

A Note on Bootstrapping in Sufficient Dimension Reduction

Jae Keun Yoo^{1,a}, Sun Jeong^b

^aDepartment of Statistics, Ewha Womans University, Korea; ^bSamsung Life Insurance, Korea

Abstract

A permutation test is the popular and attractive alternative to derive asymptotic distributions of dimension test statistics in sufficient dimension reduction methodologies; however, recent studies show that a bootstrapping technique also can be used. We consider two types of bootstrapping dimension determination, which are partial and whole bootstrapping procedures. Numerical studies compare the permutation test and the two bootstrapping procedures; subsequently, real data application is presented. Considering two additional bootstrapping procedures to the existing permutation test, one has more supporting evidence for the dimension estimation of the central subspace that allow it to be determined more convincingly.

Keywords: bootstrapping, central subspace, permutation, regression, sufficient dimension reduction

1. Introduction

Sufficient dimension reduction (SDR) is a statistical technique of dimension reduction in regression without losing information on $Y \in \mathbb{R}^1 | \mathbf{X} \in \mathbb{R}^p$. Its usual goal is to replace the p -dimensional original predictors \mathbf{X} by its a lower-dimensional linear projection $\boldsymbol{\eta}^T \mathbf{X}$ without loss of information about $Y | \mathbf{X}$, which is equivalently stated as:

$$Y \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\eta}^T \mathbf{X}, \quad (1.1)$$

where $\perp\!\!\!\perp$ stands for independence, $\boldsymbol{\eta} \in \mathbb{R}^{p \times d}$, and $d \leq p$.

A subspace spanned by the columns of $\boldsymbol{\eta}$ to satisfy (1.1) is called a dimension reduction subspace. SDR typically seeks for the intersection of all dimension reduction subspaces, which is called the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. Classical but still dominate SDR methods to estimate $\mathcal{S}_{Y|\mathbf{X}}$ should be sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook, 2000), principal Hessian directions (pHd; Li, 1992), and polynomial ordinary least square (POLS; Yin and Cook, 2002). Before explaining the methods, we define $\boldsymbol{\Sigma} = \text{cov}(\mathbf{X})$ and $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - E(\mathbf{X}))$. We denote $\mathcal{S}(\mathbf{B})$ as a subspace spanned by the columns of $\mathbf{B} \in \mathbb{R}^{p \times q}$.

SIR constructs $E(\mathbf{Z}|Y)$ to estimate $\mathcal{S}_{Y|\mathbf{X}}$. It is known that the subspace spanned by $E(\mathbf{Z}|Y)$ is equal to $\mathcal{S}[\mathbf{M}_{\text{SIR}} = \text{cov}\{E(\mathbf{Z}|Y)\}]$. For a categorical Y , the construction of a sample version of $E(\mathbf{Z}|Y)$ is

For the corresponding author Jae Keun Yoo, this works was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education, Science & Technology (NRF-2012R1A1A1040077).

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

Published 31 May 2015 / journal homepage: <http://csam.or.kr>

© 2015 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

straightforward. If Y is many-valued or continuous, Y is categorized by dividing its range into h slices. Then, $\Sigma^{-1/2}\mathcal{S}(\mathbf{M}_{\text{SIR}}) \subseteq \mathcal{S}_{Y|X}$. While SIR focuses on $E(\mathbf{Z}|Y)$ to estimate $\mathcal{S}_{Y|X}$, SAVE pays attention onto $\text{cov}(\mathbf{Z}|Y)$. Cook (2000) showed that $\mathbf{M}_{\text{SAVE}} = E\{\mathbf{I}_p - \text{cov}(\mathbf{Z}|Y)\}^2$ is informative to $\mathcal{S}_{Y|X}$, so that $\Sigma^{-1/2}\mathcal{S}(\mathbf{M}_{\text{SAVE}}) \subseteq \mathcal{S}_{Y|X}$. The sample version of $\text{cov}(\mathbf{X}|Y)$ is constructed by the slicing scheme of Y in SIR. The method of pHd constructs $\Sigma_{yzz} = E\{(Y - E(Y))\mathbf{Z}\mathbf{Z}^T\}$ to estimate $\mathcal{S}_{Y|X}$. In practice, the response Y in Σ_{yzz} is often replaced by the OLS residuals of $\epsilon = Y - E(Y) - \beta^T\{\mathbf{X} - E(\mathbf{X})\}$ and we construct $\Sigma_{\epsilon z z} = E(\epsilon\mathbf{Z}\mathbf{Z}^T)$, where $\beta = \Sigma^{-1}\text{cov}(\mathbf{X}, Y)$. Again, we have $\Sigma^{-1/2}\mathcal{S}(\mathbf{M}_{\text{pHd}} = \Sigma_{\epsilon z z}) \subseteq \mathcal{S}_{Y|X}$. POLS constructs a $p \times q$ matrix $\mathbf{M}_{\text{POLS}} = \{\text{cov}(\mathbf{Z}, W), \text{cov}(\mathbf{Z}, W^2), \dots, \text{cov}(\mathbf{Z}, W^q)\}$, where $W = (Y - E(Y))/\sqrt{\text{var}(Y)}$. According to Yin and Cook (2002), the relation of $\Sigma^{-1/2}\mathcal{S}(\mathbf{M}_{\text{POLS}}) \subseteq \mathcal{S}_{Y|X}$ is established.

