Statistica Sinica **15**(2005), 1069-1079

ASYMPTOTIC DISTRIBUTIONS FOR DIMENSION REDUCTION IN THE SIR-II METHOD

Xiangrong Yin and Lynne Seymour

The University of Georgia

Abstract: The asymptotic distribution of the test statistic for testing the dimensionality in the SIR-II method is derived and shown to be a linear combination of χ^2 random variables under weak assumptions. This statistic is based on Li's (1991) sequential test statistic for sliced inverse regression (SIR). Also presented is a simulation study of the result.

Key words and phrases: Asymptotic distributions, Central subspaces, Dimension-reduction subspaces, Regression graphics.

1. Introduction

In this article, the asymptotic distribution of the test statistic for the dimension reduction method SIR-II (Li (1991)) is developed under the null hypothesis for the dimension. The test statistic is based on the sequential test statistic for SIR suggested by Li (1991). In Section 2, a brief review of the SIR-II method is given; in Section 3, the asymptotic distribution for the test statistic is developed. A simulation study of the result is presented in Section 4. All proofs are in the Appendix.

It is assumed throughout this work that the data $(Y_i, \mathbf{X}_i^T)$, $i = 1, \ldots, n$, are i.i.d. observations on $(Y, \mathbf{X}^T)$, and that the first four moments are finite. Here $\mathcal{S}(\mathbf{B})$ denotes the subspace of $\mathcal{R}^t$ spanned by the columns of a $t \times u$ matrix $\mathbf{B}$, $P_{\mathbf{B}}$ denotes the projection operator for $\mathcal{S}(\mathbf{B})$ with respect to the usual inner product, and $Q_{\mathbf{B}} = I - P_{\mathbf{B}}$.

2. The SIR-II Method

Li (1991) proposed an innovative method, SIR, using the inverse mean $\mathrm{E}(\mathbf{X}|Y)$ as a way to estimate the subspace of $\mathcal{R}^p$ spanned by the columns β_j of the $p \times k$ matrix $\boldsymbol{\beta}_\bullet = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ in regression models of the form

$$Y = g(\boldsymbol{\beta}_\bullet^T \mathbf{X}, \varepsilon),$$

where g is an arbitrary unknown function and the error ε is independent of $\mathbf{X}$. However, the inverse mean $\mathrm{E}(\mathbf{X}|Y)$ may fail to recover $\boldsymbol{\beta}_\bullet$. For example, consider

the case $y = x_1^2 + \epsilon$ where ϵ and x_i, $i = 1, \ldots, p$, are independent standard normal variables. Then $\mathrm{E}(\mathbf{X}|Y)$ results in a vector of zeros, when in fact $\boldsymbol{\beta}_\bullet = \mathbf{e}_1$ where $\mathbf{e}_1$ denotes the $p \times 1$ vector with 1 in the first position and 0 elsewhere. Other methods based on inverse variance, $\mathrm{Var}(\mathbf{X}|Y)$, have subsequently been suggested. Cook and Weisberg (1991) proposed a sliced average variance estimate (SAVE), and Li (1991) proposed SIR-II, which is described in more detail later in this section. Both methods can recover the true dimension, $\mathbf{e}_1$, in the previous example. Under certain conditions, these methods estimate part of the *central subspace*, $\mathcal{S}_{Y|\mathbf{X}}$, defined as the smallest subspace of $\mathcal{R}^p$ so that $Y \perp\!\!\!\perp \mathbf{X}|P_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$, where $\perp\!\!\!\perp$ denotes independence (Cook (1994, 1996)).

Let $\boldsymbol{\Sigma}_x = \mathrm{Var}(\mathbf{X}) > 0$, and let $\mathbf{Z} = \boldsymbol{\Sigma}_x{}^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$ be the standardized predictor. Then $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}_x{}^{-\frac{1}{2}}\mathcal{S}_{Y|\mathbf{Z}}$ (Cook (1998b), Proposition 6.1)). If the columns of the matrix $\boldsymbol{\eta}$ form a basis for $\mathcal{S}_{Y|\mathbf{X}}$, then the columns of $\boldsymbol{\gamma} = \boldsymbol{\Sigma}_x{}^{\frac{1}{2}}\boldsymbol{\eta}$ form a basis for $\mathcal{S}_{Y|\mathbf{Z}}$, the central subspace for the regression Y on $\mathbf{Z}$. Thus there is no loss of generality in working on the $\mathbf{Z}$ scale.

