})\} \in \mathcal{S}_{Y|\mathbf{X}} \Leftrightarrow \mathcal{S}\{E(\mathbf{X}|Y) - E(\mathbf{X})\} \subseteq \boldsymbol{\Sigma}\mathcal{S}_{Y|\mathbf{X}};\\
&\text{OLS} : \boldsymbol{\Sigma}^{-1}\,\text{cov}(\mathbf{X}, Y) \in \mathcal{S}_{Y|\mathbf{X}} \Leftrightarrow \mathcal{S}\{\text{cov}(\mathbf{X}, Y)\} \subseteq \boldsymbol{\Sigma}\mathcal{S}_{Y|\mathbf{X}};\\
&\text{cov}_k : \boldsymbol{\Sigma}^{-1}\,\text{cov}\{\mathbf{X}, W(k)\} \in \mathcal{S}_{Y|\mathbf{X}}(\text{cov}_k) \Leftrightarrow \mathcal{S}\{\text{cov}(\mathbf{X}, W(k))\} \subseteq \boldsymbol{\Sigma}\mathcal{S}_{Y|\mathbf{X}}, \text{ where } W(k) = \left(W, U^2, \ldots, W^k\right).
\end{aligned}$$

Note the quantities of $E(\mathbf{X}|Y) - E(\mathbf{X})$, $\text{cov}(\mathbf{X}, Y)$ and $\text{cov}(\mathbf{X}, W(k))$ are well known to how to construct them from data in practice, and it can be done without the inversion of $\boldsymbol{\Sigma}$. Therefore, the three are good choices as candidates for $\boldsymbol{\nu}$, and will be used as $\boldsymbol{\nu}$ throughout the rest of paper, if not specifically mentioning a new candidate for $\boldsymbol{\nu}$. For simplicity, we will assume that $\mathcal{S}(\boldsymbol{\nu}) = \boldsymbol{\Sigma}\mathcal{S}_{Y|\mathbf{X}}$.

Assume that a subspace $\mathcal{M}_{Y|\mathbf{X}}$ of $\mathbb{R}^p$ such that $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{M}_{Y|\mathbf{X}}$ is known. It is obvious that

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu} \in \mathcal{M}_{Y|\mathbf{X}} \Leftrightarrow \boldsymbol{\Sigma}^{-1}\mathcal{S}(\boldsymbol{\nu}) \subseteq \mathcal{M}_{Y|\mathbf{X}}. \tag{6.1}$$

Let $\mathbf{P}_{\mathcal{M}_{Y|\mathbf{X}}(\boldsymbol{\Sigma})} = \mathbf{R}(\mathbf{R}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{R})^{-1}\mathbf{R}^{\mathrm{T}}\boldsymbol{\Sigma}$ be an orthogonal projection operator $\mathbf{P}_{\mathcal{M}_{Y|\mathbf{X}}(\boldsymbol{\Sigma})}$ onto $\mathcal{M}_{Y|\mathbf{X}}$ relative to $< a, b >_{\boldsymbol{\Sigma}}$, where $\mathbf{R}$ is a $p \times q$ matrix such that $\mathcal{S}(\mathbf{R}) = \mathcal{M}_{Y|\mathbf{X}}$. By (6.1), the following equivalences are derived:

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu} = \mathbf{P}_{\mathcal{M}_{Y|\mathbf{X}}(\boldsymbol{\Sigma})}\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu} = \mathbf{R}\left(\mathbf{R}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{R}\right)^{-1}\mathbf{R}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu} = \mathbf{R}\left(\mathbf{R}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{R}\right)^{-1}\mathbf{R}^{\mathrm{T}}\boldsymbol{\nu}. \tag{6.2}$$

Since $\boldsymbol{\Sigma}^{-1}\mathcal{S}(\boldsymbol{\nu}) = \mathcal{S}_{Y|\mathbf{X}}$, we have $\mathcal{S}\{\mathbf{R}(\mathbf{R}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{R})^{-1}\mathbf{R}^{\mathrm{T}}\boldsymbol{\nu}\} = \mathcal{S}_{Y|\mathbf{X}}$ by the last equivalence in (6.2). Here, one crucially notable thing is that the inversion of $\mathbf{R}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{R}$, not directly that of $\boldsymbol{\Sigma}$, is required. Therefore, if the $q \times q$ matrix $\mathbf{R}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{R}$ has full-rank, it is invertible. Although it is not, $(\mathbf{R}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{R})^{-1}$ can be replaced by the Moore-Penrose inverse.