The estimation of $\mathcal{S}_{Y|X}$ through the kernel matrices of \mathbf{M}_\bullet requires two steps to estimate the true dimension and an orthonormal basis of $\mathcal{S}_{Y|X}$. These steps are usually done as follows. First, spectral-decompose \mathbf{M}_\bullet and then test the number of nonzero eigenvalues, which is an estimate of the dimension of $\mathcal{S}_{Y|X}$. Then the eigenvectors corresponding to the nonzero eigenvalues form an orthonormal basis of $\mathcal{S}_{Y|X}$. In the process of the dimension estimation of $\mathcal{S}_{Y|X}$, a test statistic and its null distribution are required. When the derivation of the null distribution is difficult, a popular alternative is a permutation test (Yin and Cook, 2002). We will later discuss about the dimension estimation and the permutation test.

Recently Ye and Weiss (2003) proposed a bootstrap approach and discussed that it was effective to choose one member from a large class of dimension reduction methods. Recent usage of bootstrapping in SDR context can be found in Zhu and Zeng (2006), Yoo (2011) and Yoo (2013b). Bootstrapping is well established in sufficient dimension reduction; however numerical studies for the dimension estimation and comparison with the permutation test have not yet been well studied. This paper represents a comprehensive study on bootstrapping in the dimension estimation of $\mathcal{S}_{Y|X}$ in sufficient dimension reduction.

The organization of the paper is as follows. In Section 2 the dimension estimation and the permutation test are reviewed. In Section 3, numerical studies on the dimension estimation of $\mathcal{S}_{Y|X}$ through the bootstrapping and the comparison with the permutation test are provided. Real data application is presented in Section 4, and we summarize our work in Section 5.

2. Dimension Estimation, Permutation and Bootstrapping

2.1. Dimension estimation in sufficient dimension reduction

For notational convenience, denote d as the true dimension of $\mathcal{S}_{Y|X}$. Then d is determined through testing a series of hypothesis. Starting with $m = 0$, the hypothesis of $H_0 : d = m$ versus $H_1 : d > m$, $m = 0, 1, \dots, (p - 1)$, is tested. If the null hypothesis is rejected with level α , increment m by 1 and redo the test. The test is terminated for the first time $H_0 : d = m$ is not rejected, and the estimate \hat{d} of d is set to m in the null hypothesis. This estimation procedure requires a test statistic. Let λ_i , $i = 1, \dots, p$, represent the ordered-eigenvalues of \mathbf{M}_\bullet such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. A test statistic Λ_m for $H_0 : d = m$ versus $H_1 : d > m$ is the sum of the eigenvalues multiplied by the sample size n , such as $\Lambda_m = n \sum_{i=m+1}^p \lambda_i$. This test statistic has been used in many SDR methods, including the four methods discussed in the Section 1.

Apart from the original works of SIR, SAVE, pHd and POLS, general results regarding their test statistics are presented in Bura and Cook (2001), Shao *et al.* (2007), Cook (1998) and Yoo (2013a) in the order; however, there exists big time gap between the original works and the derivation of the null distributions. Before the general results, a permutation test has been commonly adopted and will be

discussed in the next subsection.

2.2. Permutation

For notational conveniences, $\hat{\mathbf{M}}_{\bullet} \in \mathbb{R}^{p \times p}$ represents a sample version of \mathbf{M}_{\bullet} constructed from $\hat{\mathbf{Z}} = \hat{\Sigma}^{-1/2}(\mathbf{X} - \hat{E}(\mathbf{X}))$, and pairs of $(\hat{\lambda}_i, \hat{\gamma}_i)$, $i = 1, \dots, p$, stand for the eigen-structure of $\hat{\mathbf{M}}_{\bullet}$ such that $\hat{\mathbf{M}}_{\bullet} = \sum_{i=1}^p \hat{\lambda}_i \hat{\gamma}_i \hat{\gamma}_i^T$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p$. Then, a permutation test is done as follows.

