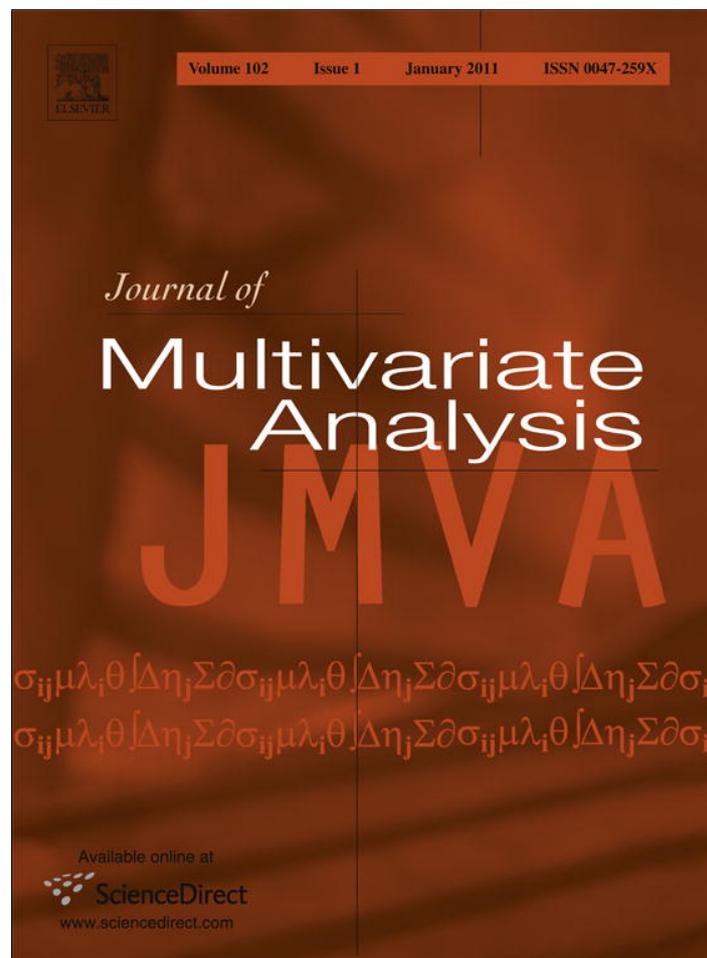


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Dimension estimation in sufficient dimension reduction: A unifying approach

E. Bura^{a,b,*}, J. Yang^c^a Department of Statistics, George Washington University, Washington, DC 20052, United States^b Vertex Pharmaceuticals, Inc., Cambridge, MA 02139, United States^c Rho, Inc., 199 Wells Avenue, Suite 302, Newton, MA 02459, United States

ARTICLE INFO

Article history:

Received 8 July 2009

Available online 20 August 2010

AMS subject classifications:

62B05

62H12

62H15

62H10

62J99

62G99

Keywords:

Random matrix

Chi-square and weighted chi-square tests

Dimension reduction

SIR

SAVE

ABSTRACT

Sufficient Dimension Reduction (SDR) in regression comprises the estimation of the dimension of the smallest (central) dimension reduction subspace and its basis elements. For SDR methods based on a kernel matrix, such as SIR and SAVE, the dimension estimation is equivalent to the estimation of the rank of a random matrix which is the sample based estimate of the kernel. A test for the rank of a random matrix amounts to testing how many of its eigen or singular values are equal to zero. We propose two tests based on the smallest eigen or singular values of the estimated matrix: an asymptotic weighted chi-square test and a Wald-type asymptotic chi-square test. We also provide an asymptotic chi-square test for assessing whether elements of the left singular vectors of the random matrix are zero. These methods together constitute a unified approach for all SDR methods based on a kernel matrix that covers estimation of the central subspace and its dimension, as well as assessment of variable contribution to the lower-dimensional predictor projections with variable selection, a special case. A small power simulation study shows that the proposed and existing tests, specific to each SDR method, perform similarly with respect to power and achievement of the nominal level. Also, the importance of the choice of the number of slices as a tuning parameter is further exhibited.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

This paper is concerned with providing a unifying approach to sufficient dimension reduction (SDR) methodology for estimating the dimension of a regression, even though our results have wider application.

The estimation of the rank of a random matrix is the central problem in all SDR methods based on a kernel matrix. Table 1 lists several SDR methods and their respective kernel matrices. Under the assumption that a root- n consistent estimator exists for an unobservable random matrix, several tests for its rank have been proposed. Gill and Lewbell [19], and Cragg and Donald [15] used a rank test based on the Lower-Diagonal-Upper triangular (LDU) decomposition. Their test has the advantage of possessing a limiting chi-square distribution, but tends to be overly conservative with type I error close to zero when the sample size is small [30]. Cragg and Donald [16] proposed another test based on a minimum chi-square criterion. The procedure needs to minimize an objective function numerically, which is often very difficult and requires the knowledge of the rank of the asymptotic variance of the estimator. Robin and Smith [31] obtained a weighted chi-square test for the rank without making such an assumption. Their test statistic is a variant of Anderson's [1] likelihood ratio statistic for the rank of a regression coefficient matrix in a multivariate normal linear model that is a functional of certain characteristic roots of a

* Corresponding author at: Department of Statistics, George Washington University, DC 20052, United States.
E-mail addresses: ebura@gwu.edu (E. Bura), Jiao_Yang@rhoworld.com (J. Yang).

Table 1
SDR Kernel matrices and their estimates.

Method	\mathbf{M}	$\hat{\mathbf{M}}$
SIR	$\text{Cov}(E(\mathbf{Z} Y))$	$\sum_{h=1}^H \hat{p}_h \bar{\mathbf{Z}}_h \bar{\mathbf{Z}}_h^T$
PIR	$E(\mathbf{Z} Y)$	$\hat{\mathbf{B}}_n = (\mathbf{F}_n^T \mathbf{F}_n)^{-1} \mathbf{F}_n^T \mathbf{Z}_n$
SAVE	$E(\mathbf{I}_p - \text{Var}(\mathbf{Z}_h))^2$	$\sum_{h=1}^H \hat{p}_h (\mathbf{I}_p - \widehat{\text{Var}}(\mathbf{Z}_h))^2$
pHd	$E((Y - E(Y))\mathbf{Z}\mathbf{Z}^T)$	$\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}) \mathbf{Z}_i \mathbf{Z}_i^T$
SIRII	$E(\text{Var}(\mathbf{Z} Y) - E(\text{Var}(\mathbf{Z} Y)))^2$	$\sum_{h=1}^H \hat{p}_h (\widehat{\text{Var}}(\mathbf{Z}_h) - \bar{\text{Var}}(\mathbf{Z}_h))^2$
SCR	$\Sigma_x^{-1/2} \mathbf{H}(c) \Sigma_x^{-1/2}$	$\widehat{\Sigma}_x^{-1/2} \widehat{\mathbf{H}}(c) \widehat{\Sigma}_x^{-1/2}$
CMS	$(\beta_{YZ}, \Sigma_{YZ} \beta_{YZ}, \dots, \Sigma_{YZ}^{p-1} \beta_{YZ})$	$(\hat{\beta}_{YZ}, \widehat{\Sigma}_{YZ} \hat{\beta}_{YZ}, \dots, \widehat{\Sigma}_{YZ}^{p-1} \hat{\beta}_{YZ})$
DR	$2E(E^2(\mathbf{Z}\mathbf{Z}^T - \mathbf{I}_p Y)) + 2E^2(E(\mathbf{Z} Y)E(\mathbf{Z}^T Y))$ $+ 2E(E(\mathbf{Z}^T Y)E(\mathbf{Z} Y)) \times E(E(\mathbf{Z} Y)E(\mathbf{Z}^T Y))$	$2 \sum_{h=1}^H \hat{p}_h E_n^2(\mathbf{Z}_h \mathbf{Z}_h^T - \mathbf{I}_p) + 2 \left(\sum_{h=1}^H E_n(\mathbf{Z}_h) E_n(\mathbf{Z}_h^T) \right)^2$ $+ 2 \sum_{h=1}^H \hat{p}_h E_n(\mathbf{Z}_h^T) E_n(\mathbf{Z}_h) \times \sum_{h=1}^H \hat{p}_h E_n(\mathbf{Z}_h) E_n(\mathbf{Z}_h^T)$
MAVE	$E(\nabla g(\mathbf{X}) \nabla^T g(\mathbf{X}))$	$\frac{1}{n} \sum_{i=1}^n \hat{\mathbf{b}}_i^T \hat{\mathbf{b}}_i$

matrix quadratic form. The latter is a quadratic form of the estimated matrix and two positive definite weighting matrices. The major disadvantage of this approach is that each application requires the selection of the two weighting matrices and the results of the test depend critically on this choice and on their interaction with the random matrix.

Most if not all tests for the rank of a random matrix are based on the fact that the rank equals the number of its non-zero eigen or singular values. The tests we propose here also fall within this class. The novelty of the proposed methods is that the only requirement is the estimate of the random matrix be unbiased and asymptotically normal with finite asymptotic second moments. No other restrictions, such as on the multiplicity of the singular values of the random matrix, nor other external quantities, such as weighting matrices, are required. In this context, we propose two tests based on the smallest eigen or singular values of the estimated matrix in Section 2. The first is an asymptotically weighted chi-square test based on a result by Eaton and Tyler [18] for the distribution of the singular values of a random matrix. From an application point of view, the second may be more important as it is easy to compute a Wald-type asymptotically chi-square test.

We also adjust and apply the asymptotic chi-square test for testing whether components of the elements of the basis of the column space of the random matrix are zero that was developed by Bura and Pfeiffer [7] to the context of general SDR in Section 3. This leads to a test for variable contribution in linear projections of the predictor vector in Section 4.2. When the variables whose contribution is tested are the same in all linear projections, this is a test for variable selection.

Dimension Reduction falls within the realm of Random Matrix Theory/Analysis as its estimation target is typically a random matrix. For example, in a regression context with response Y , Sufficient Dimension Reduction (SDR, [9]) is based on the idea that the p -dimensional predictor vector \mathbf{X} can be replaced by a smaller number of linear combinations of the predictors whose coefficients comprise basis elements of a dimension reduction subspace spanned by the columns of a kernel matrix. We use the results of Sections 2 and 3 to develop an umbrella theory for all kernel-matrix-based sufficient dimension reduction methods that generalize and unify the previous results in Section 4. We discuss SDR kernel matrices and methodology in detail in Section 4. As an aside, we also derive a straightforward proof of the asymptotic normality of the SAVE [14] kernel matrix, which was lacking from the SDR literature.

A power simulation study comparing the two proposed tests and the existing tests for SIR and SAVE is carried out in Section 5. We conclude with a discussion in Section 6. All theorem and lemma proofs are given in Appendix A.

