Extending SIR: The Weighted Chi-Square Test

Efstathia Bura[1] **R. Dennis Cook**[2]

Abstract

Sliced Inverse Regression (SIR) and an associated chi-square test for dimension were introduced by Li (1991) as a method for reducing the dimension of regression problems whose predictor variables are normal. In this paper the assumptions on the predictor distribution, under which the chi-square test was proved to apply, are relaxed and the result is extended. A general weighted chi-square test, which does not require normal regressors, for the dimension of a regression is given. Simulations exhibit that the weighted chi-square test is more reliable than the chi-square test when the regressor distribution digresses from normality significantly, and that it compares well with the chi-squared test when the regressors are normal.

KEY WORDS: Regression; Sliced inverse regression; Dimension estimation; Dimension reduction

1 Introduction

The overarching goal of a regression analysis is to understand how the conditional distribution of the univariate response Y given a vector $\mathbf{X}$ of p predictors depends on the value assumed by $\mathbf{X}$. While attention is often restricted to the mean function $\mathrm{E}(Y|\mathbf{X})$ and perhaps the variance function $\mathrm{Var}(Y|\mathbf{X})$, in full generality the object of interest is the conditional distribution of $Y|\mathbf{X}$.

Graphical displays can be quite useful for investigating $Y|\mathbf{X}$, especially when an adequate parsimoniously parameterized model is not available and when looking for

[1]Department of Statistics, The George Washington University, Washington, DC 20052

[2]School of Statistics, University of Minnesota, St. Paul, MN 55108

patterns in residuals. *Dimension reduction without loss of information* is a dominant theme of regression graphics: We try to reduce the dimension of $\mathbf{X}$ without losing information on $Y|\mathbf{X}$, and without requiring a model for $Y|\mathbf{X}$. Borrowing terminology from classical statistics, we call this *sufficient dimension reduction* which leads to the pursuit of *sufficient summary plots* containing all of the information on $Y|\mathbf{X}$ that is available from the sample.

Sliced inverse regression (SIR) is an innovative method due to Li (1991) for constructing summary plots. Informally, SIR provides a basis $\{\mathbf{b}_1, \dots, \mathbf{b}_p\}$ for $\mathbb{R}^p$ and corresponding *SIR predictors* $\{\mathbf{b}_1^T\mathbf{X}, \dots, \mathbf{b}_p^T\mathbf{X}\}$ which are ordered according to their likely importance to the regression. Thus, the first SIR predictor $\mathbf{b}_1^T\mathbf{X}$ is likely more important than the second $\mathbf{b}_2^T\mathbf{X}$, and so on. Plots of Y versus various combinations of the SIR predictors can provide useful information on the regression, the 3D plot of Y versus the first two SIR predictors often being the most informative.

It can be helpful to know if it is necessary to inspect plots involving all p SIR predictors, or if we can concentrate on the first few without loss of important information. Li (1991) provided a testing procedure to help determine the number of "significant SIR predictors." The procedure requires that $\mathbf{X}$ be normally distributed, which results in a chi-square reference distribution. This requirement can be worrisome in practice, because deviations from normality can influence the outcome.

We discuss in this article an alternative testing procedure that is based on the test statistic suggested by Li (1991), but does not require normally distributed predictors. Relaxing the normal requirement results in a reference distribution that is a weighted sum of chi-squares rather than a chi-square. The weights are unknown in general and thus must be estimated from the sample. This alternative procedure can be helpful in practice for two reasons: (a) it can confirm the indications of Li's chi-square test, increasing confidence in the results, and (b) it can help decide ambiguous situations.

In Section 2, the regression context is set out and SIR and the associated chi-square test are reviewed. In Section 3.1, the weighted chi-square test is developed, followed by a description of how it can be implemented in practice in Section 3.2. Special cases, resulting from the imposition of additional distributional assumptions, are explored in Section 3.3. There it is also shown that for the chi-square test to apply, joint normality of the predictors is not required, but rather constant second moments of the conditionals of the predictors would suffice. The mussel data are analyzed in Section 4. A simulation comparison of the two tests is given in Section 5. We conclude from those simulation results that the estimation of the weights does not compromise the advantages that the weighted chi-square test has over the chi-square test in cases with non-normal predictors. We conclude in Section 6 with a brief discussion.

2 Regression Context and SIR

Sufficient dimension reduction in regression focuses on finding $k \leq p$ linear combinations $\boldsymbol{\eta}_1^T\mathbf{X}, \ldots, \boldsymbol{\eta}_k^T\mathbf{X}$ that can replace $\mathbf{X}$ without loss of information and without requiring restrictive conditions on $Y|\mathbf{X}$. Letting $\boldsymbol{\eta}$ denote the $p \times k$ matrix with columns $\boldsymbol{\eta}_j$, we require that

$$Y \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\eta}^T\mathbf{X} \tag{1}$$

where the notation $\perp\!\!\!\perp$ indicates independence. The statement is thus that Y is independent of $\mathbf{X}$ given any value for $\boldsymbol{\eta}^T\mathbf{X}$.