- (1) Compute the sample kernel matrix $\hat{\mathbf{M}}_{\bullet}$, and, under $H_0 : d = m$, obtain $\hat{\Lambda}_m$ and partition eigenvector matrices

$$\hat{\Gamma}_1 = (\hat{\gamma}_1, \dots, \hat{\gamma}_m) \quad \text{and} \quad \hat{\Gamma}_2 = (\hat{\gamma}_{m+1}, \dots, \hat{\gamma}_p).$$

- (2) Construct the vectors $\hat{V}_i = \hat{\Gamma}_1^T \hat{\mathbf{Z}}_i$ and $\hat{U}_i = \hat{\Gamma}_2^T \hat{\mathbf{Z}}_i$.
- (3) Randomly permute the indice i of the \hat{U}_i to obtain the permuted set \hat{U}_i^{perm} .
- (4) Construct the test statistic $\hat{\Lambda}_m^{\text{perm}}$ based on the partially permuted data of Y_i and $(\hat{V}_i, \hat{U}_i^{\text{perm}})$.
- (5) Repeat Steps (3)–(4) N times. The p -value of the hypothesis testing is the fraction that $\hat{\Lambda}_m^{\text{perm}} > \hat{\Lambda}_m$.

2.3. Bootstrapping

Ye and Weiss (2003) proposed a bootstrap approach in sufficient dimension reduction to select a best one from a class of SDR methods. In the bootstrap approach, it is necessary to measure a distance between two subspaces. For this, the vector correlation coefficient q (Hotelling, 1936) and the trace correlation r (Hooper, 1959) were adopted in Ye and Weiss (2003). Suppose two k -dimensional subspaces of $\mathcal{S}(\mathbf{A} \in \mathbb{R}^{p \times k})$ and $\mathcal{S}(\mathbf{B} \in \mathbb{R}^{p \times k})$ such that $\mathbf{A}^T \mathbf{A} = \mathbf{I}_k$ and $\mathbf{B}^T \mathbf{B} = \mathbf{I}_k$. Then, let ρ_i^2 , $i = 1, \dots, k$, stand for the ordered eigenvalues of $\mathbf{B}^T \mathbf{A} \mathbf{A}^T \mathbf{B}$. Then, the vector correlation q and the trace correlation r are defined as: $q^2(\mathbf{A}, \mathbf{B}) = \prod_{i=1}^k \rho_i^2$ and $r^2(\mathbf{A}, \mathbf{B}) = (1/k) \sum_{i=1}^k \rho_i^2$. The two values of $q^2(\mathbf{A}, \mathbf{B})$ and $r^2(\mathbf{A}, \mathbf{B})$ varies from 0 to 1, and $q^2(\mathbf{A}, \mathbf{B})$ and $r^2(\mathbf{A}, \mathbf{B})$ are equal to 1, if the two subspaces of $\mathcal{S}(\mathbf{A})$ and $\mathcal{S}(\mathbf{B})$ are equivalent. To change a concept of correlation (higher and closer) to that of distance (smaller and closer), we consider the following: $q_D(\mathbf{A}, \mathbf{B}) = 1 - |\sqrt{q^2(\mathbf{A}, \mathbf{B})}|$ and $r_D(\mathbf{A}, \mathbf{B}) = 1 - |\sqrt{r^2(\mathbf{A}, \mathbf{B})}|$.

Next select candidate SDR methods and construct their corresponding sample kernel matrices $\hat{\mathbf{M}}_{\bullet}$. From the original sample (Y_i, \mathbf{X}_i) , $i = 1, \dots, n$, generate N bootstrap samples (Y_i^b, \mathbf{X}_i^b) , $b = 1, \dots, N$, and obtain bootstrap sample kernel matrices $\hat{\mathbf{M}}_{\bullet}^b$. For each method, compute $q_D^b(\hat{\mathbf{M}}_{\bullet}, \hat{\mathbf{M}}_{\bullet}^b)$ and $r_D^b(\hat{\mathbf{M}}_{\bullet}, \hat{\mathbf{M}}_{\bullet}^b)$ for $b = 1, \dots, N$. Ye and Weiss (2003) discussed that the best method among the candidates should have less distances between $\mathcal{S}(\hat{\mathbf{M}}_{\bullet})$ and $\mathcal{S}(\hat{\mathbf{M}}_{\bullet}^b)$ than others. So, for each method, compute the average distances of \bar{q}_D and \bar{r}_D such that $\bar{q}_D = (1/N) \sum_{b=1}^N q_D^b(\hat{\mathbf{M}}_{\bullet}, \hat{\mathbf{M}}_{\bullet}^b)$ and $\bar{r}_D = (1/N) \sum_{b=1}^N r_D^b(\hat{\mathbf{M}}_{\bullet}, \hat{\mathbf{M}}_{\bullet}^b)$. Then select the method to give the smallest \bar{q}_D and \bar{r}_D .

2.4. Bootstrap dimension estimation

The bootstrapping approach discussed in the previous subsection was not directly oriented for the dimension estimation because it does not give a p -value for the test. So, two versions of modification of the bootstrap approach are considered to estimate the dimension of $\mathcal{S}_{Y|\mathbf{X}}$. One of them will be called *whole bootstrap* and the other will be called *partial bootstrap*.