Let $\boldsymbol{\Sigma}_Y = \mathrm{Var}(\mathbf{Z}|Y)$. Then SIR-II (Li (1991)) is defined as

$$\mathbf{M}_{\mathrm{SIR-II}} = \mathrm{E}(\boldsymbol{\Sigma}_Y - \mathrm{E}(\boldsymbol{\Sigma}_Y))^2.$$

In fact (Li (1991)):

$$\mathbf{M}_{\mathrm{SAVE}} = \mathbf{M}_{\mathrm{SIR}}^2 + \mathbf{M}_{\mathrm{SIR-II}},$$

where $\mathbf{M}_{\mathrm{SIR}} = \mathrm{Cov}(\mathrm{E}(\mathbf{Z}|Y))$ is the SIR matrix, and $\mathbf{M}_{\mathrm{SAVE}} = \mathrm{E}(\boldsymbol{\Sigma}_Y - I_p)^2$ is the SAVE matrix. Thus SAVE combines SIR and SIR-II in a special way. It is well-known that SIR catches the inverse means and SIR-II catches the inverse variances, while SAVE catches both. There have been extensive asymptotic results for testing the dimension for SIR (e.g., Li (1991), Schott (1994), Velilla (1998), Ferré (1998) and Bura and Cook (2001)). There is no asymptotic result for testing the dimension for SAVE in general, though empirical methods such as a graphical check or an eigenvalue comparison have been used for SAVE (e.g., Cook and Critchley (2000)). So far there is no asymptotic result for testing the dimension for SIR-II either. The above matrix relationship among SIR, SIR-II and SAVE does not reduce the importance of finding such a testing method for SIR-II since the dimensions of these matrices are complicated and, as Cook and Yin (2001) pointed out, in some cases the SIR-II matrix itself may be important.

Let $\boldsymbol{\gamma}$ be a basis of $\mathcal{S}_{Y|\mathbf{Z}}$. Then under

A. $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}) = P_{\boldsymbol{\gamma}}\mathbf{Z}$ and

B. $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}) = Q_{\boldsymbol{\gamma}}$,

we have $\mathcal{S}(\mathbf{M}_{\mathrm{SIR-II}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$ (Li (1991) and Cook (1998b)). If Y is discrete, then this result can be used to justify using conditional sample variances to form $\mathbf{M}_{\mathrm{SIR-II}}$, and to estimate vectors in $\mathcal{S}_{Y|\mathbf{Z}}$. When Y is continuous, Li (1991)

suggested replacing Y with a discrete $\tilde{Y}$ based on partitioning the observed range of Y into $H+1$ fixed, non-overlapping slices. The fact that $Y \perp\!\!\!\perp \mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}$ implies that $\tilde{Y} \perp\!\!\!\perp \mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}$. Thus $\mathcal{S}_{\tilde{Y}|\mathbf{Z}} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. In addition, provided that H is sufficiently large, $\mathcal{S}_{\tilde{Y}|\mathbf{Z}} = \mathcal{S}_{Y|\mathbf{Z}}$, and there is no loss of information when Y is replaced by $\tilde{Y}$. A sample version $\hat{\mathbf{M}}_{\text{SIR}-\text{II}}$ can be constructed, and then $\mathcal{S}(\hat{\mathbf{M}}_{\text{SIR}-\text{II}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. Let $b = \dim(\mathcal{S}(\mathbf{M}_{\text{SIR}-\text{II}}))$. The span of the b eigenvectors of the sample version $\hat{\mathbf{M}}_{\text{SIR}-\text{II}}$ that correspond to its b largest eigenvalues is an estimated part of $\mathcal{S}_{Y|\mathbf{Z}}$. Recently Saracco (2001) and Gannoun and Saracco (2003) obtained the consistency and asymptotic normality of the estimators of SIR-II assuming b is known. Since in practice b is unknown, it is important to have an inference method for b, and that is the goal of this article. The only other work addressing this point is Kötter (1996), who used the cumulative proportion of the estimated eigenvalues to estimate b, empirically.

Given a random sample $\{(Y_i, \mathbf{X}_i^T)\}_{i=1}^n$ from $(Y, \mathbf{X}^T)$, construct first a discrete version $\tilde{Y}$ of Y, where $\tilde{Y} = \{0, ..., H\}$. Let $\boldsymbol{\Sigma}_s = \text{Var}\,(\mathbf{Z}|\tilde{Y} = s)$ for $s = 0, \ldots, H$. Then, form $\hat{\boldsymbol{\Sigma}}_s$, the sample version of $\boldsymbol{\Sigma}_s$. Let $\mathbf{M} = \text{E}\,(\boldsymbol{\Sigma}_{\tilde{Y}} - \boldsymbol{\Sigma}_0)^2$, where $\boldsymbol{\Sigma}_0$ is the conditional variance of $\mathbf{Z}|Y$ for Y's in the first slice of the population. Then form

$$\hat{\mathbf{M}} = \sum_{s=1}^{H} f_s(\hat{\boldsymbol{\Sigma}}_s - \hat{\boldsymbol{\Sigma}}_0)^2,$$