For any vector or matrix $\boldsymbol{\alpha}$, let $S(\boldsymbol{\alpha})$ denote its range space with dimension $\dim(S(\boldsymbol{\alpha}))$. If (1) holds then it also holds with $\boldsymbol{\eta}$ replaced by any basis for $S(\boldsymbol{\eta})$. In this sense, (1) can be regarded as a statement about $S(\boldsymbol{\eta})$ rather than a statement about a basis $\boldsymbol{\eta}$ per se. Thus, when (1) holds we follow Li (1991, 1992) and call $S(\boldsymbol{\eta})$ a *dimension-reduction subspace* for the regression of Y on $\mathbf{X}$.

The usual objective of a dimension-reduction analysis is to infer about a *smallest* dimension-reduction subspace. While there is more than one way to define such a subspace, here we rely on the *central (dimension-reduction) subspace*, denoted by $S_{Y|\mathbf{X}}$, which was introduced by Cook (1994b, 1996). $S_{Y|\mathbf{X}}$ is the intersection of all dimension-reduction subspaces for the regression. It is trivially a subspace but it is not necessarily a dimension-reduction subspace. The existence of central subspaces can be assured by placing fairly weak restrictions on aspects of the joint distribution of Y and $\mathbf{X}$ (Cook 1994a, 1996, 1998ab). We assume throughout this article that the central subspace exists.

The central subspace $S_{Y|\mathbf{X}}$ is in effect a meta-parameter that is used to index the conditional distribution of Y given $\mathbf{X}$. The columns of the $p \times k$ matrix $\boldsymbol{\eta}$ will be used to represent a basis for the central space $S_{Y|\mathbf{X}}$. The plot of Y versus $\boldsymbol{\eta}^T\mathbf{X}$ is a *minimal sufficient summary plot* (Cook 1998a) for the regression of Y on $\mathbf{X}$.

There are several methods for estimating $S_{Y|\mathbf{X}}$, or portions thereof. In this article we are concerned only with SIR which we have found to be a useful first method in any dimension reduction inquiry.

In the rest of this article we work mostly in terms of the standardized predictor

$$\mathbf{Z} = \boldsymbol{\Sigma}_x^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$$

where $\boldsymbol{\Sigma}_x = \mathrm{Var}(\mathbf{X}) > 0$. This facilitates development and involves no loss of generality because we can always back-transform to the original scale since $S_{Y|\mathbf{X}} = \boldsymbol{\Sigma}_x^{-1/2} S_{Y|\mathbf{Z}}$. Additionally, the SIR predictors reviewed in this section are invariant under full rank linear transformations of $\mathbf{X}$. Consequently, summary plots will look the same whether we work in the scale of $\mathbf{X}$ or $\mathbf{Z}$.

Let the columns $\boldsymbol{\beta}_j$ of the $p \times k$ matrix $\boldsymbol{\beta}$ be a basis for the central subspace $S_{Y|\mathbf{Z}}$. For use in practice, $\mathbf{Z}$ is constructed by replacing $\boldsymbol{\Sigma}_x$ and $\mathrm{E}(\mathbf{X})$ with the usual moment estimates. SIR requires that the conditional expectation $\mathrm{E}(\mathbf{Z}|\boldsymbol{\beta}^T\mathbf{Z})$ be linear. This is equivalent to requiring that $\mathrm{E}(\mathbf{Z}|\boldsymbol{\beta}^T\mathbf{Z}) = \mathbf{P}_{\boldsymbol{\beta}}\mathbf{Z}$ where $\mathbf{P}_{\boldsymbol{\beta}}$ is the orthogonal projection operator for $S(\boldsymbol{\beta}) = S_{Y|\mathbf{Z}}$ with respect to the usual inner product (Cook 1998b, p. 57). We will refer to this restriction as the *linearity condition*. The linearity condition involves only the marginal distribution of $\mathbf{Z}$, and is required to hold only for the basis $\boldsymbol{\beta}$. While minor deviations from the linearity condition may not matter much, substantial deviations can produce misleading results.

SIR makes use of the *inverse regression subspace* $S_{\mathrm{E}(\mathbf{Z}|Y)}$ defined as the span of $\mathrm{E}(\mathbf{Z}|Y)$ as Y varies in its marginal sample space. Under the linearity condition,

$$S_{\mathrm{E}(\mathbf{Z}|Y)} \subseteq S_{Y|\mathbf{Z}} \tag{2}$$

If Y is discrete, this result can be used to justify using conditional sample means $\hat{\mathrm{E}}(\mathbf{Z}|Y)$ as estimates of vectors in the central subspace. When Y is continuous, Li (1991) suggested replacing Y with a discrete version $\tilde{Y}$ based on partitioning the observed range of Y into H fixed, non-overlapping slices. By (1), $\tilde{Y} \perp\!\!\!\perp \mathbf{Z}|\boldsymbol{\beta}^T\mathbf{Z}$ and thus $S_{\tilde{Y}|\mathbf{Z}} \subseteq S_{Y|\mathbf{Z}}$. In addition, provided that H is sufficiently large, $S_{\tilde{Y}|\mathbf{Z}} = S_{Y|\mathbf{Z}}$, and there is no loss of information when Y is replaced by $\tilde{Y}$. Using (2), SIR can be justified on the following line of reasoning,