The partial bootstrap approach is to replace permutation samples in the permutation test by bootstrap samples, and follow the same procedure as the permutation test. The whole bootstrap approach for the dimension estimation can be done as follows.

- (1) From the original sample, compute the sample kernel matrix $\hat{\mathbf{M}}_{\bullet}$. Then, under $H_0 : d = m$, obtain a set of the eigenvectors $\hat{\mathbf{\Gamma}}_m = (\hat{\gamma}_1, \dots, \hat{\gamma}_m)$ of $\hat{\mathbf{M}}_{\bullet}$, where $\hat{\gamma}_i$, $i = 1, \dots, m$, is the eigenvector corresponding to the i th largest eigenvalue of $\hat{\mathbf{M}}_{\bullet}$.
- (2) Construct N bootstrap samples, and construct bootstrap sample kernel matrices of $\hat{\mathbf{M}}_{\bullet}^b$. Then, under $H_0 : d = m$, obtain $\hat{\mathbf{\Gamma}}_m^b = (\hat{\gamma}_1^b, \dots, \hat{\gamma}_m^b)$ from $\hat{\mathbf{M}}_{\bullet}^b$.
- (3) Compute \bar{q}_D and \bar{r}_D from pairs of $(\hat{\mathbf{\Gamma}}_m, \hat{\mathbf{\Gamma}}_m^b)$, $b = 1, \dots, N$, and denote them as \bar{q}_D^m and \bar{r}_D^m , respectively.
- (4) Do Steps (2)–(3) for $m = 1, \dots, p$.
- (5) Set \hat{d} to m to give the smallest \bar{q}_D^m and \bar{r}_D^m .

In Step (5), the estimation of d is done, following the guidance of Ye and Weiss (2003). The distance between $\hat{\mathbf{\Gamma}}_m$ and $\hat{\mathbf{\Gamma}}_m^b$ is expected to be the smallest under the true dimension. Therefore, it is natural to choose m to have the smallest values of \bar{q}_D^m and \bar{r}_D^m as the estimate of the true dimension. It should be pointed out that the whole bootstrap approach cannot be done for $H_0 : d = 0$, because $\hat{\mathbf{\Gamma}}_0$ does not exist. However, it is no issue in practice, because $d = 0$ is not of main interest in most regression problems.

3. Numerical Studies

For numerical studies, four artificial models were considered. Commonly, five-dimensional predictors $\mathbf{X} = (X_1, \dots, X_5)^T$ and a random error ε were independently generated from $N(0, 1)$.

$$\text{Model 1: } Y|\mathbf{X} = X_1 + X_1X_2 + \varepsilon.$$

$$\text{Model 2: } Y|\mathbf{X} = X_1 + \exp(X_2) + \varepsilon.$$

$$\text{Model 3: } Y|\mathbf{X} = X_1^2 + \varepsilon.$$

$$\text{Model 4: } Y|\mathbf{X} = X_1^2 + X_2^2 + \varepsilon.$$

In the three models except Model 3, two columns of $((1, 0, 0, 0, 0), (0, 1, 0, 0, 0))^T$ span $\mathcal{S}_{Y|\mathbf{X}}$, and hence the true dimension d is equal to two, while $\mathcal{S}_{Y|\mathbf{X}}$ is spanned by $(1, 0, 0, 0, 0)^T$ in Model 3, so $d = 1$.

The methods of SIR and POLS work well to detect linear trends, while SAVE and pHd do to detect non-linear trends such as quadratic relationship. So, for Models 1–2 to have linear trend, SIR and POLS with $q = 3$ were applied, while Models 3–4 were fitted with SAVE and pHd. The number N of permutation and bootstrapping were 500 and 1000 for each sample size of 50, 100, 200 and 400. The total number of iterations for each model was 500; subsequently, level $\alpha = 0.05$ was used. Tables 1–8 reports the results Results of the numerical studies.

Tables 1 and 2 show the dimension determination results for Model 1 are interesting. The permutation and partial bootstrapping tests provide similar dimension test results regardless of the application of SIR and POLS. However, the results for the whole bootstrapping are quite different from the choice

