Journal of Statistical Planning and Inference 140 (2010) 3121–3131

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Sliced inverse moment regression using weighted chi-squared tests for dimension reduction

Zhishen Ye [a], Jie Yang [b,*]

[a] Amgen Inc., Thousand Oaks, CA 91320-1799, USA
[b] Department of Mathematics, Statistics, and Computer Science (MC 249), University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045, USA

ARTICLE INFO

Article history:
Received 16 February 2009
Accepted 15 April 2010
Available online 20 April 2010

Keywords:
Dimension reduction in regression
pHd
SAVE
SIMR
SIR
Weighted chi-squared test

ABSTRACT

We propose a new method for dimension reduction in regression using the first two inverse moments. We develop corresponding weighted chi-squared tests for the dimension of the regression. The proposed method considers linear combinations of sliced inverse regression (SIR) and the method using a new candidate matrix which is designed to recover the entire inverse second moment subspace. The optimal combination may be selected based on the p-values derived from the dimension tests. Theoretically, the proposed method, as well as sliced average variance estimate (SAVE), is more capable of recovering the complete central dimension reduction subspace than SIR and principle Hessian directions (pHd). Therefore it can substitute for SIR, pHd, SAVE, or any linear combination of them at a theoretical level. Simulation study indicates that the proposed method may have consistently greater power than SIR, pHd, and SAVE.

Published by Elsevier B.V.

1. Introduction

The purpose of the regression of a univariate response y on a p-dimensional predictor vector $\mathbf{x}$ is to make inference on the conditional distribution of $y|\mathbf{x}$. Following Cook (1998b), $\mathbf{x}$ can be replaced by its standardized version

$$\mathbf{z} = [\Sigma_{\mathbf{x}}]^{-1/2}(\mathbf{x}-\mu_{\mathbf{x}}), \tag{1}$$

where $\mu_{\mathbf{x}}$ and $\Sigma_{\mathbf{x}}$ denote the mean and covariance matrix of $\mathbf{x}$, respectively, assuming non-singularity of $\Sigma_{\mathbf{x}}$.

The goal of *dimension reduction in regression* is to find out a $p \times d$ matrix γ such that

$$y \perp\!\!\!\perp \mathbf{z}|\gamma'\mathbf{z}, \tag{2}$$

where "$\perp\!\!\!\perp$" indicates independence. Then the p-dimensional $\mathbf{z}$ can be replaced by the d-dimensional vector $\gamma'\mathbf{z}$ without specifying any parametric model and without losing any information on predicting y. The column space Span$\{\gamma\}$ is called a *dimension reduction subspace*. The smallest applicable d is called the *dimension of the regression*.

Based on the inverse mean E$(\mathbf{z}|y)$, Li (1991a) proposed sliced inverse regression (SIR) for dimension reduction in regression. It is realized that SIR cannot recover the symmetric dependency (Li, 1991b; Cook and Weisberg, 1991). After SIR, many dimension reduction methods have been introduced. Sliced average variance estimate (SAVE) proposed by Cook and Weisberg (1991) and principle Hessian directions (pHd) proposed by Li (1992) are another two popular ones. Both pHd and SAVE refer to the second inverse moment, centered or non-centered. Compared with SAVE, pHd cannot detect certain

* Corresponding author. Tel.: +1 3124133748; fax: +1 3129961491.
E-mail address: jyang06@math.uic.edu (J. Yang).

0378-3758/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.jspi.2010.04.015

dependency hidden in the second moment (Yin and Cook, 2002; Ye and Weiss, 2003) and the linear dependency (Li, 1992; Cook, 1998a). Among those dimension reduction methods using only the first two inverse moments, SAVE seems to be the preferred one. Nevertheless, SAVE is not always the winner. For example, Ye and Weiss (2003) implied that a linear combination of SIR and pHd may perform better than SAVE in some cases. It is not surprising since Li (1991b) already suggested that a suitable combination of two different methods might sharpen the dimension reduction results. Ye and Weiss (2003) further proposed that a bootstrap method could be used to pick up the "best" linear combination of two known methods, as well as the dimension of the regression, in the sense of the variability of the estimators, although lower variability under the bootstrap procedure does not necessarily lead to a better estimator. Li and Wang (2007) pointed out that linear combinations of two known methods selected by the bootstrap criterion may not perform as well as a single new method, e.g. their directional regression (DR) method, even though the bootstrap one is computationally intensive.

This article aims to develop a new class of, instead of a single one, dimension reduction methods using only the first two inverse moments, as well as the corresponding large sample tests for the dimension of the regression and an efficient criterion for selecting a suitable candidate from the class. Theoretically, it can cover SIR, pHd, SAVE and their linear combinations. Practically, it can achieve higher power in recovering the dimension reduction subspace. In Section 2, we review the necessary dimension reduction context. In Section 3, we introduce a simple candidate matrix $M_{\mathbf{zz}'|y}$ which targets the entire inverse second moment subspace. It is indeed the candidate matrix of an intermediate method between pHd and SAVE. In Section 4, we propose a new class of dimension reduction methods called sliced inverse moment regression (SIMR), along with weighted chi-squared tests for the dimension of the regression. In Section 5, we use SIMR to analyze a simulated example and illustrate how to select a good candidate of SIMR. Simulation study shows that SIMR may have consistently greater power than SIR, pHd, and SAVE, as well as DR and another new method inverse regression estimator (Cook and Ni, 2005). In Section 6, a real example is used to illustrate how the proposed method works. It is implied that a class of dimension reduction methods, along with a suitable criterion for choosing a good one among them, may be preferable in practice to any single method. We conclude this article with discussion and proofs of the results presented.

2. Dimension reduction context

2.1. Central dimension reduction subspace (CDRS)

Cook (1994, 1996) introduced the notion of *central dimension reduction subspace* (CDRS), denoted by $S_{y|\mathbf{z}}$, which is the intersection of all dimension reduction subspaces. Under fairly weak restrictions, the CDRS $S_{y|\mathbf{z}}$ is still a dimension reduction subspace.

In this article, we always assume that $S_{y|\mathbf{z}}$ is a dimension reduction subspace and that the columns of γ is an orthonormal basis of $S_{y|\mathbf{z}}$. In practice, we usually first transform the original data $\{\mathbf{x}_i\}$ into their standardized version $\{\mathbf{z}_i\}$ by replacing $\Sigma_{\mathbf{x}}$ and $\mu_{\mathbf{x}}$ in (1) with their usual sample estimates $\hat{\Sigma}_{\mathbf{x}}$ and $\hat{\mu}_{\mathbf{x}}$. Then we can estimate $S_{y|\mathbf{x}}$ by

$$\hat{S}_{y|\mathbf{x}} = [\hat{\Sigma}_{\mathbf{x}}]^{-1/2}\hat{S}_{y|\mathbf{z}},$$

where $\hat{S}_{y|\mathbf{z}}$ is an estimate of $S_{y|\mathbf{z}}$. Therefore, the goal of dimension reduction in regression is to find out the dimension of the regression d and the CDRS $S_{y|\mathbf{z}} = \mathrm{Span}\{\gamma\}$.

Following Li (1991a) and Cook (1998b), we also assume: (1) $\mathrm{E}(\mathbf{z}|\gamma'\mathbf{z}) = P_\gamma\mathbf{z}$, where $P_\gamma = \gamma\gamma'$, known as the *linearity condition*; (2) $\mathrm{Var}(\mathbf{z}|\gamma'\mathbf{z}) = Q_\gamma$, where $Q_\gamma = \mathrm{I} - P_\gamma$, known as the *constant covariance condition*. These two conditions hold if $\mathbf{z}$ is normally distributed, although the normality is not necessary.

2.2. Candidate matrix

Ye and Weiss (2003) introduced the concept of *candidate matrix*, which is a $p \times p$ matrix A satisfying $A = P_\gamma A P_\gamma$. They showed that any eigenvector corresponding to any nonzero eigenvalue of A belongs to the CDRS $\mathrm{Span}\{\gamma\}$. Besides, the set of all candidate matrices, denoted by $\mathcal{M}$, is closed under scalar multiplication, transpose, addition, multiplication, and thus under linear combination and expectation.