The matrix $\mathbf{R}$ is assumed to be known; therefore, it has to be constructed so that its column spans a subspace large enough to contain $\mathcal{S}_{Y|\mathbf{X}}$ but reasonably estimable from data. For this, Cook *et al.* (2007) proposed iterative projections of $\boldsymbol{\nu}$ onto $\boldsymbol{\Sigma}$:

$$\mathbf{R}_u \equiv \left(\boldsymbol{\nu}, \boldsymbol{\Sigma}\boldsymbol{\nu}, \ldots, \boldsymbol{\Sigma}^{u-1}\boldsymbol{\nu}\right), \quad u = 1, 2, \ldots u^*. \tag{6.3}$$

The letter u in (6.3) is called a termination index of projections. It is obvious that $\mathcal{S}(\mathbf{R}_{u-1}) \subseteq \mathcal{S}(\mathbf{R}_u)$ for any $u \geq 2$, so $\mathcal{S}(\mathbf{R}_u)$ forms a nondecreasing sequence. Therefore, it is important to determine when to stop the projections and find a proper value of the termination index u that is small enough to guarantee that $\mathcal{S}(\mathbf{R}_u) = \mathcal{S}_{Y|\mathbf{X}}$. Recently Yoo (2013a) suggests bootstrap coefficients of variations to determine the termination index, which does not require any asymptotics and is implemented in a simple way.

Yoo (2013a) define bootstrap coefficient variation (BCV) as a selection criteria:

$$\text{BCV} = \frac{\sum_{\ell=1}^{b} \left(\hat{F}_u^{\ell} - \hat{F}_u^{\text{ref}}\right)^2 \Big/ b}{\left(\bar{F}_u^{\text{b}} - \hat{F}_u^{\text{ref}}\right)^2},$$

where b stands for the number of bootstrap samples and $\hat{F}_u^{\ell}$ and $\hat{F}_u^{\text{ref}}$ represent sample increments between $\mathcal{S}(\hat{\mathbf{R}}_{u-1})$ and $\mathcal{S}(\hat{\mathbf{R}}_u)$ computed from the ℓth bootstrap sample and the original sample, respectively. For $u = 1, \ldots, u^{\max}$, the BCV is constructed from bootstrap samples of $(\mathbf{X}, Y)$, where $u^{\max}$ is user-selected. Then, choose u^* to give the smallest among all BCVs considered.

In practice, first, choose $\boldsymbol{\nu}$ among the three candidates discussed above. Then $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\nu}}$ are constructed by their usual moment estimator and sample quantities, respectively. Then a proper value of u, saying $u*$, is determined. Then the sample version $\hat{\mathbf{R}}_{u*}$ is constructed such that

$$\hat{\mathbf{R}}_{u*} = \left(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\nu}}, \ldots, \hat{\boldsymbol{\Sigma}}^{u*-1}\hat{\boldsymbol{\nu}}\right).$$

Finally $\hat{\mathbf{R}}_{u^*}(\hat{\mathbf{R}}_{u^*}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{R}}_{u^*})^{-1}\hat{\mathbf{R}}_{u^*}^{\mathrm{T}}\hat{\boldsymbol{\nu}}$ becomes an estimate of a basis of $\mathcal{S}_{Y|\mathbf{X}}$. The sufficient dimension reduction through the successive projection of seed matrices is called a *seeded dimension reduction*.

7. Discussion

In the paper, we discuss SDR methodologies popularly used to estimate the central subspace, the central mean subspace and the central k^{th}-moment subspace. A permutation test to determine the structural dimension is also introduced. The permutation test approach can be applied to the methodologies discussed in the paper and do not require large-sample distributions. The theoretical relationships between the methodologies are investigated and a seeded dimension reduction approach is introduced for the methodologies to apply to large p small n regressions.