Table 1: Percentages of dimension determination for Model 1 with SIR application

SIR		N = 500				N = 1000			
Type	n	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$
Permutation	50	34.8	54.2	9.8	1.2	34.0	54.8	10.2	1.0
	100	6.8	65.6	25.8	1.8	6.8	66.8	24.2	2.2
	200	0.0	42.2	55.0	2.8	0.0	41.4	56.2	2.4
	400	0.0	6.2	89.8	4.0	0.0	5.6	90.2	4.2
Partial bootstrapping	50	35.2	54.0	9.8	1.0	33.6	55.8	9.6	1.0
	100	6.6	65.4	25.4	2.2	6.8	66.4	24.8	2.0
	200	0.0	42.0	55.4	2.6	0.0	42.0	55.2	2.8
	400	0.0	5.6	90.2	4.2	0.0	6.0	89.4	4.6
Whole bootstrapping \bar{q}_D	50	N/A	98.4	1.6	0.0	N/A	98.0	1.6	0.4
	100	N/A	97.2	2.6	0.2	N/A	97.4	2.4	0.2
	200	N/A	98.6	1.4	0.0	N/A	98.4	1.6	0.0
	400	N/A	97.2	2.8	0.0	N/A	97.8	2.2	0.0
Whole bootstrapping \bar{r}_D	50	N/A	5.8	1.4	92.8	N/A	5.6	1.0	93.4
	100	N/A	28.0	6.2	65.8	N/A	27.2	6.2	66.6
	200	N/A	67.0	15.0	18.0	N/A	67.4	14.4	18.2
	400	N/A	79.0	17.8	3.2	N/A	79.4	17.6	3.0

SIR = sliced inverse regression.

Table 2: Percentages of dimension determination for Model 1 with POLS application

POLS		N = 500				N = 1000			
Type	n	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$
Permutation	50	9.2	48.2	38.6	4.0	9.4	48.0	38.6	4.0
	100	1.4	18.8	72.4	7.4	1.4	18.0	72.6	8.0
	200	0.2	4.8	86.4	8.6	0.2	4.6	85.6	9.6
	400	0.0	0.0	92.0	8.0	0.0	0.0	92.0	8.0
Partial bootstrapping	50	9.8	46.4	39.8	4.0	9.6	48.0	38.4	4.0
	100	1.4	19.0	71.6	8.0	1.4	18.4	72.6	7.6
	200	0.2	4.8	86.2	8.8	0.2	4.8	86.0	9.0
	400	0.0	0.0	92.6	7.4	0.0	0.0	92.6	7.4
Whole bootstrapping \bar{q}_D	50	N/A	96.2	3.6	0.2	N/A	94.2	4.4	1.4
	100	N/A	88.4	17.6	0.0	N/A	88.8	11.2	0.0
	200	N/A	52.6	47.4	0.0	N/A	52.5	47.5	0.0
	400	N/A	17.0	83.0	0.0	N/A	17.0	83.0	0.0
Whole bootstrapping \bar{r}_D	50	N/A	1.8	2.0	96.2	N/A	1.6	2.1	96.3
	100	N/A	1.2	13.4	85.4	N/A	1.5	13.8	84.7
	200	N/A	0.0	36.2	63.8	N/A	0.0	36.4	63.6
	400	N/A	0.0	68.0	32.0	N/A	0.0	68.1	31.9

POLS = polynomial ordinary least square.

of the SDR methods. This implies that the whole bootstrapping is sensitive to the choice of the SDR methods and is why it is used to select a best SDR method among many candidates. From Table 1 by the SIR application, the permutation and partial bootstrapping tests show even better performance than the whole bootstrapping with both \bar{q}_D and \bar{r}_D . However, the differences become narrower in Table 2 by the POLS application.

Tables 3 and 4 for Model 2 indicate that the whole bootstrapping has better performance in the estimation of d than the permutation and partial bootstrapping tests for all sample sizes considered. For the SIR application, \bar{q}_D provides the best results, while \bar{r}_D is the best for the POLS application.

Tables 5 and 6 for Model 3, in the pHd application, the permutation and partial bootstrapping tests and the whole bootstrapping with \bar{q}_D give similar results, while the whole bootstrapping with \bar{r}_D suffers the overestimation of d with 50 sample sizes. However, the permutation and partial bootstrapping

Table 3: Percentages of dimension determination for Model 2 with SIR application

SIR		N = 500				N = 1000			
Type	n	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$
Permutation	50	0.4	88.8	9.6	1.2	0.6	88.2	9.8	1.4
	100	0.0	81.4	16.6	2.0	0.0	79.8	18.4	1.8
	200	0.0	58.6	40.0	1.4	0.0	57.8	41.0	1.2
	400	0.0	19.6	76.8	3.6	0.0	20.4	76.0	3.6
Partial bootstrapping	50	1.0	88.0	9.4	1.6	0.2	89.0	9.4	1.4
	100	0.0	80.4	17.6	2.0	0.0	80.8	17.6	1.6
	200	0.0	57.0	41.4	1.6	0.0	58.2	40.0	1.8
	400	0.0	21.0	75.2	3.8	0.0	20.6	75.8	3.6
Whole bootstrapping \bar{q}_D	50	N/A	96.0	4.0	0.0	N/A	96.4	3.4	0.2
	100	N/A	82.2	17.8	0.0	N/A	82.4	17.6	0.0
	200	N/A	51.4	48.4	0.2	N/A	53.2	46.6	0.2
	400	N/A	17.6	82.0	0.4	N/A	16.8	82.8	0.4
Whole bootstrapping \bar{r}_D	50	N/A	1.8	1.8	96.4	N/A	2.0	1.8	96.2
	100	N/A	1.2	13.4	85.4	N/A	1.6	13.0	85.4
	200	N/A	0.0	35.2	64.8	N/A	0.0	36.4	63.6
	400	N/A	0.0	68.0	32.0	N/A	0.0	67.4	32.6

SIR = sliced inverse regression.