They also showed that the matrices $[\mu_1(y)\mu_1(y)']$ and $[\mu_2(y) - \mathrm{I}]$ belong to $\mathcal{M}$ for all y, where $\mu_1(y) = \mathrm{E}(\mathbf{z}|y)$ and $\mu_2(y) = \mathrm{E}(\mathbf{z}\mathbf{z}'|y)$. They proved that the symmetric matrices that SIR, SAVE, and y-pHd estimate all belong to $\mathcal{M}$:

$$M_{\mathrm{SIR}} = \mathrm{Var}(\mathrm{E}(\mathbf{z}|y)) = \mathrm{E}[\mu_1(y)\mu_1(y)'],$$

$$M_{\mathrm{SAVE}} = \mathrm{E}[(\mathrm{I} - \mathrm{Var}(\mathbf{z}|y))^2] = \mathrm{E}([\mu_1(y)\mu_1(y)']^2 + [\mu_2(y) - \mathrm{I}]^2 - [\mu_1(y)\mu_1(y)'][\mu_2(y) - \mathrm{I}] - [\mu_2(y) - \mathrm{I}][\mu_1(y)\mu_1(y)']),$$

$$M_{y\text{-pHd}} = \mathrm{E}[(y - \mathrm{E}(y))\mathbf{z}\mathbf{z}'] = \mathrm{E}[y(\mu_2(y) - \mathrm{I})].$$

3. Candidate matrix $M_{\mathbf{zz}'|y}$

3.1. A simple candidate matrix

The matrices $[\mu_1(y)\mu_1(y)']$ and $[\mu_2(y)-\mathrm{I}]$ are actually two fundamental components of M_{SIR}, M_{SAVE}, and $M_{y\text{-pHd}}$ (see Section 2.2). M_{SIR} only involves the first component $[\mu_1(y)\mu_1(y)']$, while both M_{SAVE} and $M_{y\text{-pHd}}$ share the second component $[\mu_2(y)-\mathrm{I}]$. Realizing that this common feature may lead to the connection between SAVE and pHd, we investigate the behavior of the matrix $[\mu_2(y)-\mathrm{I}]$. To avoid the inconvenience due to $\mathrm{E}([\mu_2(y)-\mathrm{I}])=0$, we define

$$M_{\mathbf{zz}'|y}=\mathrm{E}([\mathrm{E}(\mathbf{zz}'-\mathrm{I}|y)]^2)=\mathrm{E}([\mu_2(y)-\mathrm{I}]^2).$$

Note that $M_{\mathbf{zz}'|y}$ takes a simpler form than the rescaled version of sirII (Li, 1991b, Remark R.3) while still keeping the theoretical comprehensiveness. It also appears as a component in one expression of the *directional regression* matrix G (Li and Wang, 2007, Eq. (4)). We choose its form as simple as possible for less complicated large sample test and potentially greater test power. To establish the relationship between $M_{y\text{-pHd}}$ and $M_{\mathbf{zz}'|y}$, we need:

Lemma 1. *Let M be a $p\times q$ random matrix defined on a probability space $(\Omega,\mathcal{F},\mathrm{P})$, then there exists an event $\Omega_0\in\mathcal{F}$ with probability* 1, *such that*,

$$\mathrm{Span}\{E(MM')\}=\mathrm{Span}\{M(\omega),\omega\in\Omega_0\}.$$

A similar result can also be found in Yin and Cook (2003, Proposition 2(i)). The lemma here is more general. By the definition of $M_{\mathbf{zz}'|y}$,

Corollary 1. $\mathrm{Span}\{M_{\mathbf{zz}'|y}\}=\mathrm{Span}\{[\mu_2(y)-\mathrm{I}],y\in\Omega(y)\}$, *where $\Omega(y)$ is the support of y.*

Based on Corollary 1, Ye and Weiss (2003)'s Lemma 3, and the fact that $[\mu_2(y)-\mathrm{I}]\in\mathcal{M}$ for all y, matrix $M_{\mathbf{zz}'|y}$ is in fact a candidate matrix too. Corollary 1 also implies a strong connection between $M_{y\text{-pHd}}$ and $M_{\mathbf{zz}'|y}$:

Corollary 2. $\mathrm{Span}\{M_{y\text{-pHd}}\}\subseteq\mathrm{Span}\{M_{\mathbf{zz}'|y}\}$.

To further understand the relationship between $M_{y\text{-pHd}}$ and $M_{\mathbf{zz}'|y}$, recall the *central k-th moment dimension reduction subspace* (Yin and Cook, 2003), $S^{(k)}_{y|\mathbf{z}}=\mathrm{Span}\{\eta^{(k)}\}$. The corresponding random vector $(\eta^{(k)})'\mathbf{z}$ contains all the available information about y from the first k conditional moments of $y|\mathbf{z}$. In other words, $y\perp\!\!\!\perp\{\mathrm{E}(y|\mathbf{z}),\ldots,\mathrm{E}(y^k|\mathbf{z})\}|(\eta^{(k)})'\mathbf{z}$. Similar to

$$\mathrm{Span}\{\mathrm{E}(y\mathbf{z}),\ldots,\mathrm{E}(y^k\mathbf{z})\}=\mathrm{Span}\{\mathrm{E}(y\mu_1(y)),\ldots,\mathrm{E}(y^k\mu_1(y))\}\subseteq S^{(k)}_{y|\mathbf{z}}\subseteq S_{y|\mathbf{z}},$$

the subspace $\mathrm{Span}\{\mathrm{E}(y[\mu_2(y)-\mathrm{I}]),\ldots,\mathrm{E}(y^k[\mu_2(y)-\mathrm{I}])\}$ is also contained in $S^{(k)}_{y|\mathbf{z}}$. Parallel to Yin and Cook (2002, Proposition 4), the result on $M_{\mathbf{zz}'|y}$ is:

Proposition 1. (a) *If y has finite support $\Omega(y)=\{a_0,\ldots,a_k\}$, then*

$$\mathrm{Span}\{M_{\mathbf{zz}'|y}\}=\mathrm{Span}\{\mathrm{E}[y^i(\mu_2(y)-\mathrm{I})],\ i=1,\ldots,k\}.$$

(b) *If y is continuous and $\mu_2(y)$ is continuous on y's support $\Omega(y)$, then*

$$\mathrm{Span}\{M_{\mathbf{zz}'|y}\}=\mathrm{Span}\{\mathrm{E}[y^i(\mu_2(y)-\mathrm{I})],\ i=1,2,\ldots\}.$$

According to Proposition 1 and Yin and Cook (2002, Proposition 4), the relationship between $\mathrm{E}[y(\mu_2(y)-\mathrm{I})]=M_{y\text{-pHd}}$ and $M_{\mathbf{zz}'|y}$ is fairly comparable with the relationship between $\mathrm{E}(y\mu_1(y))=\mathrm{E}(y\mathbf{z})$ and M_{SIR}. Both $\mathrm{E}(y\mathbf{z})$ and $M_{y\text{-pHd}}$ actually target the central mean (first moment) dimension reduction subspace (Cook and Li, 2002), while M_{SIR} and $M_{\mathbf{zz}'|y}$ target the central k-th moment dimension reduction subspace given any k, or equivalently the CDRS $S_{y|\mathbf{z}}$ as k goes to infinite. In order to understand the similarity from another perspective, recall the *inverse mean subspace* of $S_{y|\mathbf{z}}$ (Yin and Cook, 2002):

$$S_{\mathrm{E}(\mathbf{z}|y)}=\mathrm{Span}\{\mathrm{E}(\mathbf{z}|y),y\in\Omega(y)\}.$$

Similarly, we define the *inverse second moment subspace* of $S_{y|\mathbf{z}}$:

$$\mathrm{Span}\{\mathrm{E}(\mathbf{zz}'|y)-\mathrm{I},y\in\Omega(y)\}.$$

By definition, matrices M_{SIR} and $M_{\mathbf{zz}'|y}$ are designed to recover the entire inverse mean subspace and the entire inverse second moment subspace, respectively, while $\mathrm{E}(y\mathbf{z})$ and $M_{y\text{-pHd}}$ are only able to recover portions of those subspaces. We are therefore interested in combining matrices M_{SIR} and $M_{\mathbf{zz}'|y}$ because they are both comprehensive.