2. Estimating the rank of a random matrix

To estimate the rank k of a random $p \times q$ matrix \mathbf{M} we consider the sequential testing of hypotheses of the form

$$H_0 : \text{rank}(\mathbf{M}) = j \quad \text{versus} \quad H_1 : \text{rank}(\mathbf{M}) > j \tag{1}$$

starting with $j = 0$. The smallest value of j for which the null is not rejected, for a fixed α level, is the estimate of the rank k of \mathbf{M} .

Let $k = \text{rank}(\mathbf{M}) \leq \min(p, q)$. The singular value decomposition (SVD) of \mathbf{M} is

$$\mathbf{M} = \mathbf{U}^T \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{R}, \tag{2}$$

where the orthogonal matrix $\mathbf{U}^T = (\mathbf{U}_1, \mathbf{U}_0)$ is $p \times p$ with $\mathbf{U}_1 : p \times k$, $\mathbf{U}_0 : p \times (p - k)$, $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$, is a diagonal matrix of the descending singular values of \mathbf{M} , $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k > 0$, and $\mathbf{R}^T = (\mathbf{R}_1, \mathbf{R}_0)$ is orthogonal with $\mathbf{R}_1 : q \times k$, $\mathbf{R}_0 : q \times (q - k)$. The k left singular vectors $\mathbf{U}_1 = (\mathbf{u}_1, \dots, \mathbf{u}_k)$ of \mathbf{M} , which correspond to its k non-zero singular values $\lambda_1 \geq \dots \geq \lambda_k > 0$, span $\mathcal{S}(\mathbf{M})$.

Let $\widehat{\mathbf{M}}$ denote an estimate of \mathbf{M} based on a random sample of size n . Let us assume that the estimator $\widehat{\mathbf{M}}$ satisfies

$$\sqrt{n} \operatorname{vec}(\widehat{\mathbf{M}} - \mathbf{M}) \xrightarrow{\mathcal{D}} N_{pq}(0, \mathbf{V}). \tag{3}$$

The SVD of $\widehat{\mathbf{M}}$ is

$$\widehat{\mathbf{M}} = \widehat{\mathbf{U}}^T \widehat{\mathbf{D}} \widehat{\mathbf{R}} = \widehat{\mathbf{U}}^T \begin{pmatrix} \widehat{\mathbf{D}}_1 & 0 \\ 0 & \widehat{\mathbf{D}}_0 \end{pmatrix} \widehat{\mathbf{R}} \tag{4}$$

with $\widehat{\mathbf{U}}^T = (\widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_0)$, $\widehat{\mathbf{R}}^T = (\widehat{\mathbf{R}}_1, \widehat{\mathbf{R}}_0)$, where the partitions conform to those in the SVD of \mathbf{M} in (2). In terms of the singular values of $\widehat{\mathbf{M}}$, $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_{\min(p,q)}$, the singular value decomposition in (4) can be written as follows:

$$\begin{aligned} \widehat{\mathbf{D}}_1 &= \widehat{\mathbf{U}}_1^T \widehat{\mathbf{M}} \widehat{\mathbf{R}}_1 = \operatorname{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_k) : k \times k \\ \widehat{\mathbf{D}}_0 &= \widehat{\mathbf{U}}_0^T \widehat{\mathbf{M}} \widehat{\mathbf{R}}_0 = (\hat{d}_{ij}^{(0)}) : (p-k) \times (q-k) \end{aligned} \tag{5}$$

where $\hat{d}_{ij}^{(0)} = 0$ for $i \neq j$ and $\hat{d}_{ii}^{(0)} = \hat{\lambda}_{k+i}$ for $i = 1, \dots, m-k$, $m = \min(p, q)$. When $k = \operatorname{rank}(\mathbf{M})$, $\widehat{\mathbf{D}}_0$ tends to zero as the sample size increases. The development of the proposed two tests for the rank of \mathbf{M} is based on this fact.

2.1. Weighted chi-square test

Let

$$\Lambda_1(k) = n \operatorname{trace}(\widehat{\mathbf{D}}_0^T \widehat{\mathbf{D}}_0) = n \operatorname{vec}(\widehat{\mathbf{D}}_0)^T \operatorname{vec}(\widehat{\mathbf{D}}_0) = n \sum_{i=k+1}^{\min(p,q)} \hat{\lambda}_i^2 \tag{6}$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_{\min(p,q)} \geq 0$ are the singular values of $\widehat{\mathbf{M}}$.

Theorem 1. Assume $\operatorname{rank}(\mathbf{M}) = k$ and that $\widehat{\mathbf{M}}$ satisfies (3). Then,

$$\Lambda_1(k) \xrightarrow{\mathcal{D}} \sum_{i=1}^s w_i \chi_i^2, \tag{7}$$

where the χ_i^2 's are independent chi-square random variables each with 1 degree of freedom, and $w_1 \geq w_2 \geq \dots \geq w_s$ are the ordered eigenvalues of $\mathbf{Q} = (\mathbf{R}_0^T \otimes \mathbf{U}_0^T) \mathbf{V} (\mathbf{R}_0 \otimes \mathbf{U}_0)$, with $s = \min(\operatorname{rank}(\mathbf{V}), (p-k)(q-k))$.

Let $\widehat{\mathbf{V}}$ be a consistent estimate of \mathbf{V} . Also, let \hat{w}_i , $i = 1, 2, \dots, s$, $s = \min(\operatorname{rank}(\mathbf{V}), (p-k)(q-k)) = \min(\operatorname{rank}(\widehat{\mathbf{V}}), (p-k)(q-k))$, be the eigenvalues of $\widehat{\mathbf{Q}} = (\widehat{\mathbf{R}}_0^T \otimes \widehat{\mathbf{U}}_0^T) \widehat{\mathbf{V}} (\widehat{\mathbf{R}}_0 \otimes \widehat{\mathbf{U}}_0)$ in descending order. If $\operatorname{rank}(\mathbf{M}) = k$, $\sum_{i=1}^s \hat{w}_i \chi_i^2$ is a consistent estimate of $\sum_{i=1}^s w_i \chi_i^2$ and the limiting distribution of $\Lambda_1(k)$ is consistently estimated by $\sum_{i=1}^s \hat{w}_i \chi_i^2$. To approximate a linear combination of chi-square random variables, one may use Wood's [34] statistic. In practice, the computationally less expensive Satorra and Bentler's [32] scaled and adjusted chi-square approximations to the weighted chi-square distribution are frequently used.

2.2. Chi-square test

The estimated kernel matrix can be expressed as

$$\widehat{\mathbf{M}} = \mathbf{M} + (\widehat{\mathbf{M}} - \mathbf{M}) = \mathbf{M} + \epsilon \frac{\widehat{\mathbf{M}} - \mathbf{M}}{\epsilon} = \mathbf{M} + \epsilon \mathbf{B},$$

where $\epsilon \mathbf{B}$ is the perturbation of the matrix \mathbf{M} [22]. Using (3) we obtain that the perturbation matrix is asymptotically normal with zero mean and standard deviation of order $n^{-1/2}$; that is, $\epsilon \mathbf{B} = O_p(n^{-1/2})$. Let

$$\Lambda_2(k) = n \operatorname{vec}(\widehat{\mathbf{D}}_0^T) \widehat{\mathbf{Q}}^+ \operatorname{vec}(\widehat{\mathbf{D}}_0^T) \tag{8}$$

where $\widehat{\mathbf{D}}_0$ is defined in (5) and $\widehat{\mathbf{Q}} = (\widehat{\mathbf{R}}_0^T \otimes \widehat{\mathbf{U}}_0^T) \widehat{\mathbf{V}} (\widehat{\mathbf{R}}_0 \otimes \widehat{\mathbf{U}}_0)$. The notation \mathbf{A}^+ signifies the inverse of a matrix \mathbf{A} if it is non-singular, or its Moore–Penrose generalized inverse otherwise. This is a Wald-type test statistic [28] for testing (1). It has the attractive feature of being asymptotically chi-square, in contrast to (7), as shown next.

Theorem 2. Assume $\operatorname{rank}(\mathbf{M}) = k$ and that $\widehat{\mathbf{M}}$ satisfies (3). Then

$$\Lambda_2(k) \xrightarrow{\mathcal{D}} \chi_s^2 \tag{9}$$

for Λ_2 defined in (8), where $s = \min(\operatorname{rank}(\mathbf{V}), (p-k)(q-k))$.

Remark. When the rank k random matrix \mathbf{M} is symmetric, $\mathbf{U}_0 = \mathbf{R}_0$ and $\widehat{\mathbf{U}}_0 = \widehat{\mathbf{R}}_0$, so that $\widehat{\mathbf{U}}_0^T \widehat{\mathbf{M}} \widehat{\mathbf{R}}_0$ is $(p-k) \times (p-k)$ symmetric. Hence, its variance, $\widehat{\mathbf{Q}}$, has at most $s = (p-k)(p-k+1)/2$ non-zero eigenvalues, which is the value of s in both (7) and (9).

3. A chi-square test for assessing component significance

If the rank of \mathbf{M} is k , $\mathcal{S}(\mathbf{M}) = \mathcal{S}(\mathbf{U}_1) = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$, where \mathbf{u}_i , $i = 1, \dots, p$, are the left $p \times 1$ singular vectors of \mathbf{M} . In general, we can *simultaneously* test whether any component or set of components of any basis element of $\mathcal{S}(\mathbf{M})$ equals zero by selecting an appropriate matrix \mathbf{C} and testing

$$H_0 : \mathbf{C} \text{vec}(\mathbf{U}) = 0 \quad \text{vs.} \quad H_1 : \mathbf{C} \text{vec}(\mathbf{U}) \neq 0 \tag{10}$$

where the matrix \mathbf{C} is a pre-specified matrix of zeros and ones of dimension $r \times pk$ and rank r , where r is the number of components whose contribution is being tested, p is the dimension of the basis elements, and $k = \text{rank}(\mathbf{M})$. The unit entry in each row of the matrix \mathbf{C} corresponds to the component(s) of the basis elements $\mathbf{u}_1, \dots, \mathbf{u}_k$ tested for being zero.

The test we propose requires the computation of the asymptotic distribution of the k ($=\text{rank}(\mathbf{M})$) left singular vectors of $\widehat{\mathbf{M}}$. Bura and Pfeiffer [7] showed that for any random matrix \mathbf{M} with an asymptotically normal sample estimate $\widehat{\mathbf{M}}$, as in (3), the asymptotic distribution of $\widehat{\mathbf{U}}_1$ is given by

$$n^{1/2} \text{vec}(\widehat{\mathbf{U}}_1 - \mathbf{U}_1) \xrightarrow{\mathcal{D}} N_{pk}(0, (\mathbf{D}^{-1} \mathbf{R}_1^T \otimes \mathbf{I}) \mathbf{V} (\mathbf{R}_1 \mathbf{D}^{-1} \otimes \mathbf{I})) \tag{11}$$

where $\widehat{\mathbf{U}}_1$ is defined in (4) and \mathbf{D} and \mathbf{R}_1 in (2). A general Wald-type test for (10) is given in the next theorem. Its proof can be found in [7, Theorem 2].