Table 4: Percentages of dimension determination for Model 2 with POLS application

POLS		N = 500				N = 1000			
Type	n	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$
Permutation	50	5.4	53.4	37.4	3.8	5.8	52.4	38.0	3.8
	100	0.2	34.4	57.8	7.6	0.0	35.0	57.6	7.4
	200	0.0	15.2	74.8	10.0	0.0	14.8	75.2	10.0
	400	0.0	8.6	72.2	19.2	0.0	8.6	73.8	17.6
Partial bootstrapping	50	5.6	53.2	36.8	4.4	5.2	53.8	36.8	4.2
	100	0.0	35.2	57.4	7.4	0.0	35.4	57.8	6.8
	200	0.0	14.8	75.2	10.0	0.0	15.0	74.6	10.4
	400	0.0	8.8	72.6	18.6	0.0	8.2	73.2	18.6
Whole bootstrapping \bar{q}_D	50	N/A	45.6	53.6	0.8	N/A	45.8	53.6	0.6
	100	N/A	29.8	69.4	0.8	N/A	29.0	70.4	0.6
	200	N/A	17.0	83.0	0.0	N/A	16.6	83.4	0.0
	400	N/A	11.0	88.8	0.2	N/A	11.4	88.4	0.2
Whole bootstrapping \bar{r}_D	50	N/A	2.8	55.4	41.8	N/A	2.6	56.0	41.4
	100	N/A	1.4	78.2	20.4	N/A	1.2	78.2	20.6
	200	N/A	0.0	89.4	10.6	N/A	0.0	89.4	10.6
	400	N/A	0.0	90.0	10.0	N/A	0.0	90.0	10.0

POLS = polynomial ordinary least square.

tests suffer the underestimation of d for 50 sample sizes in the SAVE application and the whole bootstrapping with \bar{r}_D does the overestimation of d for the same sample sizes. The whole bootstrapping with \bar{q}_D alone results in the consistently correct estimation of d .

Tables 7 and 8 for Model 4 indicate that, regardless of pHd and SAVE, the permutation and partial bootstrapping tests and the whole bootstrapping with \bar{q}_D often underestimate the true dimension with smaller sample sizes such as $n = 50$ and 100, while the whole bootstrapping with \bar{r}_D overestimates d in sample sizes. However, with moderate sample sizes, the four approaches similarly determine the true dimension correctly.

From the numerical studies, what we can commonly observe is as follows. There is no difference between the permutation and partial bootstrapping tests. Usage of \bar{q}_D often yields more percentages of the correct decisions than that of \bar{r}_D . And, \bar{q}_D tends to underestimate the true dimension, while \bar{r}_D

Table 5: Percentages of dimension determination for Model 3 with pHd application

pHd		$N = 500$			$N = 1000$		
Type	n	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} > 1$	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} > 1$
Permutation	50	6.2	90.0	3.8	6.8	89.2	4.0
	100	0.0	95.4	4.6	0.0	95.8	4.2
	200	0.0	95.2	4.8	0.0	94.8	5.2
	400	0.0	95.0	5.0	0.0	95.0	5.0
Partial bootstrapping	50	4.4	91.0	4.6	4.8	91.2	4.0
	100	0.0	94.0	6.0	0.0	94.8	5.2
	200	0.0	94.6	5.2	0.0	94.2	5.8
	400	0.0	94.8	5.2	0.0	95.0	5.0
Whole bootstrapping \bar{q}_D	50	N/A	99.8	0.2	N/A	100.0	0.0
	100	N/A	100.0	0.0	N/A	100.0	0.0
	200	N/A	100.0	0.0	N/A	100.0	0.0
	400	N/A	100.0	0.0	N/A	100.0	0.0
Whole bootstrapping \bar{r}_D	50	N/A	51.2	48.8	N/A	51.4	48.6
	100	N/A	92.2	7.85	N/A	92.6	7.4
	200	N/A	100.0	0.0	N/A	100.0	0.0
	400	N/A	100.0	0.0	N/A	100.0	0.0

pHd = principal Hessian directions.