3.2. SAVE versus SIR and pHd

Ye and Weiss (2003) showed that

$$\mathrm{Span}\{M_{\mathrm{SIR}}\}\subseteq\mathrm{Span}\{M_{\mathrm{SAVE}}\}. \tag{3}$$

We then prove further the following proposition:

Proposition 2. $\mathrm{Span}\{M_{\mathrm{SAVE}}\} = \mathrm{Span}\{M_{\mathrm{SIR}}\} + \mathrm{Span}\{M_{\mathbf{zz}'|y}\}$.

A straightforward result following Proposition 2 and Corollary 2 is:

Corollary 3. $\mathrm{Span}\{M_{y\text{-pHd}}\}, \mathrm{Span}\{M_{\mathrm{SIR}}\}, \mathrm{Span}\{M_{\mathbf{zz}'|y}\} \subseteq \mathrm{Span}\{M_{\mathrm{SAVE}}\}$.

Corollary 3 explains why SAVE is able to provide better estimates of the CDRS than SIR and *y*-pHd in many cases.

4. Sliced inverse moment regression using weighted chi-squared tests

4.1. Sliced inverse moment regression

In order to simplify the candidate matrices using the first two inverse moments and still keep the comprehensiveness of SAVE, a natural idea is to combine $M_{\mathbf{zz}'|y}$ with M_{SIR} as follows:

$$\alpha M_{\mathrm{SIR}} + (1-\alpha) M_{\mathbf{zz}'|y} = \mathrm{E}(\alpha[\mu_1(y)\mu_1(y)'] + (1-\alpha)[\mu_2(y) - \mathrm{I}]^2),$$

where $\alpha \in (0,1)$. We call this matrix $M_{\mathrm{SIMR}}^{(\alpha)}$ and the corresponding dimension reduction method *sliced inverse moment regression* (SIMR or SIMR_α). Note that the combination here is simpler than the SIR_α method (Li, 1991b; Gannoun and Saracco, 2003) while retaining the least requirement on comprehensiveness. Actually, for any $\alpha \in (0,1)$, SIMR_α is as comprehensive as SAVE at a theoretical level based on the following proposition:

Proposition 3. $\mathrm{Span}\{M_{\mathrm{SIMR}}^{(\alpha)}\} = \mathrm{Span}\{M_{\mathrm{SAVE}}\}$, $\forall \alpha \in (0,1)$.

Combined with Corollary 3, we know that any linear combination of SIR, pHd and SAVE can be covered by SIMR_α:

Corollary 4. $\mathrm{Span}\{aM_{\mathrm{SIR}} + bM_{y\text{-pHd}} + cM_{\mathrm{SAVE}}\} \subseteq \mathrm{Span}\{M_{\mathrm{SIMR}}^{(\alpha)}\}$, *where a, b, and c are arbitrary real numbers.*

Note that the way of constructing SIMR_α makes it easier to develop a corresponding large sample test for the dimension of the regression (Section 4.3).

From now on, we assume that the data $\{(y_i, \mathbf{x}_i)\}_{i=1,\ldots,n}$ are i.i.d. from a population which has finite first four moments and conditional moments.

4.2. Algorithm for SIMR_α

Given i.i.d. sample $(y_1,\mathbf{x}_1),\ldots,(y_n,\mathbf{x}_n)$, first standardize $\mathbf{x}_i$ into $\hat{\mathbf{z}}_i$, sort the data by y, and divide the data into H slices with intraslice sample sizes n_h, $h=1,\ldots,H$. Second construct the intraslice sample means $\overline{(\mathbf{zz}')}_h$ and $\overline{\mathbf{z}}_h$:

$$\overline{(\mathbf{zz}')}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} \hat{\mathbf{z}}_{ih}\hat{\mathbf{z}}_{ih}',$$

$$\overline{\mathbf{z}}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} \hat{\mathbf{z}}_{ih},$$

where $\hat{\mathbf{z}}_{ih}$'s are predictors falling into slice h. Third calculate

$$\hat{M}_{\mathrm{SIMR}}^{(\alpha)} = \sum_{h=1}^{H} \hat{f}_h((1-\alpha)[\overline{(\mathbf{zz}')}_h - \mathrm{I}_p][\overline{(\mathbf{zz}')}_h - \mathrm{I}_p]' + \alpha[\overline{\mathbf{z}}_h][\overline{\mathbf{z}}_h]') = \hat{U}_n\hat{U}_n',$$

where $\hat{f}_h = n_h/n$ and

$$\hat{U}_n = (\ldots, \sqrt{1-\alpha}\,[\overline{(\mathbf{zz}')}_h - \mathrm{I}_p]\sqrt{\hat{f}_h}, \ldots, \ldots, \sqrt{\alpha}\,\overline{\mathbf{z}}_h\sqrt{\hat{f}_h}, \ldots)_{p\times(pH+H)}.$$

Finally calculate the eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ of $\hat{M}_{\mathrm{SIMR}}^{(\alpha)}$ and the corresponding eigenvectors $\hat{\gamma}_1,\ldots,\hat{\gamma}_p$. Then $\mathrm{Span}\{\hat{\gamma}_1,\ldots,\hat{\gamma}_d\}$ is an estimate of the CDRS $\mathrm{Span}\{\gamma\}$, where d is determined by the weighted chi-squared test described in the next section.

4.3. A weighted chi-squared test for SIMR_α

Define the population version of $\hat{U}_n$:

$$B = (\ldots, \sqrt{1-\alpha}\,[\mathrm{E}(\mathbf{zz}'|\tilde{y}=h) - \mathrm{I}_p]\sqrt{f_h}, \ldots, \sqrt{\alpha}\mathrm{E}(\mathbf{z}|\tilde{y}=h)\sqrt{f_h}, \ldots)$$
$$= ((\Gamma_{11})_{p\times d}, (\Gamma_{12})_{p\times(p-d)}) \begin{pmatrix} D_{d\times d} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} (\Gamma_{21}')_{d\times(pH+H)} \\ (\Gamma_{22}')_{(pH+H-d)\times(pH+H)} \end{pmatrix}, \quad (4)$$

where $\tilde{y}$ is a slice indicator with $\tilde{y} \equiv h$ for all observations falling into slice h, $f_h = P(\tilde{y} = h)$ is the population version of $\hat{f}_h$, and (4) is the *singular value decomposition* of B.

Denote $\tilde{U}_n = \sqrt{n}(\hat{U}_n - B)$. By the multivariate central limit theorem and the multivariate version of Slutsky's theorem, $\tilde{U}_n$ converges in distribution to a certain random $p \times (pH+H)$ matrix U as n goes to infinity (Gannoun and Saracco, 2003). Note that the singular values are invariant under right and left multiplication by orthogonal matrices. Based on Eaton and Tyler (1994, Theorems 4.1 and 4.2), the asymptotic distribution of the smallest $(p-d)$ singular values of $\sqrt{n}\hat{U}_n$ is the same as the asymptotic distribution of the corresponding singular values of the following $(p-d) \times (pH+H-d)$ matrix:

$$\sqrt{n}\Gamma'_{12}\hat{U}_n\Gamma_{22}. \tag{5}$$

Construct statistic

$$\hat{\Lambda}_d = n \sum_{h=d+1}^{p} \hat{\lambda}_h,$$

which is the sum of the squared smallest $(p-d)$ singular values of $\sqrt{n}\hat{U}_n$. Then the asymptotic distribution of $\hat{\Lambda}_d$ is the same as that of the sum of the squared singular values of (5). That is

$$n\,\mathrm{Trace}([\Gamma'_{12}\hat{U}_n\Gamma_{22}][\Gamma'_{12}\hat{U}_n\Gamma_{22}]') = n[\mathrm{Vec}(\Gamma'_{12}\hat{U}_n\Gamma_{22})]'[\mathrm{Vec}(\Gamma'_{12}\hat{U}_n\Gamma_{22})],$$

where $\mathrm{Vec}(A_{r\times c})$ denotes $(a'_1,\ldots,a'_c)'_{rc\times 1}$ for any matrix $A=(a_1,\ldots,a_c)$. By central limit theorem and Slutsky's theorem again,

$$\mathrm{Vec}(\tilde{U}_n) \xrightarrow{\mathcal{L}} N_{(p^2H+pH)}(0,V)$$

for some nonrandom $(p^2H+pH)\times(p^2H+pH)$ matrix V. Thus,

$$\sqrt{n}[\mathrm{Vec}(\Gamma'_{12}\hat{U}_n\Gamma_{22})] \xrightarrow{\mathcal{L}} N_{(p-d)(pH+H-d)}(0,W),$$

where $W = [\Gamma'_{22}\otimes\Gamma'_{12}]V[\Gamma'_{22}\otimes\Gamma'_{12}]'$ is a $(p-d)(pH+H-d) \times (p-d)(pH+H-d)$ matrix. Combined with Slutsky's theorem, it yields the following theorem:

Theorem 1. *The asymptotic distribution of $\hat{\Lambda}_d$ is the same as that of*

$$\sum_{i=1}^{(p-d)(pH+H-d)} \alpha_i K_i,$$

where the K_i's are independent χ^2_1 random variables, and α_i' s are the eigenvalues of the matrix W.

Clearly, a consistent estimate of W is needed for testing the dimension of the regression based on Theorem 1. The way we define $M^{(\alpha)}_{\mathrm{SIMR}}$ allows us to partition $\hat{U}_n$ into

$$\hat{U}_{n,1} = (\ldots,\sqrt{1-\alpha}[\overline{(\mathbf{z}\mathbf{z}')}_h - \mathrm{I}_p]\sqrt{\hat{f}_h},\ldots)_{p\times pH},$$
$$\hat{U}_{n,2} = (\ldots,\sqrt{\alpha}\overline{\mathbf{z}}_h\sqrt{\hat{f}_h},\ldots)_{p\times H}.$$

The asymptotic distribution of the matrix $\hat{U}_{n,2}$ has been fully explored by Bura and Cook (2001), resulting in a weighted chi-squared test for SIR. The similar techniques can also be applied on the matrix $\hat{U}_{n,1}$, and therefore the matrix $\hat{U}_n$ as a whole, although the details are much more complicated.

Define the population versions of $\hat{U}_{n,1}$ and $\hat{U}_{n,2}$,

$$B_1 = (\ldots,\sqrt{1-\alpha}[\mathrm{E}(\mathbf{z}\mathbf{z}'|\tilde{y}=h) - \mathrm{I}_p]\sqrt{f_h},\ldots)_{p\times pH},$$
$$B_2 = (\ldots,\sqrt{\alpha}\ \mathrm{E}(\mathbf{z}|\tilde{y}=h)\sqrt{f_h},\ldots)_{p\times H}.$$

Then $\hat{U}_n = (\hat{U}_{n,1},\hat{U}_{n,2})$, and $B=(B_1,B_2)$.

Let $f, \hat{f}$ and 1_H be $H \times 1$ vectors with elements f_h, $\hat{f}_h$ and 1, respectively; let G and $\hat{G}$ be $H\times H$ diagonal matrices with diagonal entries $\sqrt{f_h}$ and $\sqrt{\hat{f}_h}$, respectively; and let

$$\hat{F} = (\mathrm{I}_H - \hat{f}1'_H), \quad F = (\mathrm{I}_H - f1'_H),$$

$$\begin{pmatrix}(\Gamma'_{21})\\(\Gamma'_{22})\end{pmatrix} = \begin{pmatrix}(\Gamma'_{211})_{d\times pH} & (\Gamma'_{212})_{d\times H}\\(\Gamma'_{221})_{(pH+H-d)\times pH} & (\Gamma'_{222})_{(pH+H-d)\times H}\end{pmatrix}.$$

Finally, define four matrices

$$M = (\ldots,\mathrm{E}(\mathbf{x}|\tilde{y}=h),\ldots)_{p\times H},$$

$$N = (\ldots, \mathrm{E}(\mathbf{x}'|\tilde{y}=h), \ldots)_{1\times pH} = \mathrm{Vec}(M)',$$
$$O = (\ldots, \mathrm{E}(\mathbf{x}\mathbf{x}'|\tilde{y}=h), \ldots)_{p\times pH},$$
$$C = [O - M(\mathrm{I}_H \otimes \mu_{\mathbf{x}}') - \mu_{\mathbf{x}} N]_{p\times pH},$$

and their corresponding sample versions M_n, N_n, O_n, and C_n. By the central limit theorem,

$$\sqrt{n}\,\mathrm{Vec}([(C_n, M_n) - (C, M)]) \xrightarrow{\mathcal{L}} N_{(p^2H+pH)}(0, \Delta)$$

for a nonrandom $(p^2H+pH)\times(p^2H+pH)$ matrix Δ. As a result,

Theorem 2. *The covariance matrix W in Theorem 1 is*

$$W = (K\Gamma_{22})' \otimes (\Gamma_{12}'\Sigma_{\mathbf{x}}^{-1/2})\Delta(K\Gamma_{22}) \otimes (\Gamma_{12}'\Sigma_{\mathbf{x}}^{-1/2})',$$

where

$$K = \begin{pmatrix} \sqrt{1-\alpha}(FG) \otimes \Sigma_{\mathbf{x}}^{-1/2} & 0 \\ 0 & \sqrt{\alpha}\, FG \end{pmatrix}.$$

The only difficulty left now is to obtain a consistent estimate of Δ. By the central limit theorem,

$$\sqrt{n}\,\mathrm{Vec}([(O_n, M_n, \hat{\mu}_{\mathbf{x}}) - (O, M, \mu_{\mathbf{x}})]) \xrightarrow{\mathcal{L}} N_{(p^2H+pH+p)}(0, \Delta_0),$$

where Δ_0 is a nonrandom $(p^2H+pH+p)\times(p^2H+pH+p)$ matrix, with details shown in the Appendix. On the other hand,

$$\begin{aligned}\mathrm{Vec}(C_n, M_n) &= \begin{pmatrix} \mathrm{I}_{p^2H} & -\mathrm{I}_H \otimes \hat{\mu}_{\mathbf{x}} \otimes \mathrm{I}_p - \mathrm{I}_{pH} \otimes \hat{\mu}_{\mathbf{x}} & 0 \\ 0 & \mathrm{I}_{pH} & 0 \end{pmatrix} \mathrm{Vec}(O_n, M_n, \hat{\mu}_{\mathbf{x}}) \\ &= g([\mathrm{Vec}(O_n, M_n, \hat{\mu}_{\mathbf{x}})])\end{aligned}$$

for a certain mapping $g : \mathcal{R}^{(p^2H+pH+p)} \to \mathcal{R}^{(p^2H+pH)}$ such that

$$\mathrm{Vec}(C, M) = g([\mathrm{Vec}(O, M, \mu_{\mathbf{x}})]).$$

Thus the close form of Δ can be obtained by Cramér's (1946) theorem:

$$\Delta = [\dot{g}([\mathrm{Vec}(O, M, \mu_{\mathbf{x}})])]\Delta_0[\dot{g}([\mathrm{Vec}(O, M, \mu_{\mathbf{x}})])]', \tag{6}$$

where the $(p^2H+pH)\times(p^2H+pH+p)$ derivative matrix

$$\dot{g}[\mathrm{Vec}(O, M, \mu_{\mathbf{x}})] = \begin{pmatrix} \mathrm{I}_{p^2H} & -\mathrm{I}_H \otimes \mu_{\mathbf{x}} \otimes \mathrm{I}_p - \mathrm{I}_{pH} \otimes \mu_{\mathbf{x}} & \dot{g}_{13} \\ 0 & \mathrm{I}_{pH} & 0 \end{pmatrix} \tag{7}$$

with $\dot{g}_{13} = -(\ldots, \mathrm{I}_p \otimes \mathrm{E}(\mathbf{x}'|\tilde{y}=h), \ldots)' - \mathrm{Vec}(M) \otimes \mathrm{I}_p$.