where $f_s = n_s/n$ is the fraction of observations falling in slice s. Note that this $\mathbf{M}$ is different from $\tilde{\mathbf{M}}_{\text{SIR}-\text{II}} = \text{E}\,(\boldsymbol{\Sigma}_{\tilde{Y}} - \text{E}\,(\boldsymbol{\Sigma}_{\tilde{Y}}))^2$, the corresponding version of $\mathbf{M}_{\text{SIR}-\text{II}}$; however, they are equivalent since $\mathcal{S}(\tilde{\mathbf{M}}_{\text{SIR}-\text{II}}) = \mathcal{S}(\mathbf{M})$. $\mathbf{M}$ is used instead of $\tilde{\mathbf{M}}_{\text{SIR}-\text{II}}$ because $\text{E}\,(\boldsymbol{\Sigma}_{\tilde{Y}} - \text{E}\,(\boldsymbol{\Sigma}_{\tilde{Y}})) = 0$, making one of the terms in $\tilde{\mathbf{M}}_{\text{SIR}-\text{II}}$ redundant.

Note that when Y is binary, so that $H = 1$, $\mathbf{M}$ reduces to the Difference Of Covariances (DOC) method (Cook and Lee (1999)). In this article, we assume that $H \geq 1$ is fixed. If $H = 0$, then both population and sample kernel matrices reduce to the null matrix. Let $d = \dim(\mathcal{S}(\mathbf{M}))$. When Y is discrete, testing H_0: $d = 0$ is a by-product of the results in Section 3. This can be considered as an alternative method to Zhu, Ng and Jing (2002) and Zhang and Boos (1992) for testing homogeneity of the covariance matrices for multi-groups without multi-normality, and without reference to the likelihood ratio or union intersection principle.

3. Asymptotic Distribution for The Test Statistic

Suppose that $\hat{\mathbf{M}}$ has eigenvalues, $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$. The test statistic is the sequential test statistic suggested by Li (1991) for SIR:

$$\hat{\Lambda}_d = n \sum_{j=d+1}^{p} \hat{\lambda}_j. \tag{1}$$

Once the distribution of this test statistic under a null hypothesis is known, the number d of nonzero eigenvalues of $\mathbf{M}$ can be estimated as follows: the null hypothesis that $d = m$ is rejected in favor of the alternative that $d \geq m+1$ when $\hat{\Lambda}_m$ is bigger than the percentage point of its distribution. If $\hat{\Lambda}_{m+1}$ is less than the same percentage point of its distribution, one may then infer $d = m + 1$. Cook and Lee (1999) developed the asymptotic distribution of (1) when Y is binary. A completely different test statistic, proposed by Schott (1994) for an alternative form of SIR, uses $\mathbf{\Sigma}_Y$ in a sequence test with a test statistic that requires elliptically symmetric regressors, and for which the tuning constant is the number of observations per slice (rather than the number of slices).

Our result is developed under very weak conditions of finite first four moments, without any restrictions on the predictor distribution and without requiring conditions A and B given in Section 2. However, to ensure that the dimensions found by this method belong to the central subspace, the two conditions are required. The sequential test statistic of Li (1991) has been used as a typical test statistic for moment dimension reduction methods when H is fixed (e.g., Li (1991, 1992), Cook (1998a), Bura and Cook (2001), Cook and Yin (2002) and Yin and Cook (2004)). The general logic behind our development is similar to that used by Cook (1998a) to derive the asymptotic distribution of statistics used in the method of principal Hessian directions (PHD, Li (1992)). This kind of idea has also been used by Bura and Cook (2001), Cook and Yin (2002) and Yin and Cook (2004). The procedure has become a preferred one for moment dimension reduction methods assuming a fixed number of slices, but the same result can be obtained by using perturbation theory such as employed by Li (1991) for SIR.

3.1. General case

Write $\hat{\mathbf{M}} = \mathbf{K}_n\mathbf{K}_n^T$, where $\mathbf{K}_n = ((\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}}_0)\sqrt{f_1}, \ldots, (\hat{\mathbf{\Sigma}}_H - \hat{\mathbf{\Sigma}}_0)\sqrt{f_H})$ is a $p\times(pH)$ matrix, and $\mathbf{M} = \mathbf{K}\mathbf{K}^T$, where $\mathbf{K} = ((\mathbf{\Sigma}_1-\mathbf{\Sigma}_0)\sqrt{p_1}, \ldots, (\mathbf{\Sigma}_H-\mathbf{\Sigma}_0)\sqrt{p_H})$ is also a $p \times (pH)$ matrix, with p_s denoting the probability of Y in slice s.

