Biometrika (2008), **95**, 2, *pp.* 469–479
doi: 10.1093/biomet/asn002
Advance Access publication 15 April 2008

Determining the dimension of the central subspace and central mean subspace

BY PENG ZENG

Department of Mathematics and Statistics, Auburn University, 221 Parker Hall, Auburn, Alabama 36849, U.S.A.

zengpen@auburn.edu

SUMMARY

The central subspace and central mean subspace are two important targets of sufficient dimension reduction. We propose a weighted chi-squared test to determine their dimensions based on matrices whose column spaces are exactly equal to the central subspace or the central mean subspace. The asymptotic distribution of the test statistic is obtained. Simulation examples are used to demonstrate the performance of this test.

Some key words: Central mean subspace; Central subspace; Sufficient dimension reduction; Weighted chi-squared test.

1. INTRODUCTION

Sufficient dimension reduction in regression reduces the dimension of predictors by identifying a subspace that contains all information about regression, and therefore the regression can be conducted in the identified subspace instead of using the original predictors (Li, 1991; Cook, 1998b).

Suppose that the response Y is univariate and the predictor $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ is a vector of continuous explanatory variables. One objective of sufficient dimension reduction is to identify a subspace $\mathcal{S} \subset \mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp X \mid P_{\mathcal{S}} X, \tag{1}$$

where $\perp\!\!\!\perp$ denotes 'is independent of' and $P_{\mathcal{S}}$ is the orthogonal projection matrix on to $\mathcal{S}$ in the usual inner product. Under model (1), the conditional distribution of $Y|X$ is equal to the conditional distribution of $Y|P_{\mathcal{S}}X$, which implies that $P_{\mathcal{S}}X$ contains all information in X about Y. The subspace $\mathcal{S}$ is called a dimension reduction subspace. If the intersection of all dimension reduction subspaces also satisfies (1), it is called the central subspace, which is denoted by $\mathcal{S}_{Y|X}$ (Cook, 1996). The central subspace exists and is unique under weak conditions (Cook, 1996), which are assumed throughout this article. For example, consider a heteroscedastic model $Y = \alpha^{\mathrm{T}} X + \varepsilon \beta^{\mathrm{T}} X$, where α and β are two vectors and ε is a random error independent of X. The central subspace $\mathcal{S}_{Y|X}$ is spanned by α and β.

When only the mean response $E(Y|X)$ is of primary interest, the objective of sufficient dimension reduction is to find a subspace $\mathcal{S} \subset \mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp E(Y|X) \mid P_{\mathcal{S}} X. \tag{2}$$

The intersection of all subspaces satisfying (2) is called the central mean subspace, which is denoted by $\mathcal{S}_{E(Y|X)}$ (Cook & Li, 2002). The central mean subspace contains all information in X about $E(Y|X)$, and is always a subspace of the corresponding central subspace. For the heteroscedastic model mentioned above, the central mean subspace $\mathcal{S}_{E(Y|X)}$ is spanned by α.

One important problem in sufficient dimension reduction is to determine the dimension of the central subspace or central mean subspace. All existing tests for this problem are associated with methods for estimating these two subspaces. For example, Li (1991), Schott (1994) and Bura & Cook (2001) discussed a chi-squared test based on sliced inverse regression, which can be used to determine the dimension of the central subspace. Cook & Ni (2005) discussed a weighted chi-squared test in an inverse regression family. Li (1992), Cook (1998a) and Cook & Li (2004) studied tests for the central mean subspace, which are based on principal Hessian directions (Li, 1992) or iterative Hessian transformation (Cook & Li, 2002).

The methods mentioned above usually construct a matrix M, called a candidate matrix, whose columns span a subspace of $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$. Hence the column space of M, denoted by $\mathcal{S}(M)$, is an approximation of $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$. Each test is essentially intended to determine the rank of M. After imposition of a coverage condition that $\mathcal{S}(M) = \mathcal{S}_{Y|X}$ or $\mathcal{S}(M) = \mathcal{S}_{E(Y|X)}$, the test for the rank of M can be used to infer the dimension of $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$. The coverage condition is introduced mainly for technical convenience. There is no systematic study on when the coverage condition holds for these methods and how it can be verified. Consequently, the tests associated with these methods may underestimate the dimension of $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$.

Recently, Zhu & Zeng (2006) proposed a Fourier method by constructing two candidate matrices, M_{FC} and M_{FM}, whose column spaces are exactly equal to the central subspace and the central mean subspace, respectively. The coverage condition automatically holds for the Fourier method. Therefore, it is expected that a test based on M_{FC} or M_{FM} will minimize the risk of underestimating or overestimating the dimension. Following the idea of Ye & Weiss (2003), Zhu & Zeng (2006) discussed a bootstrap method for estimating dimension, which is computationally intensive and may be impractical when the sample size is large. In this article, we derive a weighted chi-squared test to determine dimension.

2. The Fourier method and its candidate matrices

The Fourier method proposed by Zhu & Zeng (2006) constructs two candidate matrices M_{FM} and M_{FC} such that $\mathcal{S}(M_{\text{FM}}) = \mathcal{S}_{E(Y|X)}$ and $\mathcal{S}(M_{\text{FC}}) = \mathcal{S}_{Y|X}$. As a result of the similarity between M_{FM} and M_{FC}, they can be expressed in the common form

$$M = E\left(c_{12}\left[\sigma_\omega^2 I_p + \{G(X_1) - \sigma_\omega^2(X_1 - X_2)\}\{G(X_2) + \sigma_\omega^2(X_1 - X_2)\}^{\mathrm{T}}\right]\right), \tag{3}$$