Theorem 3. Let \mathbf{C} be a matrix of order $r \times pk$ and rank r , $\boldsymbol{\theta} = \text{vec}(\mathbf{C} \text{vec}(\mathbf{U}_1))$ and $\widehat{\boldsymbol{\theta}} = \text{vec}(\mathbf{C} \text{vec}(\widehat{\mathbf{U}}_1))$, both $rp \times 1$ vectors. Also, let $\mathbf{A} = \mathbf{C}(\mathbf{D}^{-1} \mathbf{R}_1^T \otimes \mathbf{I}) \mathbf{V} (\mathbf{R}_1 \mathbf{D}^{-1} \otimes \mathbf{I}) \mathbf{C}^T$.

a. If \mathbf{V} is positive definite, then when $\boldsymbol{\theta} = 0$

$$L = n \widehat{\boldsymbol{\theta}}^T \widehat{\mathbf{A}}^{-1} \widehat{\boldsymbol{\theta}} \xrightarrow{\mathcal{D}} \chi^2(r) \tag{12}$$

where $\widehat{\mathbf{A}}$ is a consistent estimate of \mathbf{A} .

b. If \mathbf{V} is positive semi-definite with $\text{rank}(\mathbf{V}) \geq r$, then when $\boldsymbol{\theta} = 0$

$$L = n \widehat{\boldsymbol{\theta}}^T \widehat{\mathbf{A}}^+ \widehat{\boldsymbol{\theta}} \xrightarrow{\mathcal{D}} \chi^2(r) \tag{13}$$

where $\widehat{\mathbf{A}}$ is a consistent estimate of \mathbf{A} .

Bura and Pfeiffer [7] showed that the sample moment based estimate $\widehat{\mathbf{A}} = \mathbf{C}(\widehat{\mathbf{D}}^{-1} \widehat{\mathbf{R}}_1^T \otimes \mathbf{I}) \widehat{\mathbf{V}} (\widehat{\mathbf{R}}_1 \widehat{\mathbf{D}}^{-1} \otimes \mathbf{I}) \mathbf{C}^T$ is consistent for \mathbf{A} , and that $\widehat{\mathbf{A}}^{-1}$ is consistent for \mathbf{A}^{-1} , if \mathbf{V} is positive definite. If \mathbf{V} is positive semi-definite then $\widehat{\mathbf{A}}^+$ is consistent for \mathbf{A}^+ , where $+$ indicates the Moore–Penrose inverse.

4. Unifying sufficient dimension reduction

4.1. Kernel based SDR

Let $\mathbf{X} = (X_1, \dots, X_p)^T$ denote a predictor vector and Y a response variable. Sufficient dimension reduction (SDR) is based on the idea that \mathbf{X} can be replaced with a lower-dimensional projection $\mathbf{P}_\mathcal{S} \mathbf{X}$ without loss of information about the conditional distribution of $Y|\mathbf{X}$, where $\mathbf{P}_\mathcal{S}$ is the orthogonal projection onto the vector space \mathcal{S} in the usual inner product. No pre-specified model for $Y|\mathbf{X}$ is required. The intersection of all subspaces $\mathcal{S} \subset R^p$ with $F(Y|\mathbf{X}) = F(Y|\mathbf{P}_\mathcal{S} \mathbf{X})$, where $F(\cdot|\cdot)$ is the conditional distribution function of the response Y given the second argument, is the *central* subspace, $\mathcal{S}_{Y|\mathbf{X}}$ [10,9]. The dimension $k = \text{dim}(\mathcal{S}_{Y|\mathbf{X}})$ is called the structural dimension of the regression of Y on \mathbf{X} and can take on any value in the set $\{0, 1, \dots, p\}$. When $k < p$, the *structural* dimension of the regression is smaller than the number of predictors. If $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_k)$ is a basis for $\mathcal{S}_{Y|\mathbf{X}}$, $\mathbf{P}_\boldsymbol{\eta} \mathbf{X}$, or equivalently, the k linear combinations $\boldsymbol{\eta}^T \mathbf{X} = (\boldsymbol{\eta}_1^T \mathbf{X}, \dots, \boldsymbol{\eta}_k^T \mathbf{X})$ contain all the information in \mathbf{X} about Y .

If $\boldsymbol{\Sigma}_x$ denotes the covariance matrix of \mathbf{X} , $\mathbf{Z} = \boldsymbol{\Sigma}_x^{-1/2} (\mathbf{X} - E(\mathbf{X}))$ is its standardized version. There is no loss of generality in working in the \mathbf{Z} -scale since $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}_x^{-1/2} \mathcal{S}_{Y|\mathbf{Z}}$. The estimation of the central subspace in most sufficient dimension reduction (SDR) techniques is based on finding a kernel matrix \mathbf{M} so that

$$\mathcal{S}(\mathbf{M}) \subset \mathcal{S}_{Y|\mathbf{Z}}. \tag{14}$$

Suppose the kernel matrix \mathbf{M} in (14) is of order $p \times q$. Let $k = \text{rank}(\mathbf{M})$ and $\lambda_1 \geq \dots \geq \lambda_k$ be the non-zero singular values of \mathbf{M} , and $\mathbf{u}_1, \dots, \mathbf{u}_k$ be its corresponding left singular vectors. The estimation of the possibly lower-dimensional subspace $\mathcal{S}(\mathbf{M})$ in (14) can be formulated as an eigen-decomposition problem where estimating the dimension of $\mathcal{S}(\mathbf{M})$ amounts to estimating the rank of the kernel matrix \mathbf{M} , k , and estimation of the subspace itself amounts to estimating the k left singular vectors of \mathbf{M} , $\mathbf{u}_1, \dots, \mathbf{u}_k$, since $\text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k) = \mathcal{S}(\mathbf{M})$. The SDR predictors $(\mathbf{Z}_1^*, \dots, \mathbf{Z}_r^*) = (\mathbf{u}_1^T \mathbf{Z}, \dots, \mathbf{u}_r^T \mathbf{Z})$ are the projections of \mathbf{Z} onto $\mathcal{S}(\mathbf{M})$. The SDR predictors in the \mathbf{X} scale are $(\mathbf{X}_1^*, \dots, \mathbf{X}_r^*) = (\boldsymbol{\Sigma}_x^{-1/2} \mathbf{u}_1^T \mathbf{X}, \dots, \boldsymbol{\Sigma}_x^{-1/2} \mathbf{u}_r^T \mathbf{X})$. They replace the original \mathbf{X} predictor vector in modeling the response as a function of \mathbf{X} .

For the span of a kernel matrix to be a subspace of $\mathcal{S}_{Y|Z}$ at least one of two conditions on the marginal moments of the predictors must hold. For first moment based kernel methods, such as Sliced Inverse Regression (SIR) [23] and Parametric Inverse Regression (PIR) [5], the following linearity condition is needed:

$$E(\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|Z}}\mathbf{Z}) = \mathbf{P}_{\mathcal{S}_{Y|Z}}\mathbf{Z}. \tag{15}$$

For second moment based kernel methods, such as Sliced Average Variance Estimation (SAVE) [14] and principal Hessian directions (pHd) [23,10], condition (15) and also the constant variance condition

$$\text{Var}(\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|Z}}\mathbf{Z}) = \mathbf{I} - \mathbf{P}_{\mathcal{S}_{Y|Z}} \tag{16}$$

are required.

To estimate $k = \dim(\mathcal{S}_{Y|Z})$, the test statistic for dimension is generally of the form $L_k = n \sum_i f(\hat{\lambda}_i)$, where $\hat{\lambda}_i$ are the singular or eigenvalues of $\hat{\mathbf{M}}$ in decreasing order and f is a smooth non-negative function. The dimension is usually estimated via sequential testing $H_0 : k = r$ against $H_a : k > r$, starting at $r = 0$, which corresponds to independence of Y and \mathbf{Z} . Assessment of the accuracy of the estimation requires knowledge of the asymptotic distribution of the test statistic, the computation of which comprises an important aspect of all SDR techniques.

4.2. Estimation methods

Suppose a random sample of size n is available on (Y, \mathbf{X}) resulting in the $n \times 1$ vector \mathbf{Y} of responses, and the $n \times p$ matrix \mathbf{X}_n of observations on the predictors. The standardized version of the \mathbf{X} matrix is $\mathbf{Z}_n = \hat{\Sigma}_x^{-1/2}(\mathbf{X}_n - \bar{\mathbf{X}}_n)$, where $\bar{\mathbf{X}}_n = \sum_{i=1}^n \mathbf{X}_i/n$ and $\hat{\Sigma}_x = \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)^T/n$. Table 1 lists the kernel matrices of several SDR methods and their respective estimates. The two most popular kernel based SDR methods, SIR and SAVE, will be discussed in detail in Section 4.3. The notation used in the sample estimates of their respective kernel matrices is as follows. Throughout, H indicates the number of slices used, n_h is the number of observations in slice h , \mathbf{Z}_h denotes the $n_h \times p$ matrix of the standardized predictors in slice h , $\bar{\mathbf{Z}}_h = \sum_{i=1}^{n_h} \mathbf{Z}_{ih}/n_h$ denotes their $p \times 1$ intra-slice mean, and $\hat{p}_h = n_h/n$ is the fraction of observations falling in slice h .

For PIR, $\hat{\mathbf{M}} = \hat{\mathbf{B}}_n$ is the least squares estimate of the $q \times p$ parameter matrix $\mathbf{B} = (\beta_{ij})$ in the linear model $\mathbf{Z}_n|Y = \mathbf{F}_n\mathbf{B} + \mathbf{E}_n$, where $\mathbf{F}_n = (\tilde{f}_{il})$ is an $n \times q$ fixed matrix with $\tilde{f}_{il} = \tilde{f}_i(Y_i) = f_{il} - \sum_{i=1}^n f_{il}/n$, the centered version of $f_{il} = f_i(Y_i)$.