Suppose that the singular value decomposition of $\mathbf{K}$ is given by

$$\mathbf{K} = \mathbf{\Gamma}^T \begin{pmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{pmatrix} \mathbf{\Psi},$$

where $\mathbf{\Gamma}^T = (\mathbf{\Gamma}_1, \mathbf{\Gamma}_0)$ is an orthonormal $p \times p$ matrix, $\mathbf{\Gamma}_0$ is a $p \times (p - d)$ matrix, $\mathbf{\Psi}^T = (\mathbf{\Psi}_1, \mathbf{\Psi}_0)$ is an orthonormal $(pH) \times (pH)$ matrix, $\mathbf{\Psi}_0$ is a $(pH) \times (pH - d)$ matrix, and $\mathbf{D}$ is a $d \times d$ diagonal matrix with the positive singular values of $\mathbf{K}$ along its diagonal. Using an argument similar to that of Gannoun and Saracco (2003) (see also Saracco (2001)), the $p^2H \times 1$ vector $\mathrm{vec}(\mathbf{V}_n) = \mathrm{vec}(\sqrt{n}(\mathbf{K}_n - \mathbf{K}))$ converges to a multivariate normal distribution with mean 0 and finite covariance

matrix. It then follows from Eaton and Tyler (1994) that the limiting distribution of the smallest $\min(p-d, pH-d)$ singular values of $\sqrt{n}(\mathbf{K}_n - \mathbf{K})$ is the same as that of the corresponding singular values of the $(p-d) \times (pH-d)$ matrix

$$\sqrt{n}\mathbf{U}_n = \sqrt{n}\boldsymbol{\Gamma}_0^T(\mathbf{K}_n - \mathbf{K})\boldsymbol{\Psi}_0 = \sqrt{n}\boldsymbol{\Gamma}_0^T\mathbf{K}_n\boldsymbol{\Psi}_0.$$

The asymptotic distribution of $\hat{\Lambda}_d$ is the same as that of the sum of the squares of the singular values of $\sqrt{n}\mathbf{U}_n$, which is the same as that of $n \times \text{trace}(\mathbf{U}_n\mathbf{U}_n^T)$. Therefore the asymptotic distribution of $\hat{\Lambda}_d$ is the same as the asymptotic distribution of $n\text{vec}(\mathbf{U}_n)^T\text{vec}(\mathbf{U}_n)$. The distribution of $\sqrt{n}\mathbf{U}_n$ is given in the following Lemma.

Lemma 1. $\text{vec}(\sqrt{n}\mathbf{U}_n) \to N(0, \Delta^u)$, *where* $\Delta^u = (\boldsymbol{\Psi}_0^T \otimes \boldsymbol{\Gamma}_0^T)\Delta(\boldsymbol{\Psi}_0 \otimes \boldsymbol{\Gamma}_0), \Delta = (\mathbf{g}\mathbf{g}^T) \otimes \Delta_0 + diag(p_s\Delta_s)$, $\mathbf{g} = (\sqrt{p_1}, \ldots, \sqrt{p_H})^T, \Delta_s = (1/p_s)\text{Var}\left((\mathbf{Z} - \boldsymbol{\mu}_s) \otimes (\mathbf{Z} - \boldsymbol{\mu}_s)|s\right)$, $\boldsymbol{\mu}_s = \text{E}(\mathbf{Z}|s)$, *and* $\otimes$ *is the Kronecker product.*

Since the dimension of Δ^u is $(pH-d)(p-d) \times (pH-d)(p-d)$, based on Lemma 1 and the work of Eaton (1983, p.112), the asymptotic distribution of $\hat{\Lambda}_d$ is given in the following theorem.

Theorem 1. *Assume that* $p > d$ *and* $H \geq 1$. *The distribution of* $\hat{\Lambda}_d$ *converges to that of*

$$\Omega = \sum_{k=1}^{(pH-d)(p-d)} \omega_k\Omega_k,$$

where the Ω_k*'s are independent chi-squared random variables each with* 1 *degree of freedom, and* $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_{(pH-d)(p-d)}$ *are the ordered eigenvalues of* Δ^u.

Theorem 1 allows for a general test of dimension, provided that one can obtain a consistent estimate of Δ^u from which to construct estimates of the eigenvalues ω_k. Based on previous formulas, Δ^u can be estimated consistently by substituting the corresponding sample versions. Additionally, under a hypothesized value of d, $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Psi}_0$ can be estimated consistently by using the sample versions computed from $\mathbf{K}_n$. Let $\hat{\Delta}^u$ denote the resulting estimated version of Δ^u and $\hat{\omega}_k$ denote the eigenvalues of $\hat{\Delta}^u$. The limiting distribution of $\hat{\Lambda}_d$ is then consistently estimated by the distribution of

$$\hat{\Omega} = \sum_{k=1}^{(pH-d)(p-d)} \hat{\omega}_k\Omega_k. \tag{2}$$

The estimation of the weight $\hat{\omega}_i$ in the asymptotic distribution of $\hat{\Omega}$ could cause problems if the sample size is relatively small, making the estimates quite variable. This possibility might be avoided by approximating the distribution of $\hat{\Omega}$

with a scaled chi-square statistic. Bentler and Xie (2000) reported success with this approach in the context of the PHD method (Li (1992); Cook (1998a)). Their general conclusions are expected to apply in the present context as well. If $H = 1$, then $\mathbf{\Psi}_0 = \mathbf{\Gamma}_0$ and $\Delta = p_1\Delta_0 + p_1\Delta_1$, yielding the following result.