where $G(x) = \partial \log f_X(x)/\partial x$, $f_X(x)$ is the marginal density function of X, and (X_1, Y_1) and (X_2, Y_2) are independent and follow the same distribution as (X, Y). Note that $M = M_{\text{FM}}$ when $c_{12} = Y_1 Y_2 \exp(-\sigma_\omega^2 \|X_1 - X_2\|^2/2)$, while $M = M_{\text{FC}}$ when $c_{12} = \exp\{-\sigma_t^2(Y_1 - Y_2)^2/2\}\exp(-\sigma_\omega^2\|X_1 - X_2\|^2/2)$, where σ_ω^2 and σ_t^2 are two positive constants. The only difference between M_{FM} and M_{FC} lies in the expression for c_{12}, which represents different ways of determining the weights in expectation. For ease of presentation, we use M and $\mathcal{S}$ to denote a candidate matrix and the targeted subspace. When the central subspace is of interest, they are represented by M_{FC} and $\mathcal{S}_{Y|X}$; when the central mean subspace is of interest, they are represented by M_{FM} and $\mathcal{S}_{E(Y|X)}$.

The density function $f_X(x)$ is needed to evaluate the candidate matrix in (3). In practice, users may be able to choose a reasonable family of distributions for X. For example, they may have some prior knowledge about $f_X(x)$ from historical data, or they may intentionally make X follow some specific distribution when collecting data, as in computer experiments.

In this article, we assume that X follows a normal distribution, as is at least approximately valid in many applications. If the normality assumption is violated, we may apply the Voronoi weighting method proposed by Cook & Nachtsheim (1994) to assign different weights to different points in order to alleviate the violation of normality. Hence the proposed method is still applicable.

When $X \sim N(0, I_p)$, $G(x) = -x$ and

$$M = E\{c_{12}(\sigma_\omega^2 I_p + aX_1X_2^{\mathrm{T}} - bX_1X_1^{\mathrm{T}})\},$$

where $a = \sigma_\omega^4 + (\sigma_\omega^2 + 1)^2$ and $b = 2\sigma_\omega^2(\sigma_\omega^2 + 1)$ are two constants depending on σ_ω^2. Given a sample $\{(x_i, y_i), i = 1, \ldots, n\}_{1 \leqslant i \leqslant n}$, M can be estimated by replacing the expectation by the sample average,

$$\hat{M}_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(\sigma_\omega^2 I_p + ax_ix_j^{\mathrm{T}} - bx_ix_i^{\mathrm{T}}). \tag{4}$$

Zhu & Zeng (2006) showed that $\hat{M}_n$ asymptotically follows a normal distribution, and thus the eigenvalues and eigenvectors of $\hat{M}_n$ converge to those of M at rate $n^{1/2}$. When the dimension of $\mathcal{S}$ is known to be d, $\mathcal{S}$ can be estimated by the eigenspace of $\hat{M}_n$ corresponding to the largest d eigenvalues. However, d is usually unknown in practice, and it needs to be inferred from data.

3. Main results

This section derives a weighted chi-squared test for determining the dimension of $\mathcal{S}$ from a sample $\{(x_i, y_i), i = 1, \ldots, n\}_{1 \leqslant i \leqslant n}$. We also discuss a discrepancy measure that defines a distance between an estimated subspace and the true subspace, and obtain an asymptotic expansion of this discrepancy measure.

The dimension of $\mathcal{S}$ is equal to the rank of M because $\mathcal{S}$ is exactly equal to the column space of M. Thus testing the rank of M is equivalent to testing the dimension of $\mathcal{S}$. Consider the hypotheses

$$H_0 : \mathrm{rank}(M) = d, \quad H_a : \mathrm{rank}(M) > d.$$

Given a sample $\{(x_i, y_i)\}_{1 \leqslant i \leqslant n}$, a consistent estimator of M is $\hat{M}_n$ given in (4). Let $\hat{\lambda}_1 \geqslant \cdots \geqslant \hat{\lambda}_p \geqslant 0$ be the ordered eigenvalues of $\hat{M}_n$. When H_0 holds, M has only d nonzero eigenvalues. The smallest $p - d$ eigenvalues of $\hat{M}_n$ are therefore expected to be very small, because they are consistent estimators of zero. Hence $\hat{\Lambda}_d = n\sum_{i=d+1}^{p}\hat{\lambda}_i$ can be used as a test statistic, and we reject H_0 if $\hat{\Lambda}_d$ is larger than a threshold, which is chosen according to the sampling distribution of $\hat{\Lambda}_d$. Since it is difficult to obtain the exact distribution of $\hat{\Lambda}_d$, we use instead the limiting distribution of $\hat{\Lambda}_d$ as n goes to infinity. The following theorem gives us an expansion for $\hat{\Lambda}_d$.

THEOREM 1. *When* $\mathrm{rank}(M) = d$ *and* $X \sim N(0, I_p)$, $\hat{\Lambda}_d$ *has the expansion*

$$\hat{\Lambda}_d = n\sum_{i=d+1}^{p}\hat{\lambda}_i = n\binom{n}{2}^{-1}\sum_{i<j}U(x_i, y_i, x_j, y_j) + E\{U(X_0, Y_0, X_0, Y_0)\} + o_p(1),$$

where (X_0, Y_0) *follows the same distribution as* (X, Y)*,*

$$U(x_i, y_i, x_j, y_j) = \frac{1}{2}c_{ij}\{2\sigma_\omega^2(p-d) + 2ax_j^{\mathrm{T}}Qx_i - bx_i^{\mathrm{T}}Qx_i - bx_j^{\mathrm{T}}Qx_j\} - (2\sigma_\omega^2 + 1)^2\tau_i\tau_jx_j^{\mathrm{T}}Qx_i$$
$$\times\left(\eta_i - \frac{\sigma_\omega^2}{\sigma_\omega^2 + 1}\vartheta_iPx_i\right)^{\mathrm{T}}M^{+}\left(\eta_j - \frac{\sigma_\omega^2}{\sigma_\omega^2 + 1}\vartheta_jPx_j\right),$$