Simple Contour Regression (SCR), introduced by Li et al. [26], uses the matrix $\mathbf{H}(c)$, defined by $E[(\tilde{\mathbf{X}} - \mathbf{X})(\tilde{\mathbf{X}} - \mathbf{X})^T I(|\tilde{Y} - Y| \leq c)]$ for $c > 0$ and $(\tilde{\mathbf{X}}, \tilde{Y})$ an independent copy of (\mathbf{X}, Y) . The sample based estimate of $\mathbf{H}(c)$ is

$$\hat{\mathbf{H}}(c) = \frac{1}{\binom{n}{2}} \sum_{(i,j) \in N} (\mathbf{X}_j - \mathbf{X}_i)(\mathbf{X}_j - \mathbf{X}_i)^T I(|Y_j - Y_i| \leq c)$$

where $N = \{(i, j) : i = 2, \dots, n, j = 1, \dots, i-1\}$. Li and Wang [25] developed Directional Regression (DR) that builds upon and substantially improves the accuracy of contour regressions and decreases computing time. Moreover, Li and Wang [25] showed $\mathcal{S}_{DR} = \mathcal{S}_{SAVE}$, yet it is computationally more accurate than SAVE. In Table 1, the notation $E_n = \sum_{i=1}^n (\cdot)/n$ is used for DR.

In Cook and Li's [13] Central Mean Subspace (CMS), $\beta_{YZ} = E(Y\mathbf{Z})$ and $\Sigma_{YZZ} = E[(Y - E(Y))\mathbf{Z}\mathbf{Z}^T]$. Their estimates are the corresponding sample moment estimates. In MAVE, g stands for the unknown link function in $Y = g(\boldsymbol{\eta}^T \mathbf{X}) + \epsilon$, where $\boldsymbol{\eta}$ is a $p \times k$ orthogonal matrix so that $\mathcal{S}(\boldsymbol{\eta})$ is a dimension reduction subspace. The estimated kernel matrix uses the minimizers $\hat{\mathbf{b}}_j$ of

$$\min_{a_j, \mathbf{b}_j} \sum_{i=1}^n (y_i - (a_j + \mathbf{b}_j^T(\mathbf{X}_i - \mathbf{X}_j)))^2 w_{ij}$$

where $w_{ij} = K_h(\mathbf{X}_i - \mathbf{X}_j) / \sum_{i=1}^n K_h(\mathbf{X}_i - \mathbf{X}_j)$. K_h is a multidimensional kernel function and h is the bandwidth [35,8].

Let $\hat{\mathbf{M}}$ denote an estimate of \mathbf{M} based on a random sample of size n . For most SDR methods, the asymptotic normality of $\hat{\mathbf{M}}$ has already been established. This is intuitively true since the kernels are moments or functions of moments of the conditional distribution of the predictors given the response. The reader is referred to the provided references for each method. Moreover, a general paradigm of obtaining the asymptotic normality of functions of means and variances is provided in the proof for the asymptotic normality of the SIR and SAVE kernel matrices in Section 4.3. To the best of our knowledge, the asymptotic distribution of $\hat{\mathbf{M}}_{MAVE}$ has not been computed yet.

The only two conditions required for both Theorems 1 and 2 are (a) $\mathcal{S}(\hat{\mathbf{M}}) \subset \mathcal{S}_{Y|X}$ and (b) the existence of the fourth moments of \mathbf{X} . Condition (a) is satisfied when either (15) and/or (16) hold, depending on the SDR method used to estimate $\hat{\mathbf{M}}$. We will illustrate the application of our general results in Sections 2 and 3 using the two, arguably, most popular SDR methods, SIR and SAVE, in Section 4.3 and in the simulation section.

4.3. SIR and SAVE

So far in SDR literature, researchers have focused on developing tests for dimension tailored to the specific kernel matrix to each method. As the platform for arguing in favor of the proposed unified approach in SDR methodology, we focus on the two well-known and understood SDR methods, SIR [23] and SAVE [14].

For SIR, the test statistic for dimension is $\Lambda_{\text{SIR}} = n \sum_{i=k+1}^p \hat{\lambda}_i$, where $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p$ are the eigenvalues of the moment estimate of $\text{Cov}(E(\mathbf{Z}|Y))$. Bura [4], Bura and Cook [6] showed that Λ_{SIR} is asymptotically distributed as a sum of weighted independent chi-square random variables, each with one degree of freedom. This result was later extended to other SDR methods. Cook and Lee [12] obtained the same result for the SAVE [14] test statistic for dimension in the context of binary regression. Shao et al. [33] proposed a test statistic for dimension in the SAVE context and showed that is also a sum of weighted chi-square random variables. To obtain this result they required that the variance of the Kronecker product of the projection of the predictors onto the null space of the SAVE kernel matrix is constant given the projection of the predictors onto the central dimension reduction subspace (see [33, Th. 3]); a condition which may be rather challenging to check in practice. Others have shown similar results for their proposed SDR methods.

4.3.1. The unified approach

For SIR [23], the range of the response Y is divided into H slices and $\widehat{\mathbf{M}}_{\text{SIR}} = \widehat{\text{Cov}}(E(\mathbf{Z}|Y)) = \sum_{h=1}^H \hat{p}_h \bar{\mathbf{Z}}_h \bar{\mathbf{Z}}_h^T$ is a $p \times p$ symmetric random matrix. If we let $\bar{\mathbf{Z}}_n = (\bar{\mathbf{Z}}_1 \sqrt{\hat{p}_1}, \dots, \bar{\mathbf{Z}}_H \sqrt{\hat{p}_H})$, we can write

$$\widehat{\mathbf{M}}_{\text{SIR}} = \hat{p}_1 \bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^T + \hat{p}_2 \bar{\mathbf{Z}}_2 \bar{\mathbf{Z}}_2^T + \dots + \hat{p}_H \bar{\mathbf{Z}}_H \bar{\mathbf{Z}}_H^T = \bar{\mathbf{Z}}_n \bar{\mathbf{Z}}_n^T. \quad (17)$$

The multivariate central limit theorem (see, for example, [29, p. 15]) and the multivariate version of Slutsky's theorem yield

$$n^{1/2} \text{vec}(\bar{\mathbf{Z}}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} N_{pH}(\mathbf{0}, \mathbf{V}) \quad (18)$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_H)$, $\boldsymbol{\mu}_h = E(\mathbf{Z}_{ih})$, $i = 1, \dots, n_h$. The asymptotic covariance \mathbf{V} is given by (B.3) and (B.4) in Appendix B. We will use $\widehat{\mathbf{M}}_{\text{SIR}} = \bar{\mathbf{Z}}_n$ as our SIR kernel matrix since $E(\mathbf{Z}|Y) \in \text{span}(\text{Cov}(E(\mathbf{Z}|Y)))$ with probability 1 (see [17, Prop. 2.7, p. 75]). We can now apply the theory developed in Sections 2 and 3. The SIR weighted chi-square test statistic for dimension in (6) is

$$\Lambda_1^{\text{SIR}}(k) = n \sum_{i=k+1}^p \hat{\lambda}_i^2 \quad (19)$$

where $\hat{\lambda}_i$ are the ordered singular values of $\widetilde{\mathbf{M}}_{\text{SIR}} = \bar{\mathbf{Z}}_n$ or, equivalently, $\hat{\lambda}_i^2$ are the ordered eigenvalues of $\widehat{\mathbf{M}}_{\text{SIR}}$. The weights in (7) are estimated by the ordered eigenvalues of $\widehat{\mathbf{Q}} = (\widehat{\mathbf{R}}_0^T \otimes \widehat{\mathbf{U}}_0^T) \widehat{\mathbf{V}} (\widehat{\mathbf{R}}_0 \otimes \widehat{\mathbf{U}}_0)$, with $s = (p - k)(H - k)$. The columns of $\widehat{\mathbf{U}}_0$ are the $p - k$ right singular vectors and the columns of $\widehat{\mathbf{R}}_0$ are the $H - k$ left singular vectors of $\widetilde{\mathbf{M}}_{\text{SIR}}$, and $\widehat{\mathbf{V}}$ is the sample moment based estimate of \mathbf{V} .

In Appendix B we also prove, using the multivariate delta method, that

$$\sqrt{n} \text{vec}(\widehat{\mathbf{M}}_{\text{SIR}} - \mathbf{M}_{\text{SIR}}) \rightarrow N_{p^2}(\mathbf{0}, \mathbf{V}_{\text{SIR}}). \quad (20)$$

The $p^2 \times p^2$ asymptotic covariance matrix \mathbf{V}_{SIR} is given by (B.5) also in Appendix B. The asymptotic normality of $\widehat{\mathbf{M}}_{\text{SIR}}$ is known (see [23] for \mathbf{X} normal and [6] for general \mathbf{X}). The interest here lies in the derivation of the asymptotic covariance \mathbf{V}_{SIR} via the delta method and the use of the gradient of a matrix-valued function (see Appendix B). As will also be seen in SAVE, this can lead to a general paradigm for computing the asymptotic covariance of kernel matrix estimates of other SDR methods.

The SIR chi-square test statistic for dimension in (8) is

$$\Lambda_2^{\text{SIR}}(k) = n \text{vec}(\widehat{\mathbf{U}}_0^T \widetilde{\mathbf{M}}_{\text{SIR}} \widehat{\mathbf{R}}_0)^T \widehat{\mathbf{Q}}^+ \text{vec}(\widehat{\mathbf{U}}_0^T \widetilde{\mathbf{M}}_{\text{SIR}} \widehat{\mathbf{R}}_0) \quad (21)$$

with $\widehat{\mathbf{Q}}^+ = (\widehat{\mathbf{R}}_0^T \otimes \widehat{\mathbf{U}}_0^T) \widehat{\mathbf{V}}^+ (\widehat{\mathbf{R}}_0 \otimes \widehat{\mathbf{U}}_0)$. The chi-square degrees of freedom in (9) are $s = (p - k)(H - k)$.

It is worth mentioning that, in the context of SIR, Bai and He [2] also obtained an asymptotic chi-square test for dimension without requiring normality. Yet, the response has to be second-order uncorrelated (see [2] for the definition) with a subset of the predictor vector for the result to hold.