In summary, to compose a consistent estimate of matrix W, one can (i) substitute the usual sample moments to get the sample estimate of Δ_0; (ii) estimate Δ by substituting the usual sample estimates for $\mathrm{E}(\mathbf{x}'|\tilde{y}=h)$, $\mu_{\mathbf{x}}$ and M in (6) and (7); (iii) obtain the usual sample estimates of Γ_{12} and Γ_{22} from the singular value decomposition of $\hat{U}_n$; (iv) substitute the usual sample estimates for F, G, $\Sigma_{\mathbf{x}}$, Γ_{12} and Γ_{22} in Theorem 2 to form an estimate of W. Note that both Δ and Δ_0 do not rely on α. This fact can save a lot of computational time when multiple α's need to be checked.

To approximate a linear combination of chi-squared random variables, one may use the statistic proposed by Satterthwaite (1941), Wood (1989), Satorra and Bentler (1994), or Bentler and Xie (2000). In the next applications, we will present tests based on Satterthwaite's statistic for illustration purpose.

4.4. Choosing optimal α

Ye and Weiss (2003) proposed a bootstrap method to pick up the "best" linear combination of two known methods in terms of variability of the estimated CDRS $\hat{S}_{y|\mathbf{z}}$. The bootstrap method works reasonably well with known dimension d of the regression, although less variability may occur with a wrong d (see Section 5 for an example). Another drawback is its computational intensity (Li and Wang, 2007).

Alternative criterion for "optimal" α is based on the weighted chi-squared tests developed for SIMR. When multiple tests with different α report the same dimension d, we simply pick up the α with the smallest p-value. Given that the true dimension d is detected, the last eigenvector $\hat{\gamma}_d$ added into the estimated CDRS with such an α is the most significant one among the candidates based on different α. In the mean time, the other eigenvectors $\hat{\gamma}_1, \ldots, \hat{\gamma}_{d-1}$ with selected α tend to be more significant than other candidates too. Based on simulation studies (Section 5), the performance of the p-value criterion is comparable with the bootstrap one with known d. The advantages of the former include that it is compatible with the weighted chi-squared tests and it requires much less computation.

When a model or an algorithm is specified for the data analysis, cross-validation could be used for choosing optimal α too, just like how people did for model selection. For example, see Hastie et al. (2009, Chapter 7). It will not be covered in this paper since we aim at model-free dimension reduction.

5. Simulation study

5.1. A simulated example

Let the response $y = 2z_1\varepsilon + z_2^2 + z_3$, where $(\mathbf{z}', \varepsilon)' = (z_1, z_2, z_3, z_4, \varepsilon)'$ are i.i.d from the $N_5(0, I_5)$ distribution. Then the true dimension of the regression is 3 and the true CDRS is spanned by $(1,0,0,0)'$, $(0,1,0,0)'$, and $(0,0,1,0)'$, that is, z_1, z_2 and z_3.

Theoretically, $M_{\text{SIR}} = \text{Diag}\{0,0,\ \text{Var}(\text{E}(z_3|y)),\ 0\}$, $M_{y\text{-pHd}} = \text{Diag}\{0,2,0,0\}$, and $M_{r\text{-pHd}} = \text{Diag}\{0,2,0,0\}$ have rank one and therefore are only able to find a one-dimensional proper subspace of the CDRS. The linear combination of any two of them suggested by Ye and Weiss (2003) can at most find a two-dimensional proper subspace of the CDRS. On the contrary, both SAVE and SIMR are able to recover the complete CDRS at a theoretical level.

5.2. A single simulation

We begin with a single simulation with sample size n=400. SIR, r-pHd, SAVE and SIMR are applied to the data. Number of slices H=10 are used for SIR, SAVE, and SIMR. The `R` package `dr` (Weisberg, 2002, 2009, Version 3.0.3) is used for SIR, r-pHd, SAVE, as well as their corresponding marginal dimension tests. SIMR_α with $\alpha = 0, 0.01, 0.05, 0.1-0.9$ paced by 0.1, 0.95, 0.99, 1 are applied.

For this typical simulation, SIR identifies only the direction $(0.018,\ 0.000,\ -\mathbf{0.999},\ -0.035)'$. It is roughly z_3, the linear trend. r-pHd identifies only the direction $(0.011,\ \mathbf{0.999},\ -0.038,\ -0.020)'$, which is roughly z_2, the quadratic component. As expected, SAVE works better. It identifies z_2 and z_1. However, the marginal dimension tests for SAVE (Shao et al., 2007) fail to detect the third predictor, z_3. The p-value of the corresponding test is 0.331.

Roughly speaking, SAVE with its marginal dimension test is comparable with $\text{SIMR}_{0.1}$ in this case. The comparison between SAVE and SIMR_α suggests that the failure of SAVE might due to its weights combining the first and second inverse moments. As α increases, SIMR_α with α between 0.3 and 0.8 all succeed in detecting all the three effective predictors z_1, z_2 and z_3. The CDRS estimated by those candidate matrices are similar to each other, which implies that the results with different α are fairly consistent. The major difference among SIMR_α is that the order of the detected predictors changes roughly from $\{z_2, z_1, z_3\}$ to $\{z_3, z_2, z_1\}$ as α increases from 0.3 to 0.8. As expected, SIMR_α is comparable with SIR if α is close to 1.

For this particular simulation, SIMR_α with α between 0.3 and 0.8 are first selected. If we know the true CDRS, the optimal α is the one minimizing the distance between the estimated CDRS and the true CDRS. Following Ye and Weiss (2003, p. 974), the three distance measures $\arccos(q)$, $1-q$, $1-r$ behave similarly and imply the same $\alpha = 0.6$ for this particular simulation. Since the true CDRS is unknown, bootstrap criterion and p-value criterion (Section 4.4) are applied instead.

The left panel of Fig. 1 shows the variability of bootstrapped estimated CDRS. Distance $1-r$ is used because it is comparable across different dimensions. The minimum variability is attained at d=3 and $\alpha = 0.6$, which happens to the

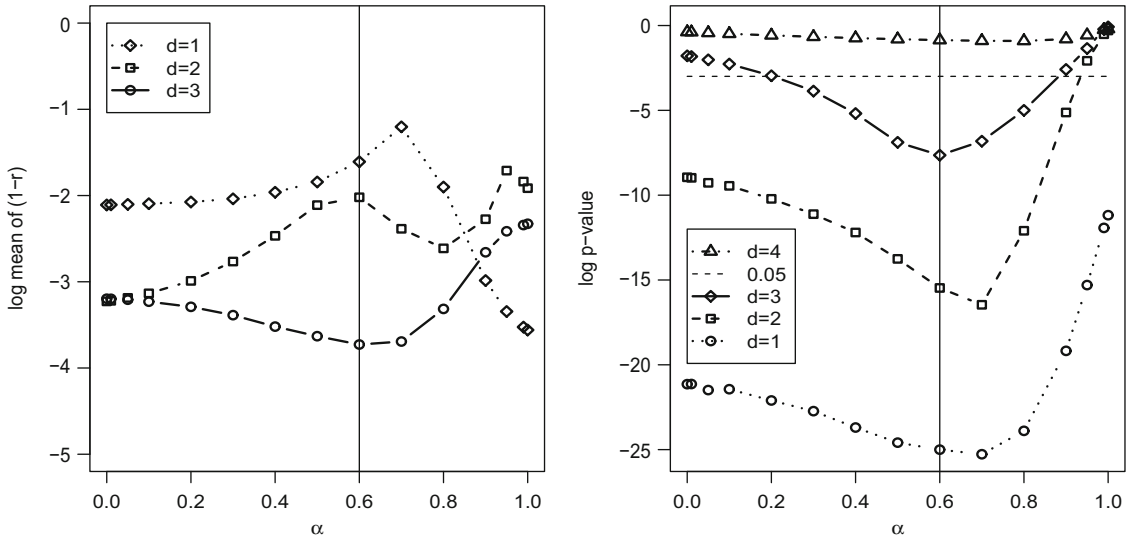

Fig. 1. Optimal α according to variability of 200 bootstrapped estimated CDRS (left panel, d=3 indicates the first three eigenvectors considered, and so on) or p-values of weighted chi-squared tests (right panel, d=3 indicates the test $d \le 2$ versus $d \ge 3$, and so on).

optimal one based on the truth. Another 200 simulations reveal that about 75% "optimal" α based on bootstrap fall in 0.5–0.6. SIMR with α chosen by bootstrap criterion attains $1-r=0.0086$ away from the true CDRS on average. Note that low variability not necessarily implies that the estimated CDRS is accurate. For example, $SIMR_1$ or SIR can only detect one direction z_3. However, the estimated one-dimensional CDRS is fairly stable under bootstrapping (see Fig. 1).