$\tau_i = (\sigma_\omega^2 + 1)^{-(p-d)/2} \exp\{-\sigma_\omega^2 x_i^T Q x_i / 2(\sigma_\omega^2 + 1)\}$, *$P$ is the projection matrix on to $\mathcal{S}$, $Q = I_p - P$, M^+ is the Moore–Penrose generalized inverse of M,*

$$\vartheta_i = E_{(X_0,Y_0)}\left[\varphi_{i0} \exp\left\{-\sigma_\omega^2 (x_i - X_0)^{\mathrm{T}} P (x_i - X_0)/2\right\}\right],$$
$$\eta_i = E_{(X_0,Y_0)}\left[\varphi_{i0} \exp\left\{-\sigma_\omega^2 (x_i - X_0)^{\mathrm{T}} P (x_i - X_0)/2\right\} P X_0\right],$$

$\varphi_{i0} = y_i Y_0$ *for the central mean subspace and* $\varphi_{i0} = \exp\{-\sigma_t^2 (Y_0 - y_i)^2/2\}$ *for the central subspace.*

Theorem 1 shows that $\hat{\Lambda}_d$ can be approximated by the sum of a U-statistic and a constant when the sample size is large. Since this U-statistic is first-order degenerate, the following theorem claims that $\hat{\Lambda}_n$ asymptotically follows a weighted chi-squared distribution based on the theory of U-statistics (Lee, 1990, Ch. 3).

THEOREM 2. *Assume that $E\{U(X_1, Y_1, X_2, Y_2)^2\} < \infty$. Under the same conditions as in Theorem 1, the limiting distribution of $\hat{\Lambda}_d = n \sum_{i=d+1}^p \hat{\lambda}_i$ is the same as $\sum_{i=1}^\infty v_i \chi_i^2(1)$, where $\chi_i^2(1)$ are independent chi-squared variables with one degree of freedom, and v_i are the eigenvalues of the integral equation*

$$\int U(x_1, y_1, x_2, y_2)\, \phi(x_1, y_1)\, dF(x_1, y_1) = v_i \phi(x_2, y_2), \tag{5}$$

in which $F(x, y)$ is the joint distribution function of (X, Y).

Similar weighted chi-squared tests have been obtained by other authors, such as Cook (1998a), Bura & Cook (2001) and Cook & Ni (2005), but in their cases the limiting distribution is a linear combination of a finite number of chi-squared distributions. In the above theorem, the number of eigenvalues of the integral equation is generally infinite, so that the limiting distribution of $\hat{\Lambda}_d$ is a linear combination of an infinite number of chi-squared distributions, which can be regarded as a price paid for achieving $\mathcal{S}(M) = \mathcal{S}$.

In order to determine the dimension of $\mathcal{S}$, we sequentially apply the test of $H_0 : \mathrm{rank}(M) = d$ for $d = 0, 1, \ldots, p-1$. Starting with $d = 0$, if the hypothesis is not rejected, then we claim that the dimension of $\mathcal{S}$ is 0. Otherwise, we then test H_0 with $d = 1$. The procedure continues until the hypothesis is not rejected, and we claim the dimension of $\mathcal{S}$ to be the corresponding value of d.

Although Theorem 2 gives the limiting distribution of $\hat{\Lambda}_d$, it is difficult to calculate all the v_i's explicitly. We have to find a simple distribution to approximate the limiting distribution of $\hat{\Lambda}_d$ for practical use. One common choice is to use a single scaled chi-squared random variable (Satterthwaite, 1941; Box, 1954). More recently, this approximation was used by Bentler & Xie (2000) in sufficient dimension reduction. Let $g = \sum v_i^2 / \sum v_i$ and $h = (\sum v_i)^2 / \sum v_i^2$. Then $T = \sum v_i \chi_i^2(1)$ can be approximated by $g\chi^2(h)$ such that they have the same first two moments. It remains to estimate g and h, or equivalently estimate $\sum v_i$ and $\sum v_i^2$, for a given sample.

Following the Fredholm theory of integral equations, there exist sequences of eigenvalues and eigenfunctions, v_i and ϕ_i, satisfying (5), and the kernel U admits the expansion

$$U(x_1, y_1, x_2, y_2) = \sum_{i=1}^\infty v_i \phi_i(x_1, y_1) \phi_i(x_2, y_2),$$

where $\int \phi_i(x, y) \phi_j(x, y) dF(x, y) = 1$ if $i = j$ and $= 0$ if $i \neq j$. Since the constant 1 is an eigenfunction corresponding to the eigenvalue zero, we have $E\{\phi_i(X, Y)\} = 0$ for $i = 1, 2, \ldots$.

It can be verified that

$$E\{U(X_1, Y_1, X_1, Y_1)\} = E\left\{\sum_{i=1}^{\infty} \upsilon_i \phi_i^2(X_1, Y_1)\right\} = \sum_{i=1}^{\infty} \upsilon_i,$$

$$E\{U(X_1, Y_1, X_2, Y_2)^2\} = E\left[\left\{\sum_{\upsilon=1}^{\infty} \upsilon_i \phi_i(X_1, Y_1)\phi_i(X_2, Y_2)\right\}^2\right] = \sum_{i=1}^{\infty} \upsilon_i^2.$$