For SAVE [14], for $h = 1, \dots, H$ and $i = 1, \dots, n_h$, we let $\text{Cov}(\mathbf{Z}_{ih}) = \boldsymbol{\Sigma}_h$, $\widehat{\boldsymbol{\Sigma}}_h = \sum_{i=1}^{n_h} (\mathbf{Z}_{ih} - \bar{\mathbf{Z}}_h)(\mathbf{Z}_{ih} - \bar{\mathbf{Z}}_h)^T / n_h$, and $\mathbf{K}_n = ((\mathbf{I}_p - \widehat{\boldsymbol{\Sigma}}_1) \sqrt{\hat{p}_1}, \dots, (\mathbf{I}_p - \widehat{\boldsymbol{\Sigma}}_H) \sqrt{\hat{p}_H})$ denote a $p \times (pH)$ matrix. Then,

$$\widehat{\mathbf{M}}_{\text{SAVE}} = \widehat{\mathbf{K}}_n \widehat{\mathbf{K}}_n^T = \hat{p}_1 (\mathbf{I}_p - \widehat{\boldsymbol{\Sigma}}_1)^2 + \dots + \hat{p}_H (\mathbf{I}_p - \widehat{\boldsymbol{\Sigma}}_H)^2. \quad (22)$$

Applying Corollary 1.2.18 in [29, p. 19], we obtain $n_h^{1/2} (\widehat{\boldsymbol{\Sigma}}_h - \boldsymbol{\Sigma}_h) \xrightarrow{\mathcal{D}} N_{p^2}(\mathbf{0}, \mathbf{Q}_h)$, where $\mathbf{Q}_h = \text{Cov}(\text{vec}(\mathbf{Z}_{1h} - \boldsymbol{\mu}_h)(\mathbf{Z}_{1h} - \boldsymbol{\mu}_h)^T)$. Let $g : \mathbf{A} \in R^{p \times p} \rightarrow (\mathbf{I}_p - \mathbf{A})^2 \in R^{p \times p}$. Using the multivariate delta method yields

$$\sqrt{n_h} \text{vec}(g(\widehat{\boldsymbol{\Sigma}}_h) - g(\boldsymbol{\Sigma}_h)) \xrightarrow{\mathcal{D}} N_{p^2}(\mathbf{0}, \mathbf{Q}_h^*). \quad (23)$$

The $p^2 \times p^2$ asymptotic covariance matrix in (23) is

$$\mathbf{Q}_h^* = \frac{\partial g(\boldsymbol{\Sigma}_h)}{\partial \boldsymbol{\Sigma}_h} \mathbf{Q}_h \left(\frac{\partial g(\boldsymbol{\Sigma}_h)}{\partial \boldsymbol{\Sigma}_h} \right)^T.$$

From (22) and (23) we have

$$\sqrt{n} \text{vec}(\widehat{\mathbf{K}}_n \widehat{\mathbf{K}}_n^T - \mathbf{K}\mathbf{K}^T) = \sqrt{n} \text{vec}(\widehat{\mathbf{M}}_{\text{SAVE}} - \mathbf{M}_{\text{SAVE}}) \implies N(\mathbf{0}, \mathbf{V}_{\text{SAVE}})$$

with $\mathbf{V}_{\text{SAVE}} = \sum_{h=1}^H p_h \mathbf{Q}_h^*$. The computation of the gradient of g at $\boldsymbol{\Sigma}_h$, necessary for calculating \mathbf{Q}_h^* , is given in Appendix B. It is interesting to observe that in SAVE, in contrast to SIR, we can work at the slice level since the summands in (22) are independent of one another. Also we note that our proof of the asymptotic normality of the SAVE kernel matrix only requires the existence of fourth moments of the predictor vector.

The SAVE weighted chi-square test statistic for dimension in (6) is

$$\Lambda_1^{\text{SAVE}}(k) = n \sum_{i=k+1}^{p(p+1)/2} \hat{\lambda}_i^2 \tag{24}$$

where $\hat{\lambda}_i$ are the ordered eigenvalues of $\widehat{\mathbf{M}}_{\text{SAVE}}$. The weights in (7) are estimated by the ordered eigenvalues of $\widehat{\mathbf{Q}} = (\widehat{\mathbf{U}}_0^T \otimes \widehat{\mathbf{U}}_0^T) \widehat{\mathbf{V}}_{\text{SAVE}} (\widehat{\mathbf{U}}_0 \otimes \widehat{\mathbf{U}}_0)$, with $s = (p - k)(p - k + 1)/2$. The columns of $\widehat{\mathbf{U}}_0$ are the $p - k$ right singular vectors of $\widehat{\mathbf{M}}_{\text{SAVE}}$ and $\widehat{\mathbf{V}}_{\text{SAVE}}$ is the sample moment based estimate of \mathbf{V}_{SAVE} .

The SAVE chi-square test statistic for dimension in (8) is

$$\Lambda_2^{\text{SAVE}}(k) = n \text{vec}(\widehat{\mathbf{U}}_0^T \widehat{\mathbf{M}}_{\text{SAVE}} \widehat{\mathbf{U}}_0)^T \widehat{\mathbf{Q}}^+ \text{vec}(\widehat{\mathbf{U}}_0^T \widehat{\mathbf{M}}_{\text{SAVE}} \widehat{\mathbf{U}}_0) \tag{25}$$

with $\widehat{\mathbf{Q}}^+ = (\widehat{\mathbf{U}}_0^T \otimes \widehat{\mathbf{U}}_0^T) \widehat{\mathbf{V}}_{\text{SAVE}}^+ (\widehat{\mathbf{U}}_0 \otimes \widehat{\mathbf{U}}_0)$. The chi-square degrees of freedom in (9) are $s = (p - k)(p - k + 1)/2$.

It should be pointed out that the only other test for dimension for SAVE, aside from a permutation test, is given in [33] and is not directly based on $\widehat{\mathbf{M}}_{\text{SAVE}}$. Their test statistic is asymptotically weighted chi-square unless the predictors are normal, in which case it is chi-square.

The SIR and SAVE test statistics defined in this section will be used in the simulation section to estimate the dimension of the regression.

4.4. A chi-square test for variable contribution or selection

Estimating the rank of a random matrix in multivariate analysis or dimension reduction regression problems provides information about the dimension of the data, but does not shed any light on which variables have significant contributions. By removing variables with insignificant contributions, the complexity of the original data set is further reduced. In dimensionality reduction methods for regression based on a kernel matrix, the rank of the kernel matrix is the dimension of the regression and its left singular vectors are the coefficients of the linear combinations of the predictor vector that are sufficient for modeling the response. That is, the original p variables $\mathbf{X} = (X_1, \dots, X_p)^T$ are replaced by the $k < p$ linear combinations, $X_i^* = \mathbf{u}_i^T \mathbf{X} = u_{i1}X_1 + u_{i2}X_2 + \dots + u_{ip}X_p$, $i = 1, \dots, k$. The coefficients u_{ij} , $j = 1, \dots, p$, of the individual variables, X_1, \dots, X_p , can be thought of as measuring the contribution of each variable to the i th linear combination, X_i^* , $i = 1, \dots, k$. Some coefficients in X_i^* may not be statistically significantly different from zero and the corresponding variables can be removed from the linear combination. Thus, testing for variable contribution to the k linear projections of the original predictors is equivalent to testing for component significance in the k left singular vectors of the kernel matrix as in Section 3.

We can *simultaneously* test the effect of any variable or set of variables in any linear combination or any set of linear combinations by selecting an appropriate matrix \mathbf{C} and testing (10), where the matrix \mathbf{C} is a pre-specified matrix of zeros and ones of dimension $r \times pk$ and rank r , with $r =$ number of variables whose contribution is being tested, p the number of variables in the data set, and $k = \text{rank}(\mathbf{M})$. The unit entry in each row of the matrix \mathbf{C} corresponds to the coefficient of the variable whose importance is being tested. For example, if variable X_2 does not contribute to the first linear combination X_1^* , then u_{21} , the coefficient of X_2 in X_1^* and the second element of \mathbf{u}_2 , should not be significantly different from zero. If we let $\mathbf{C} = (0, 1, 0, \dots, 0)$ be a $1 \times pk$ matrix, then $\mathbf{C} \text{vec}(\mathbf{U}) = u_{21}$, and testing $H_0 : \mathbf{C} \text{vec}(\mathbf{U}) = 0$ vs. $H_1 : \mathbf{C} \text{vec}(\mathbf{U}) \neq 0$ is equivalent to assessing whether X_2 has a significant contribution to X_1^* . The test statistic is the asymptotic chi-square test statistic given in (12) when the asymptotic covariance matrix of the kernel matrix is full rank or in (13) when it is not.

In particular, the special case where the effect of the same variable(s) across all k linear combinations is assessed is equivalent to variable selection. That is, the chi-square test for variable contribution can also be applied to select variables important in modeling the response.

5. Simulation study

The following two models were considered by Shao et al. [33] in their simulation study:

$$Y = X_1 + \epsilon \tag{26}$$

$$Y = X_1^2 + X_2 + \epsilon. \tag{27}$$

For both models, ϵ is an independent and normally distributed random variable with mean zero and standard deviation 0.1. The predictor vector $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$ has dimension $p = 4, 10$ in the simulations. The first model in (26)

Table 2
Estimated levels and power of nominal 5% tests based on model (26).

n			$p = 4, X_i \approx t(5)$				
			SIR	SIR*	SIR adj	SIR scaled	SIR Wald
$H = 5$	100	$d = 0$ vs. $d \geq 1$	1	1	1	1	1
		$d = 1$ vs. $d \geq 2$	0.048	0.028	0.03	0.047	0.045
	200	$d = 0$ vs. $d \geq 1$	1	1	1	1	1
		$d = 1$ vs. $d \geq 2$	0.056	0.034	0.036	0.055	0.018
$H = 10$	100	$d = 0$ vs. $d \geq 1$	1	1	1	1	1
		$d = 1$ vs. $d \geq 2$	0.056	0.029	0.031	0.065	0.381
	200	$d = 0$ vs. $d \geq 1$	1	1	1	1	1
		$d = 1$ vs. $d \geq 2$	0.043	0.03	0.029	0.056	0.13
	400	$d = 0$ vs. $d \geq 1$	1	1	1	1	1
		$d = 1$ vs. $d \geq 2$	0.051	0.042	0.042	0.054	0.065

Table 3
Estimated levels and power of nominal 5% tests based on model (26).

n		$p = 4, X_i \approx t(5), H = 5$					
		S_N	S_G	SAVE	SAVE adj	SAVE scaled	SAVE Wald
100	$d = 0$ vs. $d \geq 1$	1	1	0.176	0.178	0.38	0.995
	$d = 1$ vs. $d \geq 2$	0.499	0.059	0.048	0.045	0.067	0.202
200	$d = 0$ vs. $d \geq 1$	1	1	0.75	0.745	0.861	0.998
	$d = 1$ vs. $d \geq 2$	0.58	0.057	0.015	0.015	0.017	0.045
400	$d = 0$ vs. $d \geq 1$	1	1	0.946	0.949	0.974	1
	$d = 1$ vs. $d \geq 2$	0.637	0.047	0.006	0.006	0.006	0.02

is a one-dimensional model with $S_{Y|X} = \text{span}((1, 0, \dots, 0)^T)$. The second model in (27) is a two-dimensional model with $S_{Y|X} = \text{span}(((1, 0, \dots, 0), (0, 1, 0, \dots, 0))^T)$. We report simulation results only for non-normal predictors, X_i iid $t(5)$, $i = 1, \dots, p$, to compare our respective results. We set the number of slices to 5, used in [33], and 10 to also examine the effect of number of slices on the tests. The nominal level for all tests is 0.05.