Corollary 1. *If $H = 1$, then Theorem 1 reduces to Theorem 2 of Cook and Lee (1999).*

3.2. Normality and conditional normality

When $\mathbf{Z}$ or $\mathbf{Z}|Y$ is normally distributed, the result in Theorem 1 simplifies a little since only the first two moments need to be calculated.

Theorem 2. *Under*
(a) *$p > d$, $H \geq 1$, $\mathbf{Z}|(Y = s) \sim N(\boldsymbol{\mu}_s, \mathbf{\Sigma}_s)$ and $\boldsymbol{\mu}_s \in \mathcal{S}(\mathbf{M})$; or*
(b) *$p > d$, $H \geq 1$, $\mathbf{Z}$ is normally distributed and $\mathcal{S}_{Y|\mathbf{Z}} = \mathcal{S}(\mathbf{M})$,*
the distribution of $\hat{\Lambda}_d$ converges to that of

$$\Omega = \sum_{k=1}^{(pH-d)(p-d)} \omega_k \Omega_k,$$

where the Ω_k's are independent chi-squared random variables each with 1 degree of freedom, and $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_{(pH-d)(p-d)}$ are the ordered eigenvalues of $\Delta^u = (\mathbf{\Psi}_0^T \otimes \mathbf{\Gamma}_0^T)\Delta(\mathbf{\Psi}_0 \otimes \mathbf{\Gamma}_0)$, where

$$\Delta = (I_H + \frac{1}{p_0}\mathbf{g}\mathbf{g}^T) \otimes (I_{p^2} + K_{pp} + (\mathbf{\Sigma}_0 - I) \otimes I) + diag((\mathbf{\Sigma}_s - \mathbf{\Sigma}_0) \otimes I_p),$$

and K_{pp} is a commutation matrix (Mangus (1988)).

Note that Δ^u generally cannot be an idempotent matrix. This is in contrast to the SIR case where a simple chi-squared distribution can be obtained. But it is consistent with similar results obtained for binary Y by Cook and Lee (1999). Again, if $H = 1$, then $\mathbf{\Psi}_0 = \mathbf{\Gamma}_0$, $\mathbf{\Gamma}_0(\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1) = 0$ and $\Delta = (1/p_0)(I_{P^2} + K_{PP} + (p_1 - p_0)(\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1))$. Simple algebra shows that $\Delta^u = (1/p_0)(I_{(p-d)^2} + K_{(p-d)(p-d)})$. Then the following corollary holds.

Corollary 1. *If $H = 1$, then part (a) of Theorem 2 reduces to part (b) of Corollary 1 in Cook and Lee (1999).*

4. A Simulation Study

In this section, results of a simulation study are reported for the weighted χ^2 test for general and normal cases in Sections 3.1 and 3.2. Three examples are considered. In order to make inference about d, one needs to calculate tail probabilities of distributions of linear combinations of χ^2 random variables as in (2).

There is a substantial literature to assist in this computation (see Field (1993) for an introduction). Field's algorithm has been used by Cook (1998a) and Bura and Cook (2001) for general cases of PHD and SIR, respectively. Bentler and Xie (2000) proposed two modifications of Field's algorithm. A different approach is used in these simulations; an approach that takes advantage of modern computing power and uses standard practices in statistical computing and numerical analysis as found in texts such as Lange (1999).

In each of the examples presented below, sample sizes of $n = 100$, $n = 500$ and $n = 1,000$ are used. The simulation study follows this procedure.

Step 1. Simulate data with sample size n according to the specified model.

Step 2. Compute the test statistic $\hat{\Lambda}_d$ using the simulated data for $d = 0, \ldots, (p-1)$.

Step 3. Again for each $d = 0, \ldots, (p-1)$, construct the sampling distribution of $\hat{\Lambda}_d$, using a parametric bootstrap approach (Lange (1999)), by drawing 10,000 realizations of $\hat{\Omega}$:

a. compute the eigenvalues $\hat{\omega}_k$ using the simulated data;
b. draw the required number of $\Omega_k = \chi^2(1)$ variables;
c. compute $\hat{\Omega}$ as in Equation (2).

Step 4. The significance level is chosen to be $\alpha = 0.05$. Thus, compute the 95th percentile of these 10,000 draws of $\hat{\Omega}$.

Step 5. Compare the $\hat{\Lambda}_d$ in Step 2 to the critical value found in Step 4.

These steps are repeated 1,000 different times. All our computations were done using Matlab (Codes can be requested from the authors).