Table 6: Percentages of dimension determination for Model 3 with SAVE application

SAVE		$N = 500$			$N = 1000$		
Type	n	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} > 1$	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} > 1$
Permutation	50	59.0	35.4	5.6	58.6	36.2	5.2
	100	6.2	88.6	5.2	7.0	87.6	3.6
	200	0.0	94.0	6.0	0.0	94.2	5.8
	400	0.0	95.0	5.0	0.0	94.8	5.2
Partial bootstrapping	50	66.0	29.8	4.2	67.0	28.8	4.2
	100	7.6	87.6	4.8	8.6	86.6	4.8
	200	0.0	94.2	5.8	0.0	94.8	5.2
	400	0.0	95.6	4.4	0.0	95.8	4.2
Whole bootstrapping \bar{q}_D	50	N/A	98.2	1.8	N/A	98.8	1.2
	100	N/A	99.8	0.2	N/A	99.8	0.2
	200	N/A	100.0	0.0	N/A	100.0	0.0
	400	N/A	100.0	0.0	N/A	100.0	0.0
Whole bootstrapping \bar{r}_D	50	N/A	8.0	92.0	N/A	7.6	92.4
	100	N/A	77.0	23.0	N/A	77.0	23.0
	200	N/A	99.6	0.4	N/A	99.6	0.4
	400	N/A	100.0	0.0	N/A	100.0	0.0

SAVE = sliced average variance estimation.

does to overestimate it. Therefore, if \bar{r}_D determines a higher dimension than \bar{q}_D and the permutation or partial bootstrapping test decides a value between the two, one should consider the estimate from the permutation or partial bootstrapping test as an estimate of the true dimension.

4. Real Data Application: Soil Evaporation Data

For real data application, we consider how daily soil evaporation is influenced by air temperature area, air temperature range, humidity area and humidity range. The data set contains 46 observations, Cook (1994) has a detailed description.

According to Yin and Cook (2002), any notable non-linearities were not observed through inspecting various graphical plots. So, they determined SIR and POLS to be suitable SDR methods, and it

Table 7: Percentages of dimension determination for Model 4 with pHd application

pHd		$N = 500$				$N = 1000$			
Type	n	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$
Permutation	50	0.4	59.4	35.4	4.8	1.2	57.6	36.8	4.4
	100	0.0	7.6	88.4	4.0	0.0	8.0	88.2	3.8
	200	0.0	0.0	94.8	5.2	0.0	0.0	94.8	5.2
	400	0.0	0.0	94.0	6.0	0.0	0.0	64.4	5.6
Partial bootstrapping	50	0.4	53.4	40.0	6.2	0.4	54.8	39.4	5.4
	100	0.0	7.0	88.8	4.2	0.0	6.8	89.2	4.0
	200	0.0	0.0	94.4	5.6	0.0	0.0	94.4	5.6
	400	0.0	0.0	93.8	6.2	0.0	0.0	93.8	6.2
Whole bootstrapping \hat{q}_D	50	N/A	81.6	15.8	2.6	N/A	82.0	15.6	2.4
	100	N/A	27.4	72.0	0.6	N/A	26.6	72.8	0.6
	200	N/A	6.2	93.8	0.0	N/A	6.6	93.4	0.0
	400	N/A	0.6	99.4	0.0	N/A	0.6	99.4	0.0
Whole bootstrapping \hat{r}_D	50	N/A	5.2	9.6	85.2	N/A	4.8	10.0	85.2
	100	N/A	3.2	81.6	15.2	N/A	3.0	81.4	15.6
	200	N/A	0.6	99.2	0.2	N/A	0.6	99.2	0.2
	400	N/A	0.0	100.0	0.0	N/A	0.0	100.0	0.0

pHd = principal Hessian directions.

Table 8: Percentages of dimension determination for Model 4 with SAVE application

SAVE		$N = 500$				$N = 1000$			
Type	n	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$
Permutation	50	38.8	43.6	11.0	6.6	37.6	44.8	11.4	6.2
	100	0.8	38.6	54.8	5.8	0.2	39.4	54.6	5.8
	200	0.0	0.6	92.2	7.2	0.0	0.4	92.4	7.2
	400	0.0	0.0	95.0	5.0	0.0	0.0	95.2	4.8
Partial bootstrapping	50	49.6	36.6	9.2	4.6	48.8	37.4	8.8	5.0
	100	1.2	41.4	51.8	5.6	1.2	40.8	52.0	6.0
	200	0.0	0.4	92.6	7.0	0.0	0.4	93.0	6.6
	400	0.0	0.0	95.6	4.4	0.0	0.0	95.6	4.4
Whole bootstrapping \hat{q}_D	50	N/A	96.6	2.6	0.8	N/A	96.4	2.8	0.8
	100	N/A	52.2	46.6	1.2	N/A	51.0	48.0	1.0
	200	N/A	6.6	93.4	0.0	N/A	6.4	93.6	0.0
	400	N/A	0.2	99.8	0.0	N/A	0.2	99.8	0.0
Whole bootstrapping \hat{r}_D	50	N/A	0.8	0.2	99.0	N/A	0.8	0.2	99.0
	100	N/A	3.0	38.0	59.0	N/A	2.8	38.0	59.2
	200	N/A	0.6	97.6	1.8	N/A	0.6	97.8	1.6
	400	N/A	0.0	100.0	0.0	N/A	0.0	100.0	0.0

SAVE = sliced average variance estimation.