$$S\{\mathrm{Cov}[\mathrm{E}(\mathbf{Z}|\tilde{Y})]\} = S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})} \subseteq S_{\tilde{Y}|\mathbf{Z}} \subseteq S_{Y|\mathbf{Z}} \tag{3}$$

where the first equality follows from Proposition 2.7 in Eaton (1983, p. 75. See also Cook 1998b, Proposition 11.1.).

Thus, $S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})}$ is spanned by the eigenvectors corresponding to the nonzero eigenvalues of $\mathbf{V} = \mathrm{Cov}[\mathrm{E}(\mathbf{Z}|\tilde{Y})]$. In consequence, an estimate of $S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})}$ can be constructed from the estimated covariance matrix $\hat{\mathbf{V}}$ by using its eigenvectors corresponding to its eigenvalues which are inferred to be nonzero in the population.

2.1 SIR Algorithm

Given a random sample $\{(Y_i, \mathbf{X}_i^T)\}_{i=1}^n$ from $(Y, \mathbf{X}^T)$, begin by constructing a discrete version $\tilde{Y}$ of Y. The number of slices H is chosen much like a tuning parameter in smoothing (see Li 1991 for further discussion). Next, form the intra-slice means $\bar{\mathbf{Z}}_s$, $s = 1, \ldots, H$, of the sample version of the standardized predictor vector. Finally, form the sample version $\hat{\mathbf{V}}$ of $\mathbf{V} = \mathrm{Cov}[\mathrm{E}(\mathbf{Z}|\tilde{Y})]$,

$$\hat{\mathbf{V}} = \sum_{s=1}^{H} f_s \bar{\mathbf{Z}}_s \bar{\mathbf{Z}}_s^T = \hat{\mathbf{Z}}_n \hat{\mathbf{Z}}_n^T$$

where $f_s = n_s/n$ is the fraction of observations falling in slice s and

$$\hat{\mathbf{Z}}_n = (\bar{\mathbf{Z}}_1 \sqrt{f_1}, \ldots, \bar{\mathbf{Z}}_H \sqrt{f_H}) \tag{4}$$

is the $p \times H$ matrix of weighted intra-slice sample means of the standardized predictors. Letting $\hat{\boldsymbol{\beta}}_j$ denote the j-th eigenvector of $\hat{\mathbf{V}}$, the j-th SIR predictor is constructed as $\hat{\boldsymbol{\beta}}_j^T \mathbf{Z}$, $j = 1, ..., p$.

Let $d = \dim(S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})})$. The span of the d eigenvectors $\hat{\boldsymbol{\beta}}_j$, $j = 1, ..., d$, of $\hat{\mathbf{V}}$ that correspond to its first d eigenvalues $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_d > \ldots \geq \hat{\lambda}_p$ is an estimate of $S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})} \subseteq S_{Y|\mathbf{Z}}$. Under the assumption of normal predictors, Li (1991) proved that the test statistic

$$\hat{\Lambda}_d = n \sum_{j=d+1}^{p} \hat{\lambda}_j \tag{5}$$

has an asymptotic chi-square distribution with $(p-d)(H-d-1)$ degrees of freedom. Consequently, the number d of nonzero eigenvalues of $\mathbf{V}$, which is also the dimension of the subspace $S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})}$, can be estimated as follows: The null hypothesis that $d = j$

is rejected in favor of the alternative that $d \geq j+1$ when $\hat{\Lambda}_j$ is larger than the percentage points of a chi-square distribution with $(p-j)(H-j-1)$ degrees of freedom.

Other testing techniques that use the same simple nonparametric estimation method as Li (1991) have been developed. Schott (1994) proposed a test which requires elliptically symmetric regressors, and for which the tuning constant is the number of observations c per slice as opposed to the number of slices H in Li (1991). In order to obtain the asymptotic distribution of his test statistic, Schott lets c go to infinity. Velilla (1998) introduced another testing method, which does not impose restrictions on the regressor distribution, where c is fixed and the number of slices H varies. Ferré (1998) provided a model selection approach to the estimation of dimension by studying the quality of estimation through a measure of closeness between $S_{\mathrm{E}(\mathbf{X}|\tilde{Y})}$ and its estimates.