Example 4.1. This model is for discrete $Y = 0, 1, 2$ and $\mathbf{X} = (x_1, x_2, x_3, x_4)^T$ with $\mathbf{X}|(Y = i) \sim N(0, \mathbf{\Sigma}_i)$, for $i = 0, 1, 2$, where

$$\mathbf{\Sigma}_0 = \begin{pmatrix} 2 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathbf{\Sigma}_1 = \begin{pmatrix} 3 & 0.2 & 0 & 0 \\ 0.2 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{\Sigma}_2 = \begin{pmatrix} 1 & -0.3 & 0 & 0 \\ -0.3 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Thus $d = 2$. Since Y is discrete, the number of slices is $H = 3$. Table 1 lists the rejection rates at $\alpha = 0.05$ when replicated with 1,000 repeated simulation runs. When the sample size is small, both sample powers and sample significance levels are away from the theoretical ones in the general test of Theorem 1 and the normal test of Theorem 2. However, with large sample sizes, powers are very high while sample α levels are anti-conservative compared to the theoretical one in the general test. While achieving approximately the same powers, the sample

α levels can be significantly improved by using the normal test. The scaled test and adjusted test (Bentler and Xie (2000)) were also run for the normal test with sample size $n = 1,000$. The powers are the same as shown here, and the actual α levels are 0.099 and 0.056, respectively. But its computation is a lot faster than sampling from $\hat{\Omega}$. Clearly, the normal test is the best for normal predictors.

Table 1. Example 1. The 2nd and 3rd columns are sample powers while the last is sample significance level.

	$d = 0$ vs $d \geq 1$		$d = 1$ vs $d \geq 2$		$d = 2$ vs $d \geq 3$	
	General	Normal	General	Normal	General	Normal
$n = 100$	0.668	0.633	0.163	0.112	0.100	0.006
$n = 500$	1.000	1.000	0.936	0.956	0.099	0.053
$n = 1,000$	1.000	1.000	1.000	1.000	0.099	0.050

Example 4.2. This model is the same as Example 4.2 in Bentler and Xie (2000). Let x_1, x_2, x_3, x_4 be i.i.d. $t(5)$ random variables, and ϵ be a standard normal. The model is

$$Y = \cos(2x_2) + 0.5\epsilon.$$

Thus $d = 1$. Table 2 lists the rejection rates at $\alpha = 0.05$ when replicated with 1,000 repeated simulation runs. Since Y is continuous, $H = 3$ and $H = 6$ are chosen for discretization. In Table 2, the numbers shown are for the general test with $H = 6$, the normal test with $H = 3$, and the general test with $H = 3$. With a large sample size, power is high. However, sample significance levels are very bad for either of these methods for non-normal predictors, again anitconservative. Generally, with a large number of slices, results are worse. This makes sense since the sample size in each slice becomes small as the number of slices increases.

Table 2. Example 2. The 2nd column is sample power while the last column is sample significance level.

	$d = 0$ vs $d \geq 1$			$d = 1$ vs $d \geq 2$		
	General $H = 3$	General $H = 6$	Normal $H = 3$	General $H = 3$	General $H = 6$	Normal $H = 3$
$n = 100$	0.467	0.436	0.556	0.172	0.198	0.127
$n = 500$	0.949	0.962	0.965	0.387	0.476	0.363
$n = 1,000$	0.997	0.998	0.998	0.451	0.680	0.455

Example 4.3. This model is the same as Example 4.2, except now i.i.d. standard normal random variables are used for the predictors. Table 3 lists the rejection rates at $\alpha = 0.05$ when replicated with 1,000 repeated simulation runs by the general test with $H = 3$. These simulations showed that powers are very high

with large sample sizes, while sample α levels compare well to the theoretical one.

Table 3. Example 3. The 2nd column is sample power while the last column is sample significance level.

	$d = 0$ vs $d \geq 1$	$d = 1$ vs $d \geq 2$
$n = 100$	0.731	0.067
$n = 500$	1.000	0.037
$n = 1,000$	1.000	0.037

From this simulation, it is clear that one should always try to transform predictors to approximate normality, which is desirable under (A) and (B); the normal test in Section 3.2 performs best when sampling from $\hat{\Omega}$ empirically; the number of slices should not be too big; and if there is uncertainty in deciding d using a numerical test, graphical methods such as those developed by Cook and Weisberg (1999) should be helpful.

Acknowledgement

The authors thank the Editor and two referees for their invaluable comments and suggestions that greatly improved the paper.

Appendix. Proofs

We sketch proofs, details can be found in tech report # 2004-15, in the Department of Statistics at the University of Georgia.