Given a sample $\{(x_i, y_i)\}_{1\leqslant i\leqslant n}$, $\sum \upsilon_i$ and $\sum \upsilon_i^2$ can be estimated simply by replacing the expectations in the above expressions by their corresponding sample averages. Consequently, g and h can be estimated by

$$\hat{g} = \frac{n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} U(x_i, y_i, x_j, y_j)^2}{n^{-1}\sum_{i=1}^{n} U(x_i, y_i, x_i, y_i)}, \quad \hat{h} = \frac{\{n^{-1}\sum_{i=1}^{n} U(x_i, y_i, x_i, y_i)\}^2}{n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} U(x_i, y_i, x_j, y_j)^2}.$$

When the dimension of $\mathcal{S}$ is chosen to be d by the above testing procedure, an estimator $\hat{\mathcal{S}}$ of $\mathcal{S}$ is the space spanned by the eigenvectors of $\hat{M}_n$ corresponding to the largest d eigenvalues. We need to assess the performance of the estimator. Since a subspace uniquely corresponds to a projection matrix, the distance between $\mathcal{S}$ and $\hat{\mathcal{S}}$ can be defined in terms of projection matrices. Let P be the projection matrix on to $\mathcal{S} = \mathcal{S}(M)$, and let $\hat{P}$ be the projection matrix on to $\hat{\mathcal{S}}$. A distance function is defined to be $D(\hat{\mathcal{S}}, \mathcal{S}) = 1 - d^{-1}\mathrm{tr}(\hat{P}P)$. It is easy to see that $D(\hat{\mathcal{S}}, \mathcal{S})$ is always between 0 and 1. $D(\hat{\mathcal{S}}, \mathcal{S}) = 0$ if $\hat{\mathcal{S}} = \mathcal{S}$, and $D(\hat{\mathcal{S}}, \mathcal{S}) = 1$ if $\hat{\mathcal{S}} \perp \mathcal{S}$. The following theorem provides an expansion of the expectation of the distance function.

THEOREM 3. *Under the same conditions as in Theorem 1,*

$$E\{D(\hat{\mathcal{S}}, \mathcal{S})\} = \frac{(p-d)\zeta}{nd} E\left\{\left\| M^{+}\left(\eta_i - \frac{\sigma_\omega^2}{\sigma_\omega^2 + 1}\vartheta_i P X\right)\right\|^2\right\} + o(n^{-1}),$$

where $\zeta = (2\sigma_\omega^2 + 1)^2(3\sigma_\omega^2 + 1)^{-(p-d+2)/2}(\sigma_\omega^2 + 1)^{-(p-d-2)/2}$ *and* $\|\cdot\|$ *is the Euclidean norm of a vector.*

4. IMPLEMENTATION AND EXAMPLE

Given a sample $\{(x_i, y_i)\}_{1\leqslant i\leqslant n}$, the algorithm below should be followed to test the dimension of the central subspace or the central mean subspace.

Step 1. Choose proper values of σ_ω^2 and σ_t^2, if applicable.

Step 2. Standardize data by calculating $z_i = \hat{\Sigma}^{-1/2}(x_i - \bar{x})$ and $\tilde{y}_i = (y_i - \bar{y})/s_y$, where $\bar{x}$ and $\hat{\Sigma}$ are the sample mean and the sample covariance matrix of the x_i's, and $\bar{y}$ and s_y are the sample mean and the sample standard deviation of the y_i's.

Step 3. Calculate $\hat{M}_n$ using $\{(z_i, \tilde{y}_i), i = 1, \ldots, n\}_{1\leqslant i\leqslant n}$.

Step 4. Perform the spectral decomposition of $\hat{M}_n$ to obtain its eigenvalues $\hat{\lambda}_1 \geqslant \cdots \geqslant \hat{\lambda}_p \geqslant 0$.

Step 5. Test the hypothesis regarding the rank of M:

(i) set $d = 0$;

(ii) test H_0 : rank$(M) = d$ versus H_a : rank$(M) > d$, by calculating $\hat{\Lambda}_d$, $\hat{g}$ and $\hat{h}$, rejecting H_0 if $\hat{\Lambda}_d/\hat{g} > \chi_\alpha^2(\hat{h})$, where α is a significance level, and otherwise accepting H_0;

(iii) if H_0 is rejected, set $d = d + 1$ and repeat (ii), otherwise go to step 6.

Step 6. The dimension of $\mathcal{S}$ is d.

The first four steps of the algorithm also lead to an estimate of $\mathcal{S}$, which is the space spanned by $\hat{\Sigma}^{-1/2}\hat{\alpha}_1, \ldots, \hat{\Sigma}^{-1/2}\hat{\alpha}_d$, where $\hat{\alpha}_1, \ldots, \hat{\alpha}_d$ are eigenvectors of $\hat{M}_n$ corresponding to the largest d eigenvalues (Zhu & Zeng, 2006).

Since $\mathcal{S}(M) = \mathcal{S}$ for any positive σ_ω^2 and σ_t^2, Theorem 2 ensures that the above algorithm works for virtually any positive σ_ω^2 and σ_t^2 as long as the sample size is large enough. However, when the sample size is small, the choice of σ_ω^2 and σ_t^2 may still affect the performance of the weighted chi-squared test. In general we should avoid excessively large or small values of σ_ω^2 and σ_t^2 in order to ensure that the test has a large power against the alternative hypothesis. We consider some extreme cases. If $\sigma_\omega^2 = 0$, $M_{\text{FC}} = E[\exp\{-\sigma_t^2(Y_1 - Y_2)^2/2\}\, E(X_1|Y_1)E(X_2|Y_2)^{\mathrm{T}}]$ and $M_{\text{FM}} = E\{YE(X|Y)\}\, E\{YE(X|Y)\}^{\mathrm{T}}$. Both M_{FC} and M_{FM} are zero if $E(X|Y) = 0$, which happens, for example, when $Y = (\beta^{\mathrm{T}}X)^2$ and $X \sim N(0, I_p)$. Hence $\mathcal{S}(M_{\text{FC}})$ or $\mathcal{S}(M_{\text{FM}})$ may be proper subspaces of $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$ when $\sigma_\omega^2 = 0$. When $\sigma_t^2 = 0$, M_{FC} becomes an expectation of a function of X_1 and X_2, which is independent of the response. In these two cases, therefore, $\mathcal{S}(M)$ cannot always be equal to $\mathcal{S}$. When σ_ω^2 and σ_t^2 are too large, c_{12} is close to 0, in which case the performance of the test also deteriorates. Based on intensive simulations, we recommend the use of $\sigma_\omega^2 = 0{\cdot}5$ when testing the dimension of $\mathcal{S}_{E(Y|X)}$, and $\sigma_\omega^2 = 0{\cdot}8$ and $\sigma_t^2 = 4{\cdot}0$ when testing the dimension of $\mathcal{S}_{Y|X}$; see Zhu & Zeng (2006) for more discussion.