For the weighted chi-square test in Theorem 1, we use Wood's [34] numerical approximation to the exact distribution of a linear combination of independent chi-square variates, as implemented in the *dr* package in R, as well as two simple approximations based on the chi-square distribution [3]. The re-scaled version of $T = \sum_{i=1}^s w_i \chi_i^2$ in (7) is $T_{sc} = T/c \sim \chi_s^2$, where $c = \sum_{i=1}^s w_i/s$. The adjusted version of T is $T_{adj} = T/a \sim \chi_b^2$, where $a = \sum_{i=1}^s w_i^2 / \sum_{i=1}^s w_i$ and $b = (\sum_{i=1}^s w_i)^2 / \sum_{i=1}^s w_i^2$. T_{sc} is a mean corrected version of T , whereas T_{adj} matches the first two moments of T with those of $a\chi_b^2$. Shao et al. [33] used T_{sc} in place of their weighted chi-square test statistic in their simulation study.

In Table 2 we report the estimated level and power for several tests of dimension based on SIR applied only to model (26). For model (27), SIR will estimate the central dimension reduction to be 0, since $E(X_1|Y) = 0$. The column headings denote the following tests: SIR is the original chi-square test statistic proposed by Li [23]; SIR* is the weighted chi-square test statistic and SIR Wald is the chi-square test statistic, both derived in this paper, as described in Section 4.1; SIR scaled is the scaled and SIR adj is the adjusted version of the weighted chi-square test statistic, as described above. In this table, the line headed by $d = 0$ vs. $d \geq 1$ corresponds to the power of the corresponding test, whereas the row headed by $d = 1$ vs. $d \geq 2$ reports the estimated level. The power of all tests is always uniformly best, i.e. 1, but one can immediately observe that the level of the tests depends on the combination of sample size and number of slices. In this simple model, when the number of slices is 5, the weighted chi-square and its adjusted chi-square version, derived in this paper, have levels smaller than the nominal. The scaled version of the weighted chi-square test achieves the nominal level. Li's SIR test has level fairly close to nominal. When the number of slices is 10, the sample size needs to be at least 200 for the scaled version to achieve the nominal level. It also appears that the Wald-type chi-square test is the least conservative and requires higher sample sizes to achieve the nominal level. The effect on estimation of the interplay between the number of slices and the sample size will also be observed in the tables reporting results for the SAVE based tests of dimension for both models.

In Tables 3–8 we report the estimated level and power for several tests of dimension based on SAVE applied to models (26) and (27). The column headings denote the following tests: S_N and S_G denote the sliced average variance estimation test assuming normality (a chi-square test) and the general (weighted chi-square test), respectively, proposed by Shao et al. [33]; SAVE denotes the weighted chi-square test statistic and SAVE Wald is the chi-square test statistic, both derived in this paper, as described in Section 4.1; SAVE scaled is the scaled and SAVE adj is the adjusted version of the weighted chi-square test statistic, as described above. In Tables 3 and 4, the line headed by $d = 0$ vs. $d \geq 1$ corresponds to the power of the corresponding test, whereas the row headed by $d = 1$ vs. $d \geq 2$ reports the estimated level. One can readily see that S_N fails to estimate the correct dimension 1, as expected as it assumes that the predictors are normal. Among the other tests, there is no clear winner as both power and level depend on the combination of sample size and number of slices. In general, the tests derived in this paper are more conservative than those proposed by Shao et al. [33]. Yet, a pattern is emerging. The

Table 4
Estimated levels and power of nominal 5% tests based on model (26).

<i>n</i>		$p = 4, X_i \approx t(5), H = 10$					
		S_N	S_G	SAVE	SAVE adj	SAVE scaled	SAVE Wald
100	$d = 0 \text{ vs. } d \geq 1$	1	1	0.034	0.022	0.145	0.996
	$d = 1 \text{ vs. } d \geq 2$	0.58	0.065	0.124	0.121	0.245	0.73
200	$d = 0 \text{ vs. } d \geq 1$	1	1	0.487	0.474	0.672	0.997
	$d = 1 \text{ vs. } d \geq 2$	0.816	0.104	0.052	0.051	0.061	0.176
400	$d = 0 \text{ vs. } d \geq 1$	1	1	0.838	0.846	0.916	0.998
	$d = 1 \text{ vs. } d \geq 2$	0.875	0.053	0.018	0.018	0.018	0.04

Table 5
Estimated levels and power of nominal 5% tests based on model (27).

<i>n</i>		$p = 4, X_i \approx t(5), H = 5$					
		S_N	S_G	SAVE	SAVE adj	SAVE scaled	SAVE Wald
100	$d = 0 \text{ vs. } d \geq 1$	1	1	0	0	0.007	0.63
	$d = 1 \text{ vs. } d \geq 2$	1	1	0.068	0.063	0.131	0.782
	$d = 2 \text{ vs. } d \geq 3$	0.99	0.965	0.066	0.067	0.082	0.294
200	$d = 0 \text{ vs. } d \geq 1$	1	1	0.086	0.085	0.609	0.819
	$d = 1 \text{ vs. } d \geq 2$	1	0.955	0.418	0.411	0.572	0.904
	$d = 2 \text{ vs. } d \geq 3$	0.905	0.52	0.053	0.054	0.057	0.113
400	$d = 0 \text{ vs. } d \geq 1$	1	1	0.892	0.869	0.968	0.961
	$d = 1 \text{ vs. } d \geq 2$	1	0.914	0.784	0.835	0.905	0.968
	$d = 2 \text{ vs. } d \geq 3$	0.518	0.045	0.015	0.016	0.016	0.022

Table 6
Estimated levels and power of nominal 5% tests based on model (27).

<i>n</i>		$p = 4, X_i \approx t(5), H = 10$					
		S_N	S_G	SAVE	SAVE adj	SAVE scaled	SAVE Wald
100	$d = 0 \text{ vs. } d \geq 1$	1	1	0	0	0	0.501
	$d = 1 \text{ vs. } d \geq 2$	1	1	0.03	0.026	0.074	0.711
	$d = 2 \text{ vs. } d \geq 3$	1	1	0.118	0.125	0.161	0.65
200	$d = 0 \text{ vs. } d \geq 1$	1	1	0.008	0.008	0.318	0.716
	$d = 1 \text{ vs. } d \geq 2$	1	0.995	0.25	0.234	0.37	0.849
	$d = 2 \text{ vs. } d \geq 3$	0.99	0.816	0.099	0.102	0.125	0.32
400	$d = 0 \text{ vs. } d \geq 1$	1	1	0.716	0.673	0.94	0.916
	$d = 1 \text{ vs. } d \geq 2$	1	0.797	0.668	0.667	0.798	0.952
	$d = 2 \text{ vs. } d \geq 3$	0.755	0.079	0.043	0.043	0.048	0.094

Table 7
Estimated levels and power of nominal 5% tests based on model (27).

<i>n</i>		$p = 10, X_i \approx t(5), H = 5$					
		S_N	S_G	SAVE	SAVE adj	SAVE scaled	SAVE Wald
200	$d = 0 \text{ vs. } d \geq 1$	1	1	0	0	0	0.214
	$d = 1 \text{ vs. } d \geq 2$	0.914	0.094	0	0	0	0.335
	$d = 2 \text{ vs. } d \geq 3$	0.437	0.015	0	0	0	0.385
600	$d = 0 \text{ vs. } d \geq 1$	1	1	0.912	0.887	0.984	0.74
	$d = 1 \text{ vs. } d \geq 2$	1	0.764	0.257	0.273	0.536	0.81
	$d = 2 \text{ vs. } d \geq 3$	0.823	0.044	0.014	0.013	0.029	0.046
800	$d = 0 \text{ vs. } d \geq 1$	1	1	0.947	0.942	0.992	0.862
	$d = 1 \text{ vs. } d \geq 2$	1	0.906	0.681	0.682	0.856	0.882
	$d = 2 \text{ vs. } d \geq 3$	0.852	0.057	0.013	0.015	0.021	0.031

scaled version of the weighted chi-square test statistic has higher power and its level is closer to the nominal as compared to either the weighted or adjusted chi-square test statistics. Moreover, the simple Wald-type chi-square test we propose in this paper has similar performance, with level slightly closer to the nominal. We also observe that the combination of the sample size and the number of slices has a big impact on the results across tests.

In Tables 5–8, the line headed by $d = 1 \text{ vs. } d \geq 2$ corresponds to the power of the corresponding test, whereas the row headed by $d = 2 \text{ vs. } d \geq 3$ reports the estimated level. The conclusions are similar to those for the one-dimensional model in Tables 3 and 4. Of course, in this case the sample size required across all tests is larger since the model is more complex.

Table 8
Estimated levels and power of nominal 5% tests based on model (27).

n		$p = 10, X_i \approx t(5), H = 10$					
		S_N	S_G	SAVE	SAVE adj	SAVE scaled	SAVE Wald
400	$d = 0$ vs. $d \geq 1$	1	1	0.056	0.05	0.684	0.358
	$d = 1$ vs. $d \geq 2$	1	0.2	0	0	0.004	0.516
	$d = 2$ vs. $d \geq 3$	0.916	0.056	0.004	0.004	0.012	0.39
600	$d = 0$ vs. $d \geq 1$	1	1	0.741	0.732	0.958	0.684
	$d = 1$ vs. $d \geq 2$	1	0.447	0.122	0.129	0.398	0.786
	$d = 2$ vs. $d \geq 3$	0.956	0.084	0.027	0.029	0.071	0.157
800	$d = 0$ vs. $d \geq 1$	1	1	0.894	0.886	0.973	0.815
	$d = 1$ vs. $d \geq 2$	1	0.726	0.431	0.432	0.687	0.867
	$d = 2$ vs. $d \geq 3$	0.973	0.069	0.039	0.039	0.055	0.096

Table 9
Estimated level and power of nominal 5% variable importance tests in SAVE predictors based on model (26).

n	$p = 4, X_i \approx t(5), H = 5$		
	$H_0 : u_{12} = u_{13} = u_{14} = 0$	$H_1 : u_{11} = u_{12} = u_{13} = u_{14} = 0$	$H_1 : u_{11} = u_{12} = 0$
100	0.066	0.999	0.998
120	0.046	0.998	1
150	0.031	1	0.999
200	0.023	1	1
$p = 4, X_i \approx t(5), H = 10$			
100	0.225	0.997	0.993
200	0.074	0.999	0.998
250	0.041	0.999	1
300	0.033	1	1

Table 10
Estimated level and power of nominal 5% variable importance tests in SAVE predictors based on model (27).

n	$p = 4, X_i \approx t(5), H = 5$	
	$H_0 : u_{ij} = 0, i = 1, 2, j = 3, 4$	$H_1 : u_{ij} = 0, i = 1, 2, j = 1, 2, 3, 4$
50	0.142	1
100	0.058	1
200	0.029	1
$p = 4, X_i \approx t(5), H = 10$		
100	0.144	0.996
200	0.084	1
300	0.044	1
400	0.039	1
500	0.023	1
800	0.006	1

In summary, these simulation results indicate that the scaled chi-square approximation to the proposed weighted chi-square test for dimension has similar performance to Shao et al.'s [33] competitor. What is more important, however, is that the simple Wald-type chi-square test we propose in this paper, is a very good, and many times even better, competitor to the weighted chi-square tests for larger sample sizes.