Our method extends SIR's applicability by lifting the normality assumption on the regressors and replacing it by the minimal requirement of finite second moments. The tuning parameter is the number of slices and is assumed to be fixed, as in SIR. The only comparable method among the ones listed above is Velilla's as he also does not restrict the regressor distribution. Nevertheless, he uses the number of points per slice as a tuning parameter and imposes several regularity conditions on both the regressor distribution and the regression curve $\mathrm{E}(Y|\mathbf{X})$.

3 The weighted chi-square test

Various results on the distribution of $\hat{\Lambda}_d$ are presented in this section. The development of the general weighted chi-square test in Section 3.1 requires only that the data be independent and identically distributed with finite first two moments. The general logic behind this development is similar to that used by Cook (1998a) in the derivation of the asymptotic distribution of statistics used in the method of principal Hessian Directions (Li 1992). However, the details here are different because the methods and statistics are different. Subsequent results in Section 3.3 require various additional conditions.

3.1 Development

Letting p_s denote the probability that Y is in slice s, the matrix of weighted intra-slice means $\hat{\mathbf{Z}}_n$ converges almost surely to

$$\mathbf{B} = (\mathrm{E}(\mathbf{Z}|\tilde{Y} = 1)\sqrt{p_1}, \dots, \mathrm{E}(\mathbf{Z}|\tilde{Y} = H)\sqrt{p_H}) \tag{6}$$

with singular value decomposition

$$\mathbf{B} = \mathbf{\Gamma}_1 \begin{bmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{\Gamma}_2^T$$

where $\mathbf{\Gamma}_1$ is an orthonormal $p \times p$ matrix, $\mathbf{\Gamma}_2$ is an orthonormal $H \times H$ matrix, and $\mathbf{D}$ is a $d \times d$ diagonal matrix with the positive singular values of $\mathbf{B}$ along its diagonal. Recalling that $d = \dim(S_{\mathrm{E}(\mathbf{Z}|Y)})$, partition $\mathbf{\Gamma}_1 = (\mathbf{\Gamma}_{11}, \mathbf{\Gamma}_{12})$, where $\mathbf{\Gamma}_{11}$: $p \times d$, $\mathbf{\Gamma}_{12}$: $p \times (p-d)$, and partition

$$\mathbf{\Gamma}_2^T = \begin{bmatrix} \mathbf{\Gamma}_{21}^T \\ \mathbf{\Gamma}_{22}^T \end{bmatrix} \tag{7}$$

where $\mathbf{\Gamma}_{21}^T : d \times H$, $\mathbf{\Gamma}_{22}^T : (H-d) \times H$. The columns of $\mathbf{\Gamma}_{11}$ form a basis for $S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})} = S(\mathbf{V})$.

Given an $r \times c$ matrix $\mathbf{A} = (a_1, \dots, a_c)$, let $\mathrm{vec}(\mathbf{A})$ denote the $rc \times 1$ vector constructed by stacking the columns of $\mathbf{A}$ one after another, i.e., $\mathrm{vec}(\mathbf{A})^T = (a_1^T, \dots, a_c^T)$.

Let $\tilde{\mathbf{Z}}_n = \sqrt{n}(\hat{\mathbf{Z}}_n - \mathbf{B})$ where $\hat{\mathbf{Z}}_n$ and $\mathbf{B}$ are defined in (4) and (6), respectively. By the multivariate central limit theorem and the multivariate version of Slutsky's theorem, the $pH \times 1$ vector $\mathrm{vec}(\tilde{\mathbf{Z}}_n)$ converges to a multivariate normal distribution with mean 0 and covariance matrix $\mathbf{\Delta}$:

$$\mathrm{vec}(\tilde{\mathbf{Z}}_n) \xrightarrow{\mathcal{D}} N_{pH}(0, \mathbf{\Delta}) \quad \text{as } n \to \infty \tag{8}$$

It follows from Eaton and Tyler (1994) that the limiting distribution of the smallest $\min(p-d, H-d)$ singular values of $\tilde{\mathbf{Z}}_n$ is the same as the limiting distribution of the corresponding singular values of the $(p-d) \times (H-d)$ matrix

$$\sqrt{n}\, \mathbf{\Gamma}_{12}^T \hat{\mathbf{Z}}_n \mathbf{\Gamma}_{22} \tag{9}$$

The operation applied to $\hat{\mathbf{Z}}_n$ in (9) selects the space spanned by the eigenvectors corresponding to the $p-d$ smallest eigenvalues of $\mathbf{V}$. This space coincides with the space spanned by the eigenvectors corresponding to the largest $p-d$ eigenvalues of $\mathrm{E}(\mathrm{Cov}(\mathbf{Z}|Y))$ (see remarks 3.1 and 5.3 in Li (1991)). Hsing and Carroll (1992) obtained the asymptotic normality of their estimate of $\mathrm{E}(\mathrm{Cov}(\mathbf{Z}|Y))$, which was constructed using two observations per slice, by imposing restrictions on both the regressor and the conditional distribution of $\mathbf{X}|Y$. Their results were generalized to more than two observations per slice by Zhu and Ng (1995).