There are still many topics that should be further discussed such as:

(1) Tests of predictor effects: testing $\text{H}_0 : \mathbf{P}_{\mathbf{H}}\mathcal{S}_{\mathbf{X}} = O_p$, where $\mathbf{H}$ is a $p \times h$ user-selected predictor matrix. For example, if $\mathbf{P}_{\mathbf{H}}\mathcal{S}_{\mathbf{X}} = O_p$ holds for $\mathbf{H} = (1, 0, \ldots, 0)$, then the first coordinate variate X_1 in $\mathbf{X}$ do not contribute to $\mathcal{S}_{\mathbf{X}}$. Then X_1 can be eliminated.

(2) Partial sufficient dimension reduction: inference on $\boldsymbol{\eta}$ such that $Y \perp\!\!\!\perp \mathbf{X}|(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}, W)$, where W is a c-level categorical predictor.

(3) Minimum discrepancy approach: inference on $\boldsymbol{\eta}$ with arguments $\hat{\mathbf{B}}$ that minimizes the objective function, $F_m(\mathbf{B}, \mathbf{C})$, under $\text{H}_0 : d = m$: $\text{argmin}_{\mathbf{B},\mathbf{C}} F_m(\mathbf{B}, \mathbf{C})$

$$F_m(\mathbf{B}, \mathbf{C}) = \left\{\text{vec}\left(\hat{\mathbf{M}}\right) - \text{vec}(\mathbf{BC})\right\}^{\mathrm{T}} \mathbf{V}_n \left\{\text{vec}\left(\hat{\mathbf{M}}\right) - \text{vec}(\mathbf{BC})\right\},$$

where $\mathbf{B}$ is a $p \times m$ matrix and $\mathbf{C}$ is a $m \times r$ matrix.

(4) Sufficient dimension reduction in multivariate regression: inference of $\mathcal{S}_{\mathbf{X}}$ under multivariate regression.

(5) Response dimension reduction in multivariate regression: dimension reduction of multi-dimensional responses without loss of information of the conditional mean.

(6) Model-based sufficient dimension reduction: inference on $\boldsymbol{\eta}$ under the following semi-parametric model:

$$\mathbf{X}_y = \mu + \boldsymbol{\eta}\boldsymbol{\beta}\mathbf{f}_y + \sigma\boldsymbol{\varepsilon}.$$

We will leave these topics for later research opportunities.

Acknowledgments

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-2014R1A2A1A11049389 and 2009-0093827).

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

References

Bura E and Cook RD (2001). Extending sliced inverse regression: the weighted chi-squared test, *Journal of the American Statistical Association*, **96**, 996–1003.

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

Cook RD (1998b). Principal Hessian directions revisited, *Journal of the American Statistical Association*, **93**, 84–94.

Cook RD and Critchley F (2000). Identifying regression outliers and mixtures graphically, *Journal of the American Statistical Association*, **95**, 781–794.

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

Cook RD and Li B (2004). Determining the dimension of iterative Hessian transformation, *Annals of Statistics*, **32**, 2501–2531.

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.

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

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.

Shao Y, Cook RD, and Weisberg S (2007). Marginal tests with sliced average variance estimation, *Biometrika*, **94**, 285–296.

Stein CM (1981). Estimation of the mean of a multivariate normal distribution, *Annals of Statistics*, **9**, 1135–1151.

Ye Z and Weiss RE (2003). Using the bootstrap to select one of a new class of dimension reduction methods, *Journal of the American Statistical Association*, **98**, 968–979.

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

Yoo JK (2013a). Advances in seeded dimension reduction: bootstrap criteria and extensions, *Computational Statistics & Data Analysis*, **60**, 70–79.

Yoo JK (2013b). Chi-squared tests in *k*th-moment sufficient dimension reduction, *Journal of Statistical Computation and Simulation*, **83**, 191–201.

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

Received January 22, 2016; Revised February 13, 2016; Accepted February 13, 2016