In Table 9 we report the level (under column headed by H_0) and power (under column(s) headed by H_1) for testing variable importance in model (26). We assess the importance of variables X_2, X_3, X_4 in the SAVE predictor versus the specific alternatives that none of the variables are important ($u_{11} = u_{12} = u_{13} = u_{14} = 0$) and that X_1 and X_2 are not important ($u_{11} = u_{12} = 0$). The power is practically 1 across sample sizes. The level is close to nominal when the sample size is about 120 for $H = 5$, and 250 for $H = 10$. Again the importance of the choice of the number of slices is noted.

For the most complex model (27), we are testing for the importance of X_3 and X_4 ; that is, whether they contribute to both SAVE predictors simultaneously ($H_0 : u_{ij} = 0, i = 1, 2, j = 3, 4$) versus the alternative that none is important in either SAVE predictor ($H_1 : u_{ij} = 0, i = 1, 2, j = 1, 2, 3, 4$). This test is equivalent to variable selection. The power is 1 even at a sample size of 50 when $H = 5$, but the level fluctuates across different sample size/number of slices combinations. The nominal of 5% is achieved at roughly 100 observations when $H = 5$ and at 300 when $H = 10$ (Table 10).

6. Discussion

We present two tests for the rank of a random matrix that is asymptotically normal. As an application of this general result, we provide a general theory that encompasses all sufficient dimension reduction methods based on kernel matrices. The two tests for dimension we propose can be applied to all such SDR methods. Moreover, the asymptotic chi-square test we developed only requires the existence of predictor fourth moments and can be used in all SDR kernel-matrix-based methods in lieu of the currently used weighted chi-square test that requires the numerical approximation of the weighted chi-square distribution quantiles when large samples are available. In contrast, the existing chi-square tests for dimension in SDR require normal predictors (e.g., [23,33]). We also note that the proposed rank or, equivalently, dimension estimates based on sequential testing is consistent for the true rank. This can be shown by using arguments similar to those of Robin and Smith ([31], Thms. 5.1 and 5.2).

We also propose an asymptotic chi-square test for assessing which components of the basis elements of a random matrix are statistically significant. In the context of dimension reduction in regression, this is a general test for variable contribution in the lower-dimensional projections of the predictor vector. Variable selection is a special case of this test. In SDR, the first formal statistical test for the significance of a subset of predictors was developed by Cook [11] (*marginal coordinate hypotheses*) for SIR. The asymptotic distribution of the test statistics he proposed was weighted chi-square under the assumption of linearity. Also, the tests developed there are variable selection tests in that they test for concurrent variable importance in *all* SIR projections of the predictor vector. That is, the marginal coordinate hypothesis procedure does not allow testing whether a variable is not significant in the first SDR predictor, yet allow for it to be retained in another SDR predictor. This is also true for the gridded chi-square test [24], a heuristic method, for assessing variable importance in SIR predictors based on residuals and also for the marginal coordinate weighted chi-square test of Shao et al. [33] for SAVE.

Our test not only can be used for variable selection but also allows testing the contribution of any variable or set of variables in the SDR projections of the predictor vector, either separately or simultaneously. In addition, only finite fourth moments are required and the method can be used to assess variable contribution to *any* linear combination of random variables. For example, this test can be used to assess variable contribution to principal components as they are simply linear combinations of the elements of a random vector with coefficients the elements of the left singular vectors of the covariance matrix of the random vector.

Acknowledgment

The authors would like to thank Prof. Liliana Forzani for her help in computing the gradient in the SAVE asymptotic covariance and for her comments on this paper.

Appendix A

Proof of Theorem 1. Observe that

$$\begin{aligned} \sqrt{n}(\mathbf{R}_0^T \otimes \mathbf{U}_0^T) \text{vec}(\widehat{\mathbf{M}} - \mathbf{M}) &= \sqrt{n} \text{vec}[\mathbf{U}_0^T(\widehat{\mathbf{M}} - \mathbf{M})\mathbf{R}_0] \\ &= \sqrt{n} \text{vec}(\mathbf{U}_0^T \widehat{\mathbf{M}} \mathbf{R}_0) \end{aligned}$$

since from (2) we have $\mathbf{U}_0^T \mathbf{M} \mathbf{R}_0 = \mathbf{0}$. Hence, from (3),

$$\sqrt{n} \text{vec}(\mathbf{U}_0^T \widehat{\mathbf{M}} \mathbf{R}_0) \xrightarrow{d} N(\mathbf{0}, (\mathbf{R}_0^T \otimes \mathbf{U}_0^T) \mathbf{V}(\mathbf{R}_0 \otimes \mathbf{U}_0)). \tag{A.1}$$

Also from (3), $\widehat{\mathbf{M}}$ is root n consistent for \mathbf{M} , which yields that $\widehat{\mathbf{U}}_0$ and $\widehat{\mathbf{R}}_0$ are also root n consistent for \mathbf{U}_0 and \mathbf{R}_0 , respectively [7,30]. That is, $\widehat{\mathbf{U}}_0 = \mathbf{U}_0 + O_p(n^{-1/2})$ and $\widehat{\mathbf{R}}_0 = \mathbf{R}_0 + O_p(n^{-1/2})$, and

$$\begin{aligned} \sqrt{n} \text{vec}(\widehat{\mathbf{U}}_0^T \widehat{\mathbf{M}} \widehat{\mathbf{R}}_0) &= \sqrt{n} \text{vec}[(\mathbf{U}_0 + O_p(n^{-1/2}))^T \widehat{\mathbf{M}} (\mathbf{R}_0 + O_p(n^{-1/2}))] \\ &= \text{vec}[\sqrt{n} \mathbf{U}_0^T \widehat{\mathbf{M}} \mathbf{R}_0 + \mathbf{U}_0^T \widehat{\mathbf{M}} O_p(1) + O_p(1) \widehat{\mathbf{M}} \mathbf{R}_0 + \widehat{\mathbf{M}} O_p(n^{-1/2})]. \end{aligned}$$

Observe that $\mathbf{U}_0^T \widehat{\mathbf{M}} \xrightarrow{p} \mathbf{U}_0^T \mathbf{M}$, and $\widehat{\mathbf{M}} \mathbf{R}_0 \xrightarrow{p} \mathbf{M} \mathbf{R}_0$. Also, $\widehat{\mathbf{M}} O_p(n^{-1/2}) = [\mathbf{M} + O_p(n^{-1/2})] O_p(n^{-1/2}) \xrightarrow{p} \mathbf{0}$, $\mathbf{U}_0^T \mathbf{M} = \mathbf{0} \Rightarrow \mathbf{U}_0^T \mathbf{M} O_p(1) = \mathbf{0}$, and $\mathbf{M} \mathbf{U}_0 = \mathbf{0} \Rightarrow O_p(1) \mathbf{M} \mathbf{U}_0 = \mathbf{0}$. Hence, $\mathbf{U}_0^T \widehat{\mathbf{M}} O_p(1) \xrightarrow{p} \mathbf{0}$, $O_p(1) \widehat{\mathbf{M}} \mathbf{U}_0 \xrightarrow{p} \mathbf{0}$, and $\widehat{\mathbf{M}} O_p(n^{-1/2}) \xrightarrow{p} \mathbf{0}$. These results together with (A.1) imply

$$\sqrt{n} \text{vec}(\widehat{\mathbf{D}}_0) = \sqrt{n} \text{vec}(\widehat{\mathbf{U}}_0^T \widehat{\mathbf{M}} \widehat{\mathbf{R}}_0) \xrightarrow{d} N(\mathbf{0}, (\mathbf{R}_0^T \otimes \mathbf{U}_0^T) \mathbf{V}(\mathbf{R}_0 \otimes \mathbf{U}_0)) \tag{A.2}$$

which in turn yields (7). The weights w_i , $i = 1, 2, \dots, s$, are the eigenvalues of $\mathbf{Q} = (\mathbf{R}_0^T \otimes \mathbf{U}_0^T) \mathbf{V}(\mathbf{R}_0 \otimes \mathbf{U}_0)$ in descending order (see, for example, [20]). \square

Proof of Theorem 2. From (A.2), $\mathbf{Q} = (\mathbf{R}_0^T \otimes \mathbf{U}_0^T) \mathbf{V} (\mathbf{R}_0^T \otimes \mathbf{U}_0)$ is the asymptotic covariance matrix of $\sqrt{n} \text{vec}(\widehat{\mathbf{D}}_0) = \sqrt{n} \text{vec}(\widehat{\mathbf{U}}_0^T \widehat{\mathbf{M}} \widehat{\mathbf{R}}_0)$. Using the consistency of $\widehat{\mathbf{U}}_0$ and $\widehat{\mathbf{R}}_0$ [7,30], we obtain that the estimate of \mathbf{Q} , $\widehat{\mathbf{Q}} = (\widehat{\mathbf{R}}_0^T \otimes \widehat{\mathbf{U}}_0^T) \widehat{\mathbf{V}} (\widehat{\mathbf{R}}_0 \otimes \widehat{\mathbf{U}}_0)$, is also consistent. Inversion is a continuous function, so that $\widehat{\mathbf{Q}}^{-1} \xrightarrow{p} \mathbf{Q}^{-1}$, when \mathbf{Q} is full rank. Also, the Moore–Penrose inverse of a matrix is unique and its entries are continuous functions of the entries of the original matrix. Thus, $\widehat{\mathbf{Q}}^+ \xrightarrow{p} \mathbf{Q}^+$, when \mathbf{Q} is not full rank. Hence, by (A.2), $\Delta_2(k) = n \text{vec}(\widehat{\mathbf{D}}_0)^T \widehat{\mathbf{Q}}^+ \text{vec}(\widehat{\mathbf{D}}_0) \xrightarrow{d} \chi_s^2$ [28]. \square

Appendix B

The asymptotic covariance \mathbf{V} in (18): The asymptotic covariance \mathbf{V} is an $H \times H$ array of $p \times p$ matrices $\mathbf{V}_{hs} = n \text{Cov}(\bar{\mathbf{Z}}_h, \bar{\mathbf{Z}}_s)$, $h, s = 1, \dots, H$. Bura and Cook [6] computed these matrices. For $h = s$,

$$\mathbf{V}_{hh} = \mathbf{I}_p p_h + (1 - 2p_h) \boldsymbol{\Sigma}_{z|h} \tag{B.3}$$

where $\boldsymbol{\Sigma}_{z|h} = \boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{\Sigma}_{x|h} \boldsymbol{\Sigma}_x^{-1/2}$, and $\boldsymbol{\Sigma}_{x|h} = \text{Cov}(\mathbf{X}_i | Y_i \text{ falls in slice } h)$. Also,

$$\mathbf{V}_{hs} = \sqrt{p_h p_s} (\mathbf{I}_p - \boldsymbol{\Sigma}_{z|h} - \boldsymbol{\Sigma}_{z|s}). \tag{B.4}$$