The asymptotic distribution of $\hat{\Lambda}_d$ is the same as that of the sum of the squares of the singular values of (9), which can be expressed as $n \times \mathrm{trace}[\mathbf{\Gamma}_{12}^T \hat{\mathbf{Z}}_n \mathbf{\Gamma}_{22} (\mathbf{\Gamma}_{12}^T \hat{\mathbf{Z}}_n \mathbf{\Gamma}_{22})^T]$. Therefore, the asymptotic distribution of $\hat{\Lambda}_d$ is the same as the asymptotic distribution of $n\, \mathrm{vec}(\mathbf{\Gamma}_{12}^T \hat{\mathbf{Z}}_n \mathbf{\Gamma}_{22})^T\, \mathrm{vec}(\mathbf{\Gamma}_{12}^T \hat{\mathbf{Z}}_n \mathbf{\Gamma}_{22})$. By (8), the limiting distribution of the vector version of (9) is

$$\sqrt{n}\ \mathrm{vec}(\mathbf{\Gamma}_{12}^T \hat{\mathbf{Z}}_n \mathbf{\Gamma}_{22}) \xrightarrow{\mathcal{D}} N_{(H-d)(p-d)}(0, (\mathbf{\Gamma}_{22}^T \otimes \mathbf{\Gamma}_{12}^T) \mathbf{\Delta} (\mathbf{\Gamma}_{22} \otimes \mathbf{\Gamma}_{12})) \tag{10}$$

where $\otimes$ denotes the Kronecker product. Finally, the asymptotic distribution of $\hat{\mathbf{\Lambda}}_d$ is given in the following theorem.

Theorem 1 *Assume that $H > d+1$ and $p > d$. Then the limiting distribution of $\hat{\Lambda}_d$ is the same as that of*

$$C = \sum_{k=1}^{(p-d)(H-d)} \omega_k C_k$$

where the C_k's are independent chi-square random variables each with 1 degree of freedom, $\omega_1 \geq \omega_2 \geq \ldots \geq \omega_{(p-d)(H-d)}$ are the ordered eigenvalues of

$$\mathbf{\Delta}_C = (\mathbf{\Gamma}_{22}^T \otimes \mathbf{\Gamma}_{12}^T)\mathbf{\Delta}(\mathbf{\Gamma}_{22} \otimes \mathbf{\Gamma}_{12}) \tag{11}$$

and $\otimes$ denotes the Kronecker product.

Theorem 1 allows for a general test of dimension provided we can obtain a consistent estimate of $\mathbf{\Delta}_C$ from which to construct estimates of the eigenvalues ω_k. To this end, it is useful to investigate the form of $\mathbf{\Delta}$.

$\mathbf{\Delta}$ can be represented as an $H \times H$ array of $p \times p$ matrices $\mathbf{\Delta}_{ts}$ by partitioning according to the intra-slice averages $\bar{\mathbf{Z}}_j$. Letting $\mathbf{\Sigma}_{u|s} = \text{Cov}(\mathbf{U}|\tilde{Y} = s)$ denote the conditional covariance matrix of a random vector $\mathbf{U}$ in slice s, the various sub-matrices $\mathbf{\Delta}_{ts}$ can be computed as follows. For $t = s$,

$$\mathbf{\Delta}_{ss} = np_s \mathbf{\Sigma}_x^{-1/2} \,\text{Cov}(\bar{\mathbf{X}}_s - \bar{\mathbf{X}}) \mathbf{\Sigma}_x^{-1/2}$$

But,

$$\begin{aligned}
\text{Cov}(\bar{\mathbf{X}}_s - \bar{\mathbf{X}}) &= \text{Cov}(\bar{\mathbf{X}}_s) - 2\,\text{Cov}(\bar{\mathbf{X}}_s, \bar{\mathbf{X}}) + \text{Cov}(\bar{\mathbf{X}}) \\
&= \frac{\mathbf{\Sigma}_{x|s}}{n_s} - 2\,\text{Cov}\left(\bar{\mathbf{X}}_s, \sum_{s=1}^{H} p_s \bar{\mathbf{X}}_s\right) + \frac{\mathbf{\Sigma}_x}{n} \\
&= \frac{1}{n}\big(\mathbf{\Sigma}_x + \frac{1-2p_s}{n_s}\mathbf{\Sigma}_{x|s}\big)
\end{aligned}$$