We use the following simulation example to demonstrate the performance of the weighted chi-squared test. Consider a quadratic model,

$$Y = \beta_1^{\mathrm{T}}X + (\beta_2^{\mathrm{T}}X)^2 + 0{\cdot}2\,\varepsilon,$$

where $X \in \mathbb{R}^5 \sim N(0, I_5)$, $\varepsilon \sim N(0, 1)$ is independent of X, $\beta_1 = (1, 1, 0, 0, 0)^{\mathrm{T}}$ and $\beta_2 = (0, 0, 0, 1, 1)^{\mathrm{T}}$. The central subspace and central mean subspace are both equal to $\mathcal{S}(\beta_1, \beta_2)$ with two dimensions.

First, consider estimating the dimension of the central mean subspace. We randomly generate 500 samples each of size n, and apply the weighted chi-squared test with significance level $\alpha = 5\%$ to each sample. Figure 1 shows how the probability of correctly determining $\dim(\mathcal{S}_{E(Y|X)}) = 2$ changes as σ_ω^2 increases for a given sample size. The lines marked by 1, 2, 3 and 4 correspond to sample sizes $n = 50$, 100, 200 and 300, respectively. When the sample size is too small, as with Line 1 in Fig. 1, it is difficult for the weighted chi-squared test to determine the dimension of $\mathcal{S}_{E(Y|X)}$ correctly, because the test is based on large-sample theory. As the sample size increases, the probability increases quickly and the performance is very good even when the sample size is as small as $n = 100$ for some σ_ω^2. The performance is best when σ_ω^2 is about 0·4. When σ_ω^2 is too small or too large, we need a much larger sample size than when σ_ω^2 is moderate in order to achieve similar performance. When $n = 300$, the choice of σ_ω^2 has a rather small influence on the performance of the test, and this supports the claim that the algorithm works for any σ_ω^2 when the sample size is large enough.

Next, we discuss the performance of the weighted chi-squared test for determining the dimension of the central subspace. We randomly generate 500 samples each of size $n = 250$. Figure 2 shows the probability of correctly determining $\dim(\mathcal{S}_{Y|X}) = 2$ for different values of σ_ω^2 and σ_t^2. In Fig. 2(a), σ_t^2 is chosen to be 0·5, 1·5, 3·0 and 8·0, corresponding to the lines marked by 1, 2, 3 and 4. The influence of σ_ω^2 demonstrates a pattern similar to that in Fig. 1. However, the performance is not as good as that for the central mean subspace, because it is much easier to claim noise falsely as a part of the central subspace than as part of the central mean subspace. Figure 2(b) displays how the performance of the test changes with the value of σ_t^2, where σ_ω^2 is

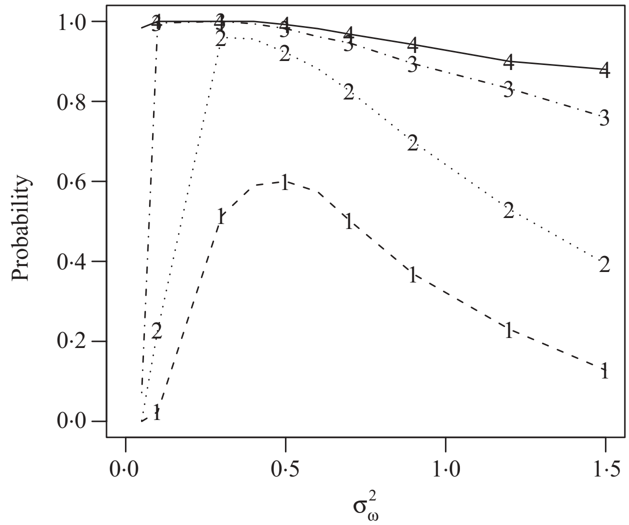

Fig. 1. The probability of correctly determining $\dim(\mathcal{S}_{E(Y|X)}) = 2$ is plotted against σ^2_ω. The four lines marked by 1, 2, 3, 4 correspond to sample size $n = 50$, 100, 200, 300, respectively.