Table 9: Dimension estimation for soil evaporation data in Section 4

Type	$H_0 : d = 0$	$H_0 : d = 1$	$H_0 : d = 2$	$H_0 : d = 3$	Decision
Permutation (p -value)	0.000	0.016	0.918	N/A	$\hat{d} = 2$
Partial bootstrap (p -value)	0.000	0.028	0.912	N/A	$\hat{d} = 2$
Whole bootstrap (\hat{q}_D)	N/A	0.495	0.503	0.841	$\hat{d} = 1$
Whole bootstrap (\hat{r}_D)	N/A	0.495	0.192	0.173	$\hat{d} = 3$

was discussed that the results of POLS was superior than SIR. Following this guidance, the method of POLS alone was considered for illustration purposes.

To estimate the true dimension of $S_{Y|X}$, POLS with $k = 3$ was applied to the data set. The results of the permutation test, the partial bootstrap test and the whole bootstrap determination with \hat{q}_D and

\bar{r}_D are reported in Table 9.

Table 9 indicates that the permutation and partial bootstrap tests determine $\hat{d} = 2$ with level $\alpha = 0.05$. However, the whole bootstrapping with \bar{q}_D provides $\hat{d} = 1$, while the whole bootstrapping with \bar{r}_D does $\hat{d} = 3$. This coincides what we observed from numerical studies with smaller sample sizes. Therefore, it is reasonably concluded that $\hat{d} = 2$.

5. Discussion

A permutation test is the popular and attractive alternative to derive the asymptotic distributions of dimension test statistics in sufficient dimension reduction methodologies; however, recent studies show that a bootstrapping technique can also be used. We consider two types of bootstrapping dimension determination, which are partial and whole bootstrapping procedures. In the former procedure, subsets of variables are bootstrap-sampled, while the whole variables are bootstrap-sampled in the latter. The partial bootstrapping is similar to the permutation test and can yield relating p -values; however, distance measures using a vector correlation coefficient q and a trace correlation r are adopted since the whole bootstrapping cannot provide p -values for the dimension determination. So, given to a hypothesized dimension, compute the average distances between a basis estimate from the original sample and basis estimates from the whole bootstrap samples, and set the dimension which gives the smallest averages to the estimate of the central subspace.

According to numerical studies, with smaller sample sizes, usage of q often underestimates the true dimension, while that of r overestimates it; in addition, the partial bootstrapping determination is almost equal to the permutation test.

Considering two additional bootstrapping procedures to the existing permutation test, one has more supporting evidences for the dimension estimation of the central subspace, so they can determine it more convincingly.

References

- Bura, E. and Cook, R. D. (2001). Extending SIR: The weighted chi-square test, *Journal of the American Statistical Association*, **96**, 996–1003.
- Cook, R. D. (1994). On the interpretation of regression plots, *Journal of the American Statistical Association*, **89** 177–189.
- Cook, R. D. (1998). Principal Hessian directions revisited. *Journal of the American Statistical Association*, **93**, 84–94.
- Cook, R. D. (2000). SAVE: A method for dimension reduction and graphics in regression, *Communications in Statistics-Theory and Methods*, **29**, 2109–2121.
- Hooper, J. W. (1959). Simultaneous equations and canonical correlation theory, *Econometrica*, **27**, 245–256.
- Hotelling, H. (1936). Relations between two sets of variates, *Biometrika*, **28**, 321–377.
- Li, K. C. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316–327.
- Li, K. C. (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, R. D. and Weisberg, S. (2007). Marginal tests of sliced average variance estimation, *Biometrika*, **94**, 285–296.
- Ye, Z. and Weiss, R. E. (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, R. D. (2002). Dimension reduction for the conditional k th moment in regression, *Journal of Royal Statistical Society Series B*, **64**, 159–175.
- Yoo, J. K. (2011). Unified predictor hypothesis tests in sufficient dimension reduction: A bootstrap approach, *Journal of the Korean Statistical Society*, **40**, 217–225.
- Yoo, J. K. (2013a). Chi-squared tests in k th-moment sufficient dimension reduction, *Journal of Statistical Computation and Simulation*, **83**, 191–201.
- Yoo, J. K. (2013b). Advances in seeded dimension reduction: Bootstrap criteria and extensions, *Computational Statistics and Data Analysis*, **60**, 70–79.
- Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression, *Journal of the American Statistical Association*, **101**, 1638–1651.

Received February 26, 2015; Revised April 27, 2015; Accepted April 29, 2015