Thus,

$$\mathbf{\Delta}_{ss} = \mathbf{I}_p p_s + (1-2p_s)\mathbf{\Sigma}_{z|s} \tag{12}$$

where $\mathbf{\Sigma}_{z|s} = \mathbf{\Sigma}_x^{-1/2}\mathbf{\Sigma}_{x|s}\mathbf{\Sigma}_x^{-1/2}$. For $t \neq s$,

$$\begin{aligned}\mathbf{\Delta}_{ts} &= n\sqrt{p_t p_s}\,\mathbf{\Sigma}_x^{-1/2}\,\mathrm{Cov}(\bar{\mathbf{X}}_t - \bar{\mathbf{X}}, \bar{\mathbf{X}}_s - \bar{\mathbf{X}})\mathbf{\Sigma}_x^{-1/2} \\ &= \sqrt{p_t p_s}(\mathbf{I}_p - \mathbf{\Sigma}_{z|t} - \mathbf{\Sigma}_{z|s}) \qquad (13)\end{aligned}$$

since

$$\mathrm{Cov}(\bar{\mathbf{X}}_t - \bar{\mathbf{X}}, \bar{\mathbf{X}}_s - \bar{\mathbf{X}}) = \frac{1}{n}(\mathbf{\Sigma}_x - \mathbf{\Sigma}_{x|t} - \mathbf{\Sigma}_{x|s})$$

Clearly, $\mathbf{\Delta}$ can be estimated consistently by substituting sample versions of p_s, $\mathbf{\Sigma}_{x|s}$, $s = 1, \dots H$, and $\mathbf{\Sigma}_x$. Additionally, under a hypothesized value of d, $\mathbf{\Gamma}_{22}$ and $\mathbf{\Gamma}_{12}$ can be estimated consistently by using the sample versions computed from $\hat{\mathbf{Z}}_n$. Let $\hat{\mathbf{\Delta}}_C$ denote the resulting estimated version of $\mathbf{\Delta}_C$, and let $\hat{\omega}_k$ denote the eigenvalues of $\hat{\mathbf{\Delta}}_C$. The limiting distribution of $\hat{\Lambda}_d$ is then consistently estimated to be the same as that of

$$\hat{C} = \sum_{k=1}^{(p-d)(H-d)} \hat{\omega}_k C_k \qquad (14)$$

These results can be used to estimate $d = \dim(S_{\mathrm{E}(\mathbf{Z}|Y)})$ just as described for Li's chi-square test procedure.

3.2 Test algorithm

In this section, we give a methodological version of the test of $d = j$ implied by the developments in the previous section.

1. From the singular value decomposition of $\hat{\mathbf{Z}}_n$, let $\hat{\mathbf{\Gamma}}_1$ denote the orthonormal $p \times p$ matrix of *left* singular vectors, and let $\hat{\mathbf{\Gamma}}_2$ denote the orthonormal $H \times H$ matrix of *right* singular vectors. This step is the same for all values of j.

2. Construct $\hat{\mathbf{\Gamma}}_{22}$ as the matrix of the last $H - j$ columns of $\hat{\mathbf{\Gamma}}_2$, and construct $\hat{\mathbf{\Gamma}}_{12}$ as the last the last $p - j$ columns of $\hat{\mathbf{\Gamma}}_1$.

3. Construct $\hat{\mathbf{\Delta}}$ by using estimates $\hat{\mathbf{\Delta}}_{st}$ of its submatrices from (12) for $s = t$ and (13) for $s \neq t$.

4. Use the results from steps 2 and 3 to construct $\hat{\mathbf{\Delta}}_C$, and find its eigenvalues $\hat{\omega}_k$.

5. Construct the p-value for the weighted chi-square test as

$$p\text{-value} = \Pr(\hat{C} > \text{observed } \hat{\Lambda}_j)$$

where the distribution of $\hat{C}$ is given by (14). There is a substantial literature to assist in computing the tail probabilities of linear combinations of chi-squared random variables. See Field (1993) for an introduction.

The linearity condition, which is an essential part of the justification for SIR, is not required for Theorem 1, which provides a general method for inferring about $\dim(S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})})$. However, the linearity condition is required for (3) which establishes the necessary relation between the inverse regression subspace and the central subspace.

When the number of observations per slice is small, the variation in the estimated weights $\hat{\omega}$ might mitigate the usefulness of the estimated distribution of $\hat{\Lambda}_d$. For this reason it may be useful to know if there are special cases in which the limiting distribution simplifies. We address such cases in the next section. These cases may help us gain a better feeling for the relative merits of the two test procedures under consideration.

3.3 Special cases

The following lemma will be useful later in this section.

Lemma 1 *Let* $\mathbf{g}$ *denote the* $H \times 1$ *vector with elements* $\sqrt{p_s}$, $s = 1, \ldots, H$. *Then* $\|\mathbf{g}\| = 1$ *and* $\mathbf{g} \in S(\mathbf{\Gamma}_{22})$, *where* $\mathbf{\Gamma}_{22}$ *is as defined in (7).*

This lemma follows from the singular value decomposition of $\mathbf{B}$ near (6) and the fact that $\mathbf{Bg} = \mathrm{E}(\mathbf{Z}) = 0$. The next result, stated as a corollary to Theorem 1, gives the first simplified result on the distribution of $\hat{\Lambda}_d$.