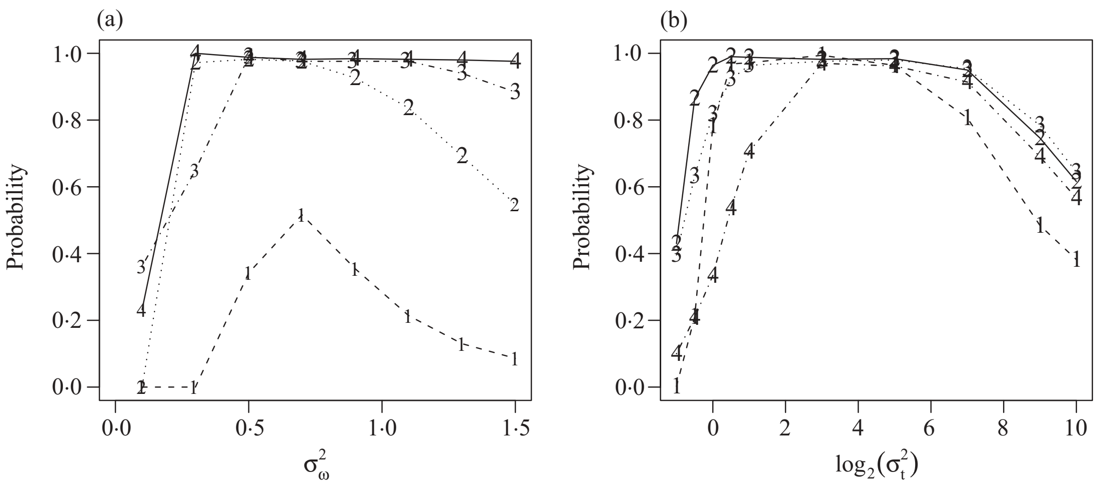

Fig. 2. (a) The probability of correctly determining $\dim(\mathcal{S}_{Y|X}) = 2$ is plotted against σ^2_ω. The four lines marked by 1, 2, 3, 4 correspond to $\sigma^2_t = 0{\cdot}5$, $1{\cdot}5$, $3{\cdot}0$, $8{\cdot}0$, respectively. (b) The probability of correctly determining $\dim(\mathcal{S}_{Y|X}) = 2$ is plotted against $\log_2(\sigma^2_t)$. The four lines marked by 1, 2, 3, 4 correspond to $\sigma^2_\omega = 0{\cdot}3$, $0{\cdot}6$, $0{\cdot}9$, $1{\cdot}5$, respectively.

chosen to be 0·3, 0·6, 0·9 and 1·5, corresponding to the lines marked by 1, 2, 3 and 4. Figure 2(b) shows a pattern similar to that of σ^2_ω, but note that the probability is plotted against $\log_2(\sigma^2_t)$ and the choice of σ^2_t is not so sensitive as the choice of σ^2_ω. A wide range of values of σ^2_t yield similar performance, especially when σ^2_ω is chosen optimally. For example, when $\sigma^2_\omega = 0{\cdot}6$, Line 2, σ^2_t can roughly be any value between 1 and 2^6.

Finally, we compare the weighted chi-squared test with a bootstrap method and a permutation test, because one motivation for deriving the former is that the latter two methods are computationally intensive and may be impracticable when the sample size is large. Recall that the permutation test uses the same test statistic $\hat{\Lambda}_d$, but evaluates p-values according to a permutation algorithm (Cook & Yin, 2001). The bootstrap method, which is not a formal test, calculates

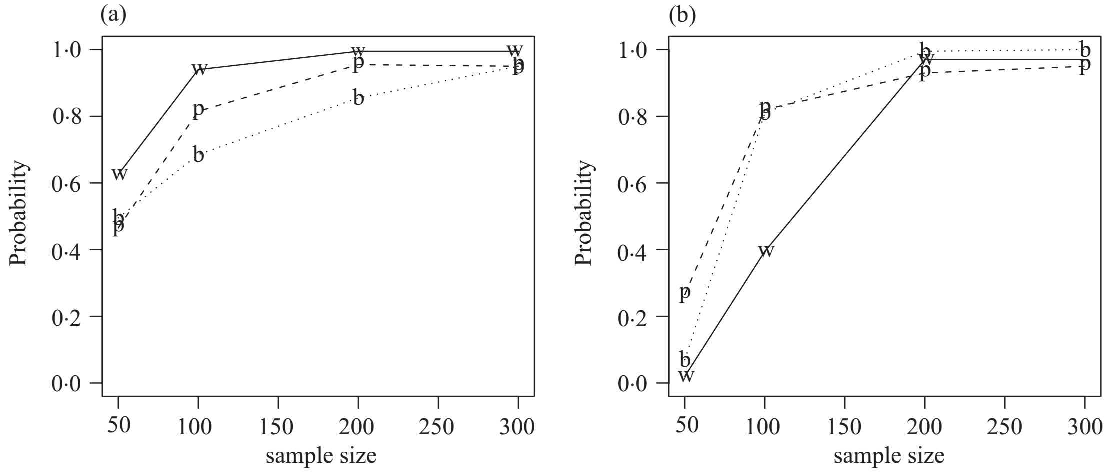

Fig. 3. (a) The percentage of correctly determining $\dim(\mathcal{S}_{E(Y|X)}) = 2$ is plotted against sample size. (b) The percentage of correctly determining $\dim(\mathcal{S}_{Y|X}) = 2$ is plotted against sample size. In both plots, the lines marked 'w', 'p' and 'b' correspond to the weighted chi-squared test, the permutation test and the bootstrap method, respectively.

discrepancies between $\mathcal{S}$ and estimates of $\mathcal{S}$ for different dimensions and chooses the dimension corresponding to the smallest discrepancy. For a fair comparison, all three methods are based on M_{FM} with $\sigma_\omega^2 = 0{\cdot}5$ when estimating the dimension of $\mathcal{S}_{E(Y|X)}$, and are based on M_{FC} with $\sigma_\omega^2 = 0{\cdot}8$ and $\sigma_t^2 = 4{\cdot}0$ when estimating the dimension of $\mathcal{S}_{Y|X}$. We randomly generate 200 samples each of size n, and apply the three methods separately to each sample to determine the dimensions of $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{Y|X}$. We use 1000 bootstrap samples and 1000 permutations for the two computational methods, respectively. Figure 3(a) shows the percentages of correctly determining $\dim(\mathcal{S}_{E(Y|X)}) = 2$ when the sample sizes are $n = 50$, 100, 200 and 300. The weighted chi-squared test performs better than the other two methods for this model. Figure 3(b) shows the probability of correctly determining $\dim(\mathcal{S}_{Y|X}) = 2$. The weighted chi-squared test is comparable to the other two methods when the sample size is large, and is inferior to them when the sample size is small. This occurs because the weighted chi-squared test is derived assuming the sample size is large, while computational methods usually have nice small-sample properties. In view of the large computing time needed by the bootstrap method and the permutation test, the weighted chi-squared test is always preferred when the sample size is large. When the sample size is moderate or small, the weighted chi-squared test is still a useful alternative when the dimension of the central mean subspace is being estimated.