Proof of (20) and computation of \mathbf{V}_{SIR} : By the multivariate version of the delta method, we have

$$n^{1/2} (\bar{\mathbf{Z}}_n \bar{\mathbf{Z}}_n^T - \boldsymbol{\mu} \boldsymbol{\mu}^T) = n^{1/2} (\widehat{\mathbf{M}}_{\text{SIR}} - \mathbf{M}_{\text{SIR}}) \xrightarrow{d} N_{p^2}(\mathbf{0}, \nabla f_{\boldsymbol{\mu}} \mathbf{V} \nabla f_{\boldsymbol{\mu}}^T)$$

where $f(\mathbf{x}) = \mathbf{x} \mathbf{x}^T$, with \mathbf{x} a $p \times H$ matrix. Let \mathbf{H}_p be the elimination matrix defined by $\text{vech } \mathbf{A} = \mathbf{H}_p \text{vec } \mathbf{A}$ for any matrix \mathbf{A} and \mathbf{G}_p be the duplication matrix defined by $\text{vec } \mathbf{A} = \mathbf{G}_p \text{vech } \mathbf{A}$ for any symmetric matrix \mathbf{A} [21, p. 352]. Then,

$$\nabla f = \frac{d \text{vec}(f(\mathbf{x}))}{d \text{vec}^T \mathbf{x}} = \mathbf{G}_p \mathbf{H}_p \frac{d \text{vec}(f(\mathbf{x}))}{d \text{vec}^T \mathbf{x}}$$

where the derivative in the right-hand side is the usual derivative without taking into account any symmetry and equals $(\mathbf{x} \otimes \mathbf{I}_p) + (\mathbf{I}_p \otimes \mathbf{x}) \mathbf{K}_{(p,H)}$, where $\mathbf{K}_{(p,H)}$ is the commutation matrix of order $pH \times pH$ that transforms $\text{vec}(\mathbf{A})$ to $\text{vec}(\mathbf{A}^T)$ for any matrix \mathbf{A} [27]. Then,

$$\nabla f_{\boldsymbol{\mu}} = \mathbf{G}_p \mathbf{H}_p ((\boldsymbol{\mu} \otimes \mathbf{I}_p) + (\mathbf{I}_p \otimes \boldsymbol{\mu}) \mathbf{K}_{(p,H)}).$$

Hence,

$$\mathbf{V}_{\text{SIR}} = \mathbf{G}_p \mathbf{H}_p ((\boldsymbol{\mu} \otimes \mathbf{I}_p) + (\mathbf{I}_p \otimes \boldsymbol{\mu}) \mathbf{K}_{(p,H)}) \mathbf{V} (\mathbf{G}_p \mathbf{H}_p ((\boldsymbol{\mu} \otimes \mathbf{I}_p) + (\mathbf{I}_p \otimes \boldsymbol{\mu}) \mathbf{K}_{(p,H)}))^T. \tag{B.5}$$

Computation of $\nabla g(\boldsymbol{\Sigma}_h)$ in \mathbf{Q}_h^* : The function $g = (\mathbf{I} - \mathbf{A})^2 = (\mathbf{I} - \mathbf{A})(\mathbf{I} - \mathbf{A})$ is symmetric and is applied to $\mathbf{A} \in \mathbb{R}^{p \times p}$, a symmetric matrix. The derivative of the symmetric g at the symmetric \mathbf{A} is

$$\nabla g(\mathbf{A}) = \frac{\partial \text{vech } g(\mathbf{A})}{\partial \text{vech}^T(\mathbf{A})} = \frac{\partial \text{vec } g(\mathbf{A})}{\partial \text{vec}^T(\mathbf{A})} \frac{\partial \text{vec}(\mathbf{A})}{\partial \text{vech}^T(\mathbf{A})}. \tag{B.6}$$

Now,

$$\frac{\partial \text{vec } g(\mathbf{A})}{\partial \text{vec}^T(\mathbf{A})} = -((\mathbf{I} - \mathbf{A}^T) \otimes \mathbf{I}) - (\mathbf{I} \otimes (\mathbf{I} - \mathbf{A}))$$

and,

$$\frac{\partial \text{vec}(\mathbf{A})}{\partial \text{vech}^T(\mathbf{A})} = \mathbf{G}_p \frac{\partial \text{vech}(\mathbf{A})}{\partial \text{vech}^T(\mathbf{A})} = \mathbf{G}_p.$$

Plugging these in (B.6) gives $g(\mathbf{A}) = -((\mathbf{I} - \mathbf{A}) \otimes \mathbf{I}) \mathbf{G}_p - (\mathbf{I} \otimes (\mathbf{I} - \mathbf{A})) \mathbf{G}_p$, so that

$$\nabla g(\boldsymbol{\Sigma}_h) = -((\mathbf{I} - \boldsymbol{\Sigma}_h) \otimes \mathbf{I}) \mathbf{G}_p - (\mathbf{I} \otimes (\mathbf{I} - \boldsymbol{\Sigma}_h)) \mathbf{G}_p \tag{B.7}$$

which is a $p^2 \times p(p + 1)/2$ matrix whose rows are specified by the $p^2 g$ components and its columns by the $p(p + 1)/2$ distinct $\boldsymbol{\Sigma}_h$ entries.

References

- [1] T.W. Anderson, The asymptotic distribution of certain characteristic roots and vectors, in: J. Neyman (Ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, pp. 103–130.
- [2] Z.D. Bai, Z. He, A chi-square test for dimensionality with non-Gaussian data, Journal of Multivariate Analysis 88 (2004) 109–117.
- [3] P.M. Bentler, J. Xie, Corrections to test statistics in principal Hessian directions, Statistics and Probability Letters 47 (2000) 381–389.

- [4] E. Bura, Dimension reduction via inverse regression, Ph.D. Thesis, Department of Statistics, University of Minnesota, 1996.
- [5] E. Bura, R.D. Cook, Estimating the structural dimension of regressions via parametric inverse regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (2001) 393–410.
- [6] E. Bura, R.D. Cook, Extending SIR: the weighted chi-square test, *Journal of the American Statistical Association* 96 (2001) 996–1003.
- [7] E. Bura, R. Pfeiffer, On the distribution of the left singular vectors of a random matrix and its applications, *Statistics and Probability Letters* 78 (2008) 2275–2280.
- [8] P. Cizek, W. Hardle, Robust estimation of dimension reduction space, *Computational Statistics and Data Analysis* 51 (2006) 545–555.
- [9] R.D. Cook, *Regression Graphics: Ideas for Studying Regressions through Graphics*, Wiley, New York, 1998.
- [10] R.D. Cook, Principal Hessian directions revisited, *Journal of the American Statistical Association* 93 (1998) 84–94.
- [11] R.D. Cook, Testing predictor contributions in sufficient dimension reduction, *The Annals of Statistics* 32 (2004) 1062–1092.
- [12] R.D. Cook, H. Lee, Dimension reduction in regressions with a binary response, *Journal of the American Statistical Association* 94 (1999) 1187–1200.
- [13] R.D. Cook, B. Li, Dimension reduction for the conditional mean in regression, *The Annals of Statistics* 30 (2002) 455–474.
- [14] R.D. Cook, S. Weisberg, Discussion of Li [23], *Journal of the American Statistical Association* 86 (1991) 328–332.
- [15] J.G. Cragg, S.G. Donald, On the asymptotic properties of LDU-based tests of the rank of a matrix, *Journal of the American Statistical Association* 91 (1996) 1301–1309.
- [16] J.G. Cragg, S.G. Donald, Inferring the rank of a matrix, *Journal of Econometrics* 76 (1997) 223–250.
- [17] M.L. Eaton, *Multivariate Statistics: A Vector Space Approach*, Wiley, New York, 1983.
- [18] M.L. Eaton, D.E. Tyler, Asymptotic distributions of singular values with applications to canonical correlations and correspondence analysis, *Journal of Multivariate Analysis* 50 (1994) 238–264.
- [19] L. Gill, A. Lewbell, Testing the rank and definiteness of estimated matrices with applications to factor, state-space, and ARMA models, *Journal of the American Statistical Association* 87 (1992) 766–776.
- [20] I. Guttman, *Linear Models: An Introduction*, Wiley, New York, 1982.
- [21] D.A. Harville, *Matrix Algebra from a Statistician's Perspective*, Springer-Verlag, New York, 1997.
- [22] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
- [23] K.-C. Li, Sliced inverse regression for dimension reduction (with discussion), *Journal of the American Statistical Association* 86 (1991) 316–342.
- [24] L. Li, R.D. Cook, C.J. Nachtsheim, Model-free variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2005) 285–299.
- [25] B. Li, S. Wang, On directional regression for dimension reduction, *Journal of the American Statistical Association* 102 (2007) 997–1008.
- [26] B. Li, H. Zha, F. Chiaromonte, Contour regression: a general approach to dimension reduction, *The Annals of Statistics* 33 (2005) 1580–1616.
- [27] J.R. Magnus, H. Neudecker, The commutation matrix: some properties and applications, *Annals of Statistics* 7 (1979) 381–394.
- [28] D.S. Moore, Generalized inverse, Wald's method, and the construction of chi-squared tests of fit, *Journal of the American Statistical Association* 72 (1977) 131–137.
- [29] R.J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.
- [30] Z. Ratsimalahelo, Rank test based on matrix perturbation theory, 2002. <http://sjepeg.univ-fcomte.fr/sjepegbis/crese/Zaka351b.pdf>.
- [31] J.M. Robin, R.J. Smith, Tests of rank, *Econometric Theory* 16 (2000) 151–175.
- [32] A. Satorra, P.M. Bentler, Corrections to tests statistics and standard errors in covariance structure analysis, in: A. von Eye, C.C. Clogg (Eds.), *Latent Variables Analysis: Applications for Developmental Research*, Sage, Newbury Park, CA, 1994, pp. 399–419.
- [33] Y. Shao, R.D. Cook, S. Weisberg, Marginal tests with sliced average variance estimation, *Biometrika* 94 (2007) 285–296.
- [34] A.T.A. Wood, An F approximation to the distribution of a linear combination of chi-squared variables, *Communications in Statistics: Simulation* 18 (1989) 1439–1456.
- [35] Y. Xia, H. Tong, W.K. Li, L.-X. Zhu, An adaptive estimation of dimension reduction space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (2002) 363–410.