The right panel in Fig. 1 shows that the p-value criterion also picks up $\alpha=0.6$ for this single simulation (check the line $d=3$, which is the highest one that still goes below the significance level 0.05). Based on the same 200 simulations, about 80% of the "best" α selected by p-value criterion fall between 0.4 and 0.7. On average, SIMR with α selected by p-values attains $1-r=0.0082$, which is comparable with the bootstrap one.

5.3. Power analysis

We conduct 1000 independent simulations and summarize in Table 1 the empirical powers and sizes of the marginal dimension tests with significance level 0.05 for SIR, SAVE, r-pHd, and $SIMR_\alpha$ with α chosen by the p-value criterion. For illustration purpose, we omit the simulation results of y-pHd because there is little difference between y-pHd and r-pHd in this case. The empirical powers and sizes with significance level 0.01 are omitted too since their pattern is similar to Table 1.

In Table 1, the rows $d\le 0$, $d\le 1$, $d\le 2$ and $d\le 3$ indicate different null hypotheses. Following Bura and Cook (2001), the numerical entries in the rows $d\le 0$, $d\le 1$, and $d\le 2$ are empirical estimates of the powers of the corresponding tests, while the entries in the row $d\le 3$ are empirical estimates of the sizes of the tests.

As expected, SIR claims $d=1$ in most cases. r-pHd works a little better. At the significance level 0.05, r-pHd has about 30% chance to find out $d\ge 2$ (Table 1). At level 0.01, the chance shrinks to about 15%. Both SAVE and SIMR perform much better than SIR and pHd. Compared with SAVE, SIMR has consistently greater powers for the null hypotheses $d\le 0$, $d\le 1$ and $d\le 2$ across different choices of sample size, number of slices and significant level. For example, under the null hypothesis $d\le 2$ with sample size 400, the empirical powers of SIMR at level 0.05 are 0.939 under 5 slices and 0.943 under 10 slices, while the corresponding powers of SAVE are only 0.399 and 0.213, respectively (Table 1). Those differences become even bigger at level 0.01. The empirical sizes of SIMR are roughly under the nominal size 0.05 although they tend to be larger than the others.

For comparison purpose, the methods *inverse regression estimator* (IRE) (Cook and Ni, 2005; Wen and Cook, 2007; Weisberg, 2009) and *directional regression* (DR) (Li and Wang, 2007) are also applied. Roughly speaking, IRE performs similar to SIR in this example. Given that the truth dimension $d=3$ is known, both DR and SIMR are among the best in terms of mean$(1-r)$. For example, at $n=600$, DR achieves mean$(1-r)$ =0.0050 with $H=5$, 0.0053 with $H=10$, and 0.0059 with $H=15$, while SIMR's are 0.0048, 0.0046, and 0.0053. Nevertheless, the powers of the marginal tests for DR are between SAVE and SIMR in this case. Roughly speaking, DR's power tests are comparable with $SIMR_\alpha$'s with α between 0.2 and 0.3. For example, at $H=10$ and level 0.05, the empirical powers of DR against $d\le 2$ are 0.247 with $n=200$, 0.800 with $n=400$, and 0.974 with $n=600$.

Table 1
Empirical power and size of marginal dimension tests for SIR, SAVE, $SIMR_\alpha$ with α chosen by p-value criterion, and r-pHd, as well as mean of $1-r$ distances between estimated 3-Dim CDRS and true CDRS, based on 1000 simulations. (Significance level: 0.05; sample size: 200, 400, 600; number of slices: 5, 10, 15.)

	SIR			SAVE			$SIMR_\alpha$			r-pHd
Slice	5	10	15	5	10	15	5	10	15	–
$n=200$										
$d\le 0$	0.996	0.967	0.933	1.000	0.994	0.885	1.000	0.999	0.985	1.000
$d\le 1$	0.050	0.053	0.102	0.561	0.379	0.152	0.892	0.855	0.760	0.277
$d\le 2$	0.004	0.003	0.003	0.061	0.025	0.007	0.489	0.441	0.354	0.027
$d\le 3$	0.001	0.000	0.000	0.003	0.001	0.000	0.032	0.022	0.026	0.005
Mean$(1-r)$	0.124	0.127	0.119	0.045	0.060	0.077	0.033	0.033	0.039	0.111
$n=400$										
$d\le 0$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$d\le 1$	0.039	0.050	0.108	0.983	0.974	0.888	1.000	1.000	0.993	0.293
$d\le 2$	0.003	0.001	0.012	0.399	0.213	0.091	0.939	0.943	0.860	0.026
$d\le 3$	0.001	0.000	0.000	0.015	0.013	0.010	0.052	0.040	0.033	0.002
Mean$(1-r)$	0.127	0.129	0.120	0.016	0.025	0.038	0.009	0.009	0.011	0.109
$n=600$										
$d\le 0$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$d\le 1$	0.054	0.062	0.053	1.000	1.000	0.998	1.000	1.000	1.000	0.328
$d\le 2$	0.001	0.000	0.002	0.841	0.601	0.371	0.996	1.000	0.992	0.040
$d\le 3$	0.001	0.000	0.000	0.021	0.019	0.013	0.048	0.034	0.031	0.006
Mean$(1-r)$	0.123	0.123	0.125	0.008	0.010	0.016	0.005	0.005	0.005	0.108

Table 2
Ozone data: estimated CDRS by r-pHd, SAVE, $SIMR_0$, and $SIMR_{0.2}$ (H=10 for SAVE and SIMR).

First	Second	Third	Fourth
r-pHd			
−0.113	0.333	(0.183)	(−0.194)
−0.049	0.084	(−0.018)	(−0.012)
0.826	0.939	(−0.642)	(−0.030)
−0.551	−0.031	(0.745)	(0.981)
$SIMR_0$			
0.652	0.169	0.092	(0.125)
−0.025	−0.032	0.015	(0.026)
−0.662	−0.803	−0.645	(0.137)
−0.369	−0.571	0.758	(−0.982)
SAVE			
0.635	0.126	0.096	(−0.124)
−0.026	−0.031	0.015	(−0.026)
−0.665	−0.621	−0.664	(−0.143)
−0.392	−0.773	0.741	(0.981)
$SIMR_{0.2}$			
0.685	0.204	0.092	(−0.125)
−0.024	−0.031	0.015	(−0.026)
−0.653	−0.708	−0.653	(−0.141)
−0.322	−0.676	0.751	(0.982)

Note: "$(\cdot)$" indicates nonsignificant direction at level 0.05.

Among the six dimension reduction methods applied, SIMR is the most reliable one. Besides, the chi-squared tests for SIMR do not seem to be very sensitive to the numbers of slices. Nevertheless, we suggest that the number of slices should not be greater than 3–5% of the sample size based on the simulation results.

6. A real example: ozone data

To examine how SIMR works in practice, we consider a data set taken from Breiman and Friedman (1985). The response `Ozone` is the daily ozone concentration in parts per million, measured in Los Angeles basin, for 330 days in 1976. For illustration purpose, the dependence of `Ozone` on the following four predictors is studied next: `Height`, Vandenburg 500 millibar height in meters; `Humidity` in percents; `ITemp`, Inverse base temperature in degrees Fahrenheit; and `STemp`, Sandburg Air Force Base temperature in degrees Fahrenheit.

To meet both the linearity condition and the constant covariance condition, simultaneously power transformations on the predictors are estimated to improve the normality of their joint distribution. After replacing `Humidity`, `ITemp`, and `STemp` with $\texttt{Humidity}^{1.68}$, $\texttt{ITemp}^{1.25}$, and $\texttt{STemp}^{1.11}$, respectively, SIR, r-pHd, SAVE and SIMR are applied to the data. For SIR, SAVE, and SIMR, various numbers of slices are applied, and the results are fairly consistent. Here we only present the outputs based on H=8.