Corollary 1 *Assume that the conditional covariance matrices* $\mathbf{\Sigma}_{z|s}$ *are constant across slices. Then the limiting distribution of* $\hat{\Lambda}_d$ *is chi-squared with* $(p-d)(H-d-1)$ *degrees of freedom.*

Justification: Since the conditional covariance matrices are constant, $\mathbf{\Sigma}_{z|s} = \mathbf{I}_p$ for $s = 1, \ldots, H$. Substituting this into the form for $\mathbf{\Delta}$ given previously yields $\mathbf{\Delta} = \mathbf{I}_{pH} - \mathbf{gg}^T \otimes \mathbf{I}_p$ and thus

$$\mathbf{\Delta}_C = \mathbf{I}_{(p-d)(H-d)} - \mathbf{\Gamma}_{22}^T \mathbf{gg}^T \mathbf{\Gamma}_{22} \otimes \mathbf{I}_{p-d}$$

which, from Lemma 1, is a symmetric idempotent matrix. The conclusion follows because $\mathrm{trace}(\mathbf{\Delta}_C) = (p-d)(H-d-1)$. $\Box$

We next develop another set of conditions under which $\hat{\Lambda}_d$ is asymptotically chi-squared. Let $\mathbf{P}_{11}$ be the orthogonal projection operator onto $S(\mathbf{\Gamma}_{11})$, where $\mathbf{\Gamma}_{11}$ is the basis for $S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})}$ defined near (7). The linearity condition of Theorem 1 is equivalent to $\mathrm{E}(\mathbf{Z}|\mathbf{\Gamma}_{11}^T\mathbf{Z}) = \mathbf{P}_{11}\mathbf{Z}$. Also, assume that

$$\mathrm{Cov}(\mathbf{Z}|\mathbf{\Gamma}_{11}^T\mathbf{Z}) = \mathbf{I}_p - \mathbf{P}_{11} \tag{15}$$

Let $\mathbf{\Gamma}_{22} = (\gamma_{ts})$, $t = 1, \ldots, H$, $s = 1, \ldots, H-d$, denote the elements of $\mathbf{\Gamma}_{22}$. Then, the (s,k)th block of $\mathbf{\Delta}_C$ in (11) is found by taking the product of the sth row block of $\mathbf{\Gamma}_{22}^T \otimes \mathbf{\Gamma}_{12}^T$ with $\mathbf{\Delta}$ and the kth column block of $\mathbf{\Gamma}_{22} \otimes \mathbf{\Gamma}_{12}$:

$$[\mathbf{\Delta}_C]_{sk} = \sum_{l}^{H} \sum_{t}^{H} \gamma_{ts}\gamma_{lk} \mathbf{\Gamma}_{12}^T \mathbf{\Delta}_{tl} \mathbf{\Gamma}_{12}$$

For $t = l$,

$$\mathbf{\Gamma}_{12}^T \mathbf{\Delta}_{tt} \mathbf{\Gamma}_{12} = \mathbf{\Gamma}_{12}^T \mathbf{\Gamma}_{12} p_t + (1 - 2p_t) \mathbf{\Gamma}_{12}^T \mathbf{\Sigma}_{z|t} \mathbf{\Gamma}_{12} \tag{16}$$

Next,

$$\begin{aligned}\mathbf{\Sigma}_{z|t} &= \mathrm{E}_{z|t}[\mathrm{Cov}(\mathbf{Z}|\mathbf{\Gamma}_{11}^T \mathbf{Z}, \tilde{Y} = t)] + \mathrm{Cov}_{z|t}[\mathrm{E}(\mathbf{Z}|\mathbf{\Gamma}_{11}^T \mathbf{Z}, \tilde{Y} = t)] \\ &= \mathrm{E}_{z|t}[\mathrm{Cov}(\mathbf{Z}|\mathbf{\Gamma}_{11}^T \mathbf{Z})] + \mathrm{Cov}_{z|t}[\mathrm{E}(\mathbf{Z}|\mathbf{\Gamma}_{11}^T \mathbf{Z})]\end{aligned}$$

where the second equality follows because $Y \perp\!\!\!\perp \mathbf{Z}|\mathbf{\Gamma}_{11}^T \mathbf{Z}$. By the linearity condition and (15), we have

$$\mathbf{\Gamma}_{12}^T \,\mathrm{Cov}_{z|t}(\mathrm{E}(\mathbf{Z}|\mathbf{\Gamma}_{11}^T \mathbf{Z})) \mathbf{\Gamma}_{12} = \mathrm{Cov}_{z|t}(\mathbf{\Gamma}_{12}^T \mathbf{P}_{11} \mathbf{Z}) = 0$$

and

$$\mathbf{\Gamma}_{12}^T \,\mathrm{Cov}_{z|t}(\mathbf{Z}|\mathbf{\Gamma}_{11}^T \mathbf{Z}) \mathbf{\Gamma}_{12} = \mathbf{\Gamma}_{12}^T (\mathbf{I}_p - \mathbf{P}_{11}) \mathbf{\Gamma}_{12} = \mathbf{I}_{p-d}$$