Proof of Lemma 1. Define $\mathbf{\Sigma}_{x|i} = \mathrm{Var}\,(\mathbf{X}|s = i)$ for the $(i+1)$th slice of Y in the population; the corresponding sample version is $\hat{\mathbf{\Sigma}}_{x|i}$, for $i = 0, \ldots, H$. Let $\mathbf{C}_n = (\hat{\mathbf{\Sigma}}_{x|1} - \hat{\mathbf{\Sigma}}_{x|0}, \ldots, \hat{\mathbf{\Sigma}}_{x|H} - \hat{\mathbf{\Sigma}}_{x|0})$. Also, define $pH \times pH$ matrices $\hat{\mathbf{G}} = diag(\sqrt{f_1} I_p, \ldots, \sqrt{f_H} I_p)$, and $\hat{\mathbf{S}} = diag(\hat{\mathbf{\Sigma}}_x^{-1/2}, \ldots, \hat{\mathbf{\Sigma}}_x^{-1/2})$. Let $\mathbf{C}, \mathbf{G}$ and $\mathbf{S}$ be the population versions of $\mathbf{C}_n, \hat{\mathbf{G}}$ and $\hat{\mathbf{S}}$, respectively. Let $\hat{\mathbf{A}} = \hat{\mathbf{\Sigma}}_x^{-1/2} \mathbf{\Sigma}_x^{1/2}$. Then

$$\sqrt{n}\mathbf{U}_n = \sqrt{n}\mathbf{\Gamma}_0^T(\hat{\mathbf{A}} - I_p + I_p)\mathbf{\Sigma}_x^{-1/2}(\mathbf{C}_n - \mathbf{C} + \mathbf{C})(I_H \otimes (\hat{\mathbf{A}} - I_p + I_p))\mathbf{S}(\hat{\mathbf{G}} - \mathbf{G} + \mathbf{G})\mathbf{\Psi}_0.$$

Collecting the terms of order $o(n^{-1/2})$, and using $\mathbf{\Gamma}_0^T \mathbf{K} = 0$ and $\mathbf{K}\mathbf{\Psi}_0 = 0$, the limiting distribution for $\sqrt{n}\mathbf{U}_n$ is the same as that for $\sqrt{n}\mathbf{\Gamma}_0^T \mathbf{\Sigma}_x^{-1/2}(\mathbf{C}_n - \mathbf{C})\mathbf{S}\mathbf{G}\mathbf{\Psi}_0$. Thus the distribution of $\mathbf{W}_n = \sqrt{n}(\mathbf{C}_n - \mathbf{C})$ must be investigated. Note that $\mathrm{vec}(\mathbf{W}_n) = (\mathbf{a}_1, \ldots, \mathbf{a}_s, \ldots, \mathbf{a}_H)'$ is a $p^2 H \times 1$ vector, where $\mathbf{a}_s = \sqrt{n}(\mathrm{vec}(\hat{\mathbf{\Sigma}}_{x|s}) - \mathrm{vec}(\mathbf{\Sigma}_{x|s}) - \mathrm{vec}(\hat{\mathbf{\Sigma}}_{x|0}) + \mathrm{vec}(\mathbf{\Sigma}_{x|0}))$. Let $\boldsymbol{\mu}_{x|s} = \mathrm{E}\,(\mathbf{X}|\tilde{Y} = s)$ and $\boldsymbol{\mu}_x = \mathrm{E}\,(\mathbf{X})$. By the Multivariate Central Limit Theorem, the multivariate version of Slutsky's Theorem, and the delta method, $\mathrm{vec}(\mathbf{W}_n) \to N(0, \Delta_x)$, where

$\Delta_x = (\Delta^x_{st})$ and $\Delta^x_{st} = \mathbf{B}_x(0) + \mathbf{B}_x(s)\delta_{st}$ with $\mathbf{B}_x(s) = (1/p_s)\mathrm{Var}\,((\mathbf{X} - \boldsymbol{\mu}_{x|s}) \otimes (\mathbf{X} - \boldsymbol{\mu}_{x|s})|s)$, $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$. Finally the result follows from some algebraic work.

Proof of Theorem 2. Under condition (a) $\mathbf{Z}|(Y = s) \sim N(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, so that $\mathrm{Var}\,((\mathbf{Z} - \boldsymbol{\mu}_s) \otimes (\mathbf{Z} - \boldsymbol{\mu}_s)|s) = 2\mathbf{N}_p(\boldsymbol{\Sigma}_s \otimes \boldsymbol{\Sigma}_s)$ (Mangus (1988)), where $2\mathbf{N}_p = I_{p^2} + \mathbf{K}_{pp}$. Thus $\Delta_s = (2\mathbf{N}_p/p_s)(\boldsymbol{\Sigma}_s \otimes \boldsymbol{\Sigma}_s)$. Let $\boldsymbol{\Psi}_0 = (\boldsymbol{\Psi}_{01}, \ldots, \boldsymbol{\Psi}_{0H})^T$. Simplify Δ^u to be

$$\Delta^u = \sum_{s,t=1}^{H} (\boldsymbol{\Psi}_{0s} \otimes \boldsymbol{\Gamma}_0^T)\sqrt{p_s p_t}\Delta_0(\boldsymbol{\Psi}_{0t}^T \otimes \boldsymbol{\Gamma}_0) + \sum_{s=1}^{H}(\boldsymbol{\Psi}_{0s} \otimes \boldsymbol{\Gamma}_0^T)p_s\Delta_s(\boldsymbol{\Psi}_{0s}^T \otimes \boldsymbol{\Gamma}_0).$$

The result follows from the basic properties of $\mathbf{K}_{pp}$, and the facts that $\boldsymbol{\Gamma}_0^T\boldsymbol{\mu}_s = 0$, $\boldsymbol{\Gamma}_0^T = \boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_s$ and $\sum_{s=1}^H \sqrt{p_s}\boldsymbol{\Psi}_{0s} = \sum_{s=1}^H \sqrt{p_s}\boldsymbol{\Sigma}_s\boldsymbol{\Psi}_{0s}$.