At significance level 0.05, SIR suggests the dimension of the regression d=1, while r-pHd claims d=2. Using the visualization tools described by Cook and Weisberg (1994) and Cook (1998b), the first pHd predictor appears to be somewhat symmetric about the response *Ozone*, and the second pHd predictor seems to be similar to the first SIR predictor, which are not shown in this article. The symmetric dependency explains why SIR is not able to find the first pHd predictor. The resulting inference based on pHd is therefore more reliable than the inference based on SIR.

When checking the predictors of SAVE, visual tools show a clear quadratic or even higher order polynomial dependency between the response and the first SAVE predictor. The second SAVE predictor is similar to the second pHd predictor, and the third SAVE predictor is similar to the first pHd predictor. Both SIR's and pHd's tests miss the first SAVE predictor.

Now apply SIMR to the ozone data. Bootstrap criterion picks up $\alpha = 0.2$ while p-value criterion suggests $\alpha = 0$. Nevertheless, both $SIMR_{0.2}$ and $SIMR_0$ lead to very similar estimated CDRS in this case (see Table 2). As expected, they recovers all the three SAVE predictors. Actually, those three estimated CDRS appear to be almost identical.

7. Discussion

$SIMR_\alpha$ and SAVE are theoretically equivalent since that the subspaces spanned by their underlying matrices are identical. Nevertheless, simulation study shows that $SIMR_\alpha$ with some chosen α may perform better than SAVE. The main reason is that SAVE is only a fixed combination of the first two inverse moments. The simulation example in Section 5

implies that any fixed combination cannot always be the winner. Apparently, $\mathrm{SIMR}_{0.6}$ cannot always be the winner either. For example, if the simulation example is changed to $y = 2z_1\varepsilon + z_2^2 + 0.1z_3$, SIMR_α with α closer to 1 will perform better. For practical use, multiple methods, as well as their combinations, should be tried and unified. SIMR_α with $\alpha \in (0,1)$ provide a simple solution to it.

As a conclusion, we propose SIMR using weighted chi-squared tests as an important class of dimension reduction methods, which should be routinely considered during the search for the central dimension reduction subspace and its dimension.

Acknowledgements

The authors thank Robert Weiss for comments on an earlier draft. They are also very grateful to Bing Li and Shaoli Wang for sharing their computer program.

Appendix

Proof of Lemma 1. By definition, $\mathrm{Span}\{\mathrm{E}(MM')\} \subseteq \mathrm{Span}\{M(\omega), \omega \in \Omega_0\}$, if $P(\Omega_0) = 1$. On the other hand, for any $v_{p\times 1} \neq 0$,

$$v'\mathrm{E}(M(\omega)M'(\omega)) = 0 \Rightarrow v'\mathrm{E}(M(\omega)M'(\omega))v = 0$$
$$\Rightarrow \mathrm{E}([v'M(\omega)][v'M(\omega)]') = 0 \Rightarrow [v'M(\omega)] \equiv 0, \text{ with probability 1.}$$

Since $\{v : v'\mathrm{E}(MM') = 0\}$ only has finite dimension, there exists an Ω_0 with probability 1, such that,

$$\dim(\mathrm{Span}\{\mathrm{E}(M(\omega)M'(\omega))\}) \geq \dim(\mathrm{Span}\{M(\omega), \omega \in \Omega_0\}).$$

Thus, $\mathrm{Span}\{\mathrm{E}(M(\omega)M'(\omega))\} = \mathrm{Span}\{M(\omega), \omega \in \Omega_0\}$. □

Proof of Corollary 2.

$$\mathrm{Span}\{M_{y-\mathrm{pHd}} = \mathrm{E}[y(\mu_2(y) - \mathrm{I})]\} \subseteq \mathrm{Span}\{[\mu_2(y) - \mathrm{I}], \forall y\} = \mathrm{Span}\{M_{\mathbf{zz}'|y}\}. \quad \square$$

Proof Proposition 1. Define $\mu_i = \mathrm{E}[(\mathbf{zz}' - \mathrm{I})|y = a_i] = \mathrm{E}(\mathbf{zz}'|y = a_i) - \mathrm{I}$ and $f_i = \Pr(y = a_i)$ for $i = 0, \ldots, k$, then $\Sigma_{i=0}^k f_i = 1$ and $\Sigma_{i=0}^k f_i\mu_i = \mathrm{E}((\mathbf{zz}' - \mathrm{I})) = 0$. The rest of the steps follow the exactly same proof as in Yin and Cook (2002, A.3. Proposition 4). □

Proof of Proposition 2:. By Lemma 1,

$$\begin{aligned}\mathrm{Span}\{M_{\mathrm{SAVE}}\} &= \mathrm{Span}\{[\mu_1(y)\mu_1(y)' + (\mu_2(y) - \mathrm{I})], \forall y\} \subseteq \mathrm{Span}\{\mu_1(y), \forall y\} + \mathrm{Span}\{(\mu_2(y) - \mathrm{I}), \forall y\} \\ &= \mathrm{Span}\{M_{\mathrm{SIR}}\} + \mathrm{Span}\{M_{\mathbf{zz}'|y}\} \subseteq \mathrm{Span}\{M_{\mathrm{SIR}}\} + [\mathrm{Span}\{\mu_1(y)\mu_1(y)' + (\mu_2(y) - \mathrm{I}), \forall y\} \\ &\quad + \mathrm{Span}\{\mu_1(y), \forall y\}] \subseteq \mathrm{Span}\{M_{\mathrm{SIR}}\} + \mathrm{Span}\{M_{\mathrm{SAVE}}\} + \mathrm{Span}\{M_{\mathrm{SIR}}\} \\ &= \mathrm{Span}\{M_{\mathrm{SAVE}}\}. \quad \square\end{aligned}$$

Proof of Proposition 3. By Lemma 1,

$$\begin{aligned}\mathrm{Span}\{M_{\mathrm{SIMR}}^{(\alpha)}\} &= \mathrm{Span}\{(\mu_1(y), [\mu_2(y) - \mathrm{I}]), \forall y\} \\ &= \mathrm{Span}\{\mu_1(y), \forall y\} + \mathrm{Span}\{[\mu_2(y) - \mathrm{I}], \forall y\} \\ &= \mathrm{Span}\{M_{\mathrm{SIR}}\} + \mathrm{Span}\{M_{\mathbf{zz}'|y}\} \\ &= \mathrm{Span}\{M_{\mathrm{SAVE}}\}. \quad \square\end{aligned}$$

Proof of Theorem 2. Actually, $B = \Sigma_{\mathbf{x}}^{-1/2}(C, M)K$,

$$\hat{U}_n = \hat{\Sigma}_{\mathbf{x}}^{-1/2}(C_n, M_n)\begin{pmatrix} \sqrt{1-\alpha}\ (\hat{F}\hat{G}) \otimes \hat{\Sigma}_{\mathbf{x}}^{-1/2} & 0 \\ 0 & \sqrt{\alpha}\hat{F}\hat{G} \end{pmatrix}.$$