5. DISCUSSION

The limiting distribution of the test statistic $\hat{\Lambda}_d$ has been obtained for normal predictors. Although the normality assumption is stronger than is needed for some existing methods, the proposed weighted chi-squared test does not need to assume the coverage condition. Recently, Li et al. (2005) proposed contour regression, which can exhaustively estimate the central subspace when X follows an elliptically contoured distribution. However, there is no available testing procedure associated with contour regression to determine the dimension of the central subspace.

When X follows a different distribution with a known density function, the limiting distribution of $\hat{\Lambda}_d$ is similar, but with different $U(x_i, y_i, x_j, y_j)$. When the distribution of X is completely unknown, a nonparametric density estimate can be used to replace $f_X(x)$ and a similar weighted chi-squared test is also expected. The candidate matrix (3) discussed in this article is a special case of Zhu & Zeng (2006) with Gaussian weight functions. It is straightforward to extend the results derived here to other weight functions. The choice of σ_ω^2 and σ_t^2 is worth further exploration; for example, Fig. 2 shows some interaction between σ_ω^2 and σ_t^2.

Acknowledgement

The author thanks Professor D. M. Titterington and two anonymous referees for constructive comments and suggestions that greatly helped improve an earlier version of this article. Research for this article was supported by a grant from U.S. National Science Foundation.

Appendix

Sketch proofs

The appendix contains a brief sketch of the proofs of Theorem 1, 2 and 3. More detailed calculations are available from the author upon request.

Proof of Theorem 1. The proof of this theorem follows from Lemma A·1 in Li (1991). Suppose the estimate $\hat{M}_n$ has expansion

$$\hat{M}_n = M + n^{-1/2}T^{(1)} + n^{-1}T^{(2)} + o_p(n^{-1}), \tag{A1}$$

where $T^{(1)}$ and $T^{(2)}$ are symmetric matrices. When the rank of M is d, the sum of the smallest $p-d$ eigenvalues of $\hat{M}_n$ has expansion

$$\sum_{i=d+1}^{p} \hat{\lambda}_i = n^{-1/2}\lambda^{(1)} + n^{-1}\lambda^{(2)} + o_p(n^{-1}),$$

where $\lambda^{(1)} = \mathrm{tr}(QT^{(1)}Q)$, $\lambda^{(2)} = \mathrm{tr}(QT^{(2)}Q - QT^{(1)}M^{+}T^{(1)}Q)$, and Q is the projection matrix on to the eigenspace of M corresponding to eigenvalue zero. Therefore, we only need to obtain an expansion of $\hat{M}_n$ and calculate $\lambda^{(1)}$ and $\lambda^{(2)}$.

By Hoeffding decomposition, the expansion of $\hat{M}_n$ in (A1) holds with

$$T^{(1)} = n^{-1/2}\sum_{i=1}^{n}\{M_1(x_i, y_i) - 2M\},$$

$$T^{(2)} = n\binom{n}{2}^{-1}\sum_{i<j}\tfrac{1}{2}\{M_2(x_i, y_i, x_j, y_j) - M_1(x_i, y_i) - M_1(x_j, y_j) + 2M\}$$
$$+ E\left\{c_{00}\left(\sigma_\omega^2 I_p + X_0X_0^{\mathrm{T}}\right)\right\} - M,$$

where

$$M_1(x_i, y_i) = E_{(X_0,Y_0)}\left\{c_{i0}\left(2\sigma_\omega^2 I_p + ax_iX_0^{\mathrm{T}} + aX_0x_i^{\mathrm{T}} - bx_ix_i^{\mathrm{T}} - bX_0X_0^{\mathrm{T}}\right)\right\},$$

$$M_2(x_i, y_i, x_j, y_j) = c_{ij}\left(2\sigma_\omega^2 I_p + ax_ix_j^{\mathrm{T}} + ax_jx_i^{\mathrm{T}} - bx_ix_i^{\mathrm{T}} - bx_jx_j^{\mathrm{T}}\right).$$

We know that $\lambda^{(1)} = 0$, which follows from $QM_1(x_i, y_i)Q = 0$ for any (x_i, y_i). For $\lambda^{(2)}$, because

$$QT^{(2)}Q = n\binom{n}{2}^{-1}\sum_{i<j}\tfrac{1}{2}QM_2(x_i, y_i, x_j, y_j)Q + E(c_{00})\left(\sigma_\omega^2 + 1\right)Q,$$

$$QM_1(x_i)P = \left(2\sigma_\omega^2 + 1\right)\tau_i Qx_i\left(\eta_i - \frac{\sigma_\omega^2}{\sigma_\omega^2 + 1}\vartheta_i Px_i\right)^{\mathrm{T}},$$

we can obtain $n^{-1}\lambda^{(2)}$ as follows:

$$n^{-1}\text{tr}\big(QT^{(2)}Q - QT^{(1)}M^{+}T^{(1)}Q\big)$$
$$= \binom{n}{2}^{-1} \sum_{i<j} \frac{1}{2} c_{ij} \{2\sigma_\omega^2(p-d) + 2ax_j^{\mathrm{T}}Qx_i - bx_i^{\mathrm{T}}Qx_i - bx_j^{\mathrm{T}}Qx_j\} + n^{-1}E(c_{00})\big(\sigma_\omega^2+1\big)(p-d)$$
$$- n^{-2}\big(2\sigma_\omega^2+1\big)^2 \sum_{i=1}^{n}\sum_{j=1}^{n} \tau_i\tau_j x_j^{\mathrm{T}}Qx_i \left(\eta_i - \frac{\sigma_\omega^2}{\sigma_\omega^2+1}\vartheta_i P x_i\right)^{\mathrm{T}} M^{+} \left(\eta_j - \frac{\sigma_\omega^2}{\sigma_\omega^2+1}\vartheta_j P x_j\right).$$