Under condition (b) $\mathbf{Z}|P\mathbf{Z} \sim N(P\mathbf{Z}, Q)$, where P is the orthogonal projection onto $\mathcal{S}(\mathbf{M})$ and $Q = I - P$. Using the form (Yin and Cook (2003)) $\mathcal{M}^{(3)}(\mathbf{Z}|Y) = \mathrm{E}\,((\mathbf{Z} - \mathrm{E}\,(\mathbf{Z}|Y)) \otimes (\mathbf{Z} - \mathrm{E}\,(\mathbf{Z}|Y))(\mathbf{Z} - \mathrm{E}\,(\mathbf{Z}|Y))^T)$, the result follows from a similar argument as in condition (a) above.

References

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

Bentler, P. M. and Xie, J. (2000). Corrections to test statistics in Principal Hessian directions. *Statist. Probab. Lett.* **47**, 381-389.

Cook, R. D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc.* **89**, 177-190.

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

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

Cook, R. D. (1998b). *Regression Graphics: Ideas for studying regressions through graphics.* Wiley, New York.

Cook, R. D. and Critchley, F. (2000). Detecting regression outliers and mixtures graphically. *J. Amer. Statist. Assoc.* **95**, 781-94.

Cook, R. D. and Lee, H. (1999). Dimension reduction in regressions with a binary response. *J. Amer. Statist. Assoc.* **94**, 1187-1200.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *J. Amer. Statist. Assoc.* **86**, 328-332.

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics.* Wiley, New York.

Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Austral. N. Z. J. Statist.* **43**, 147-199.

Cook, R. D. and Yin, X. (2002). Asymptotic Distribution for testing dimensionality in q-based PHD. *Statist. Probab. Lett.* **58**, 233-243.

Eaton, M. L.(1983). *Multivariate Statistics.* Wiley, New York.

Eaton, M. L. and Tyler, D. (1994). The asymptotic distribution of singular values with application to canonical correlations and correspondence analysis. *J. Multivariate Anal.* **50**, 238-264.

Ferré, L. (1998). Determining the dimension of sliced inverse regression and related methods. *J. Amer. Statist. Assoc.* **93**, 132-140.

Field, C. (1993). Tail areas of linear combinations of Chi-squares and non-central Chi-squares. *J. Statist. Comput. Simulation* **45**, 243-248.

Gannoun, A. and Saracco, J. (2003). Asymptotic theory for SIR_α method. *Statist. Sinica* **13**, 297-310.

Kötter, T. (1996). An asymptotic result for sliced inverse regression. *Comput. Statist.* **11**, 113-136.

Lange, K. (1999). *Numerical Analysis for Statisticians*. Springer-Verlag, New York.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-342.

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

Magnus, J. R. (1988). *Linear Structures*. Oxford University Press, New York.

Saracco, J. (1997). An asymptotic theory for sliced inverse regression. *Comm. Statist. Theory Methods* **26**, 2141-2171.

Saracco, J. (2001). Pooled slicing methods versus slicing methods. *Comm. Statist. Simulation Comput.* **30**, 489-511.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *J. Amer. Statist. Assoc.* **89**, 141-148.

Velilla, S. (1998) Assessing the number of linear components in a general regression problem. *J. Amer. Statist. Assoc.* **93**, 1088-1098.

Yin, X. and Cook, R. D. (2003). Estimating central subspace via inverse third moments. *Biometrika* **90**, 113-125.

Yin, X. and Cook, R. D. (2004). Asymptotic distribution of test statistic for the covariance dimension reduction methods in regression. *Statist. Probab. Lett.* **68**, 421-427.

Zhang, J. and Boos, D. D. (1992). Bootstrap critical values for testing homogeneity of covariance matrices. *J. Amer. Statist. Assoc.* **87**, 425-429.

Zhu, L. X. Ng, K. W. and Jing, P. (2002). Resampling methods for homogeneity tests of covariance matrices. *Statist. Sinica* **12**, 769-783.

Department of Statistics, 204 Statistics Building, University of Georgia, Athens, GA 30602, U.S.A.

E-mail: xryin@stat.uga.edu

Department of Statistics, 204 Statistics Building, University of Georgia, Athens, GA 30602, U.S.A.

E-mail: seymour@stat.uga.edu

(Received September 2003; accepted January 2005)