Note that $(\Gamma_{12}'B_1, \Gamma_{12}'B_2) = 0_{(p-d)\times(pH+H)}$, $B_1\Gamma_{221} + B_2\Gamma_{222} = 0_{p\times(pH+H-d)}$, $\mathrm{Span}\{C'\Sigma_{\mathbf{x}}^{-1/2}\Gamma_{12}\} \subseteq \mathrm{Span}\{1_H \otimes \mathrm{I}_p\}$, $\mathrm{Span}\{M'\Sigma_{\mathbf{x}}^{-1/2}\Gamma_{12}\} \subseteq \mathrm{Span}\{1_H\}$, $1_H'\hat{F} = 0, 1_H'F = 0$. Writing $\hat{\mathrm{I}}_p = \hat{\Sigma}_{\mathbf{x}}^{-1/2}\Sigma_{\mathbf{x}}^{1/2}$,

$$\begin{aligned}&\sqrt{n}\Gamma_{12}'\hat{U}_n\Gamma_{22} \\ &= \sqrt{n}\Gamma_{12}'\hat{U}_{n,1}\Gamma_{221} + \sqrt{n}\Gamma_{12}'\hat{U}_{n,2}\Gamma_{222} \\ &= \sqrt{1-\alpha}\ \sqrt{n}\Gamma_{12}'(\hat{\mathrm{I}}_p - \mathrm{I}_p + \mathrm{I}_p)\Sigma_{\mathbf{x}}^{-1/2}(C_n - C + C)[(\hat{F}\hat{G} - FG + FG) \otimes \mathrm{I}_p] \\ &\quad \times(\mathrm{I}_H \otimes \Sigma_{\mathbf{x}}^{-1/2})[\mathrm{I}_H \otimes (\hat{\mathrm{I}}_p' - \mathrm{I}_p + \mathrm{I}_p)]\Gamma_{221} + \sqrt{\alpha}\sqrt{n}\Gamma_{12}'(\hat{\mathrm{I}}_p - \mathrm{I}_p + \mathrm{I}_p) \times \Sigma_{\mathbf{x}}^{-1/2}(M_n - M + M)(\hat{F}\hat{G} - FG + FG)\Gamma_{222} \\ &= \sqrt{1-\alpha}\ \sqrt{n}\Gamma_{12}'\Sigma_{\mathbf{x}}^{-1/2}(C_n - C)[FG \otimes \mathrm{I}_p](\mathrm{I}_H \otimes \Sigma_{\mathbf{x}}^{-1/2})\Gamma_{221} + \sqrt{\alpha}\ \sqrt{n}\Gamma_{12}'\Sigma_{\mathbf{x}}^{-1/2}(M_n - M)FG\Gamma_{222} + O_p(n^{-1/2}) \\ &= \sqrt{n}\Gamma_{12}'\Sigma_{\mathbf{x}}^{-1/2}[(C_n, M_n) - (C, M)]K\Gamma_{22} + O_p(n^{-1/2}).\end{aligned}$$

Therefore, the asymptotic distribution of $\Gamma_{12}'\hat{U}_n\Gamma_{22}$ is determined only by the asymptotic distribution of (C_n, M_n).

The detail of Δ_0, $(p^2H+pH+p)\times(p^2H+pH+p)$:

$$\Delta_0=\begin{pmatrix}\Delta_0^{1,1} & \Delta_0^{1,2} & \Delta_0^{1,3}\\ \Delta_0^{2,1} & \Delta_0^{2,2} & \Delta_0^{2,3}\\ \Delta_0^{3,1} & \Delta_0^{3,2} & \Delta_0^{3,3}\end{pmatrix},$$

where

$\Delta_0^{1,1}=\mathrm{diag}\{\ldots,\mathrm{Cov}(\mathrm{Vec}(\mathbf{x}\mathbf{x}')|\tilde{y}=h)/f_h,\ldots\},p^2H\times p^2H;\Delta_0^{2,1}=\mathrm{diag}\{\ldots,\mathrm{Cov}(\mathbf{x},\mathrm{Vec}(\mathbf{x}\mathbf{x}')|\tilde{y}=h)/f_h,\ldots\},pH\times p^2H;$
$\Delta_0^{2,2}=\mathrm{diag}\{\ldots,\mathrm{Cov}(\mathbf{x}|\tilde{y}=h)/f_h,\ldots\},pH\times pH;\Delta_0^{3,1}=[\ldots,\mathrm{Cov}(\mathbf{x},\mathrm{Vec}(\mathbf{x}\mathbf{x}')|\tilde{y}=h),\ldots],p\times p^2H;$-
$\Delta_0^{3,2}=[\ldots,\mathrm{Cov}(\mathbf{x}|\tilde{y}=h),\ldots],p\times pH;\Delta_0^{3,3}=\Sigma_{\mathbf{x}},p\times p;\Delta_0^{1,2}=(\Delta_0^{2,1})';\Delta_0^{1,3}=(\Delta_0^{3,1})';\Delta_0^{2,3}=(\Delta_0^{3,2})'.$ □

References

Bentler, P.M., Xie, J., 2000. Corrections to test statistics in principal Hessian directions. Statist. Probab. Lett. 47, 381–389.

Breiman, L., Friedman, J., 1985. Estimating optimal transformations for multiple regression and correlation. J. Amer. Statist. Assoc. 80, 580–597.

Bura, E., Cook, R.D., 2001. Extending sliced inverse regression: the weighted chi-squared test. J. Amer. Statist. Assoc. 96, 996–1003.

Cook, R.D., 1994. Using dimension-reduction subspaces to identify important inputs in models of physical systems. In: Proceedings of the Section on Physical and Engineering Sciences. American Statistical Association, Alexandria, VA, pp. 18–25.

Cook, R.D., 1996. Graphics for regressions with a binary response. J. Amer. Statist. Assoc. 91, 983–992.

Cook, R.D., 1998a. Principal Hessian directions revisited (with discussion). J. Amer. Statist. Assoc. 93, 84–100.

Cook, R.D., 1998b. Regression Graphics, Ideas for Studying Regressions through Graphics. Wiley, New York.

Cook, R.D., Li, B., 2002. Dimension reduction for conditional mean in regression. Ann. Statist. 30, 455–474.

Cook, R.D., Ni, L., 2005. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. J. Amer. Statist. Assoc. 100, 410–428.

Cook, R.D., Weisberg, S., 1991. Discussion of 'sliced inverse regression for dimension reduction'. J. Amer. Statist. Assoc. 86, 328–332.

Cook, R.D., Weisberg, S., 1994. An Introduction to Regression Graphics. Wiley, New York.

Cramér, H., 1946. Mathematical Methods of Statistics. Princeton University Press, Princeton.

Eaton, M.L., Tyler, D.E., 1994. The asymptotic distributions of singular values with applications to canonical correlations and correspondence analysis. J. Multivariate Anal. 50, 238–264.

Gannoun, A., Saracco, J., 2003. Asymptotic theory for SIR_α method. Statist. Sinica 13, 297–310.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Li, B., Wang, S., 2007. On directional regression for dimension reduction. J. Amer. Statist. Assoc. 102, 997–1008.

Li, K.-C., 1991a. Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc. 86, 316–327.

Li, K.-C., 1991b. Rejoinder to 'sliced inverse regression for dimension reduction'. J. Amer. Statist. Assoc. 86, 337–342.

Li, K.-C., 1992. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. J. Amer. Statist. Assoc. 87, 1025–1039.

Satorra, A., Bentler, P.M., 1994. Corrections to test statistics and standard errors in covariance structure analysis. In: von Eye, A., Clogg, C.C. (Eds.), Latent Variables Analysis Applications for Developmental Research. Sage, Newbury Park, CA, pp. 399–419.

Satterthwaite, F.E., 1941. Synthesis of variance. Psychometrika 6, 309–316.

Shao, Y., Cook, R.D., Weisberg, S., 2007. Marginal tests with sliced average variance estimation. Biometrika 94, 285–296.

Weisberg, S., 2002. Dimension reduction regression in R. J. Statist. Software 7. Available from ⟨http://www.jstatsoft.org⟩.

Weisberg, S., 2009. The dr package. Available from ⟨http://www.r-project.org⟩.

Wen, X., Cook, R.D., 2007. Optimal sufficient dimension reduction in regressions with categorical predictors. J. Statist. Inf. Plann. 137, 1961–1979.

Wood, A., 1989. An F-approximation to the distribution of a linear combination of chi-squared random variables. Commun. Statist. Part B Simul. Comput. 18, 1439–1456.

Yin, X., Cook, R.D., 2002. Dimension reduction for the conditional k th moment in regression. J. Roy. Statist. Soc. Ser. B 64, 159–175.

Yin, X., Cook, R.D., 2003. Estimating central subspaces via inverse third moments. Biometrika 90, 113–125.

Ye, Z., Weiss, R.E., 2003. Using the bootstrap to select one of a new class of dimension reduction methods. J. Amer. Statist. Assoc. 98, 968–979.