Finally, $\hat{\Lambda}_d = \lambda^{(2)} + o_p(1)$ and we obtain an expansion of $\hat{\Lambda}_d$ as stated in Theorem 1. □

Proof of Theorem 2. It can be checked that

$$U_n = \binom{n}{2}^{-1} \sum_{i<j} U(x_i, y_i, x_j, y_j)$$

is a U-statistic with kernel $U(x_i, y_i, x_j, y_j)$ such that

$$E\{U(X_1, Y_1, X_2, Y_2)\} = E\{U(X_1, Y_1, X_2, Y_2)|(X_1 = x_1, Y_1 = y_1)\} = 0.$$

According to Theorem 1 in Lee (1990, p. 79), the normalized statistic nU_n converges in distribution to a random variable of the form $\sum_{i=1}^{\infty} v_i(z_i^2 - 1)$, where $z_1, z_2, \ldots$ are independent standard normal random variables, and the v_i's are eigenvalues of the integral equation (5). Note that $\sum_{i=1}^{\infty} v_i = E\{U(X_0, Y_0, X_0, Y_0)\}$, so that the limiting distribution of $\hat{\Lambda}_d$ is the same as $\sum_{i=1}^{\infty} v_i\ \chi_i^2(1)$. □

Proof of Theorem 3 The distance measure can be expressed as.

$$D(\hat{\mathcal{S}}, \mathcal{S}) = 1 - d^{-1}\text{tr}(\hat{P}P) = d^{-1}\text{tr}(\hat{P} - P)P(\hat{P} - P).$$

Applying Lemma 4·1 in Tyler (1981), we know that

$$\hat{Q} = Q - Q(\hat{M}_n - M)M^{+} - M^{+}(\hat{M}_n - M)Q + o_p\big(n^{-1/2}\big).$$

Since $QP = 0$ and $M^{+}P = M^{+}$, then $(\hat{P} - P)P = Q(\hat{M}_n - M)M^{+} + o_p(n^{-1/2})$. The estimator $\hat{M}_n$ has the following expansion by Hoeffding decomposition:

$$\hat{M}_n = M + n^{-1}\sum_{i=1}^{n}\{M_1(x_i, y_i) - 2M\} + o_p\big(n^{-1/2}\big),$$

where $M_1(x_i, y_i)$ is given in the proof of Theorem 1. Since $QM^{+} = 0$ and Qx_i is independent of $M^{+}x_i$, we have

$$QM_1(x_i, y_i)M^{+} = \big(2\sigma_\omega^2 + 1\big)\tau_i Qx_i \left(\eta_i - \frac{\sigma_\omega^2}{\sigma_\omega^2+1}\vartheta_i x_i\right)^{\mathrm{T}} M^{+}.$$

Therefore,

$$D(\hat{\mathcal{S}}, \mathcal{S}) = d^{-1}\text{tr}\{Q(\hat{M}_n - M)M^{+}M^{+}(\hat{M}_n - M)^{\mathrm{T}}Q\} + o_p(n^{-1})$$
$$= d^{-1}\text{tr}\left[\left\{n^{-1}\sum_{i=1}^{n} QM_1(x_i, y_i)M^{+}\right\} \times \left\{n^{-1}\sum_{i=1}^{n} M^{+}M_1(x_i, y_i)^{\mathrm{T}}Q\right\}\right] + o_p(n^{-1}),$$

and $E\{D(\hat{\mathcal{S}}, \mathcal{S})\}$ can be obtained as in the theorem. □

References

Bentler, P. M. & Xie, J. (2000). Corrections to test statistics in principal Hessian directions. *Statist. Prob. Lett.* **47**, 381–9.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in one-way classification. *Ann. Math. Statist.* **25**, 290–302.

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

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Am. Statist. Assoc.* **91**, 983–92.

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

Cook, R. D. (1998b). *Regression Graphics: Ideas For Studying Regressions Through Graphics*. New York: John Wiley & Sons.

Cook, R. D. & Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455–74.

Cook, R. D. & Li, B. (2004). Determining the dimension of iterative Hessian transformation. *Ann. Statist.* **32**, 2501–31.

Cook, R. D. & Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *J. Am. Statist. Assoc.* **89**, 592–600.

Cook, R. D. & Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Am. Statist. Assoc.* **100**, 410–28.

Cook, R. D. & Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with Discussion). *Aust. New Zeal. J. Statist.* **43**, 147–99.

Lee, A. J. (1990). *U-Statistics: Theory and Practice*. New York: Marcel Dekker.

Li, B., Zha, H. & Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.* **33**, 1580–1616.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with Discussion). *J. Am. Statist. Assoc.* **86**, 316–42.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Statist. Assoc.* **87**, 1025–39.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika* **6**, 309–16.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *J. Am. Statist. Assoc.* **89**, 141–8.

Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *Ann. Statist.* **9**, 725–36.

Ye, Z. & Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Am. Statist. Assoc.* **98**, 968–79.

Zhu, Y. & Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Am. Statist. Assoc.* **101**, 1638–51.

[*Received June* 2006. *Revised December* 2007]