Thus,

$$\mathbf{\Gamma}_{12}^T \mathbf{\Delta}_{tt} \mathbf{\Gamma}_{12} = (1 - p_t) \mathbf{I}_{p-d}$$

Similarly, $\mathbf{\Gamma}_{12}^T \mathbf{\Delta}_{tl} \mathbf{\Gamma}_{12} = -\sqrt{p_t p_l} \mathbf{I}_{p-d}$. Using Lemma 1 and $\mathbf{\Gamma}_{22}^T \mathbf{\Gamma}_{22} = \mathbf{I}_{H-d}$ we obtain:

Theorem 2 *Let $\mathbf{P}_{11}$ be the orthogonal projection operator onto $S(\mathbf{\Gamma}_{11})$, where $\mathbf{\Gamma}_{11}$ is the basis for $S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})}$ defined near (7). Assume that*

1. $Y \perp\!\!\!\perp \mathbf{Z}|\mathbf{\Gamma}_{11}^T \mathbf{Z}$
2. $\mathrm{E}(\mathbf{Z}|\mathbf{\Gamma}_{11}^T \mathbf{Z}) = \mathbf{P}_{11} \mathbf{Z}$
3. $\mathrm{Cov}(\mathbf{Z}|\mathbf{\Gamma}_{11}^T \mathbf{Z}) = \mathbf{I}_p - \mathbf{P}_{11}$
4. $H > d + 1$ *and* $p > d$

Then $\hat{\Lambda}_d$ has an asymptotic chi-squared distribution with $(p-d)(H-d-1)$ degrees of freedom.

Condition 1 of this theorem ensures that the inverse mean function depends on $\tilde{Y}$ only via $\mathbf{\Gamma}_{11}^T\mathbf{Z}$. It implies that SIR will not miss any part of the central dimension-reduction subspace. Condition 2 is the linearity assumption, expressed in terms of the projection onto $S_{\mathrm{E}(\mathbf{Z}|\tilde{Y})}$. Condition 4 is imposed to avoid trivial cases. Condition 3 is necessary to obtain an asymptotic chi-square distribution. It is always satisfied when the original predictors are independent and $\mathbf{\Gamma}_{11} = (\mathbf{I}_d, 0)$, regardless of the distribution of $\mathbf{X}$.

Conditions 1 through 3 are obviously satisfied when Y is stochastically independent of $\mathbf{X}$, so that $d = 0$. This means that SIR can be used as a diagnostic test for goodness of fit of a postulated model by using residuals in place of the response.

Li (1991) computed the asymptotic distribution of $\hat{\Lambda}_d$ for only the case in which the vector of predictors $\mathbf{X}$ has a p-variate normal distribution, and then Conditions 2 and 3 are obviously satisfied. Further, if Condition 3 holds and $\mathbf{X}$ is elliptically contoured, as might be required to ensure Condition 2 for all possible $\mathbf{\Gamma}_{11}$, then $\mathbf{X}$ must be multivariate normal (Kelker 1970; Cambanis, Huang and Simons 1981).

If, as required for Corollary 1, $\mathbf{\Sigma}_{z|s}$ is constant across slices and Conditions 1, 2 and 4 hold, then it can be shown that Condition 3 is again satisfied. This is an important observation. Li (1991) said that constant conditional variance across slices is not required for his result to hold. But, in his proof, he assumes normal regressors and non-constant variance is not an issue.

In effect, Corollary 1 and Theorem 2 state that the defining condition that needs to be met for the chi-square test to apply, is constant covariance of the conditional distribution of $\mathbf{Z}|Y$, or of the conditionals of the regressor distribution, and not normality of the regressors. In practice, both conditions can be checked by visual inspection of the scatterplot matrix of the regressors and the response.

It is also worth pointing out that if $S_{\mathrm{E}(\mathbf{Z}|Y)}$ is a proper subspace of $S_{Y|\mathbf{Z}}$, then $\hat{\Lambda}_d$ may no longer have a chi-square distribution. Of course, even in this case, if the covariance structure of $\mathbf{X}$ is constant across slices, Corollary 1 implies that $\hat{\Lambda}_d$ has an asymptotic chi-square distribution with $(p-d)(H-d-1)$ degrees of freedom.

4 An Illustrative Example: The Mussel Data

Consider a regression that arose as part of an ecological study of New Zealand mussels. The response is a mussel's muscle mass M, measured in g, the edible part of the mussel. The $p = 4$ predictors are the shell length L, height Ht, and width W, each in mm, and the shell mass S in g. There are $n = 172$ observations. Portions of these data were used for illustration by Cook and Weisberg (1994) and Cook (1998b).

We noticed substantial curvature in the cells of a scatterplot matrix of the predictors, particularly cells involving shell mass S. We compensated for this nonlinearity by simultaneously estimating power transformations of the predictors so that their joint distribution might be approximately normal on the transformed scale. This resulted in the logarithmic transformation of S and W. Of course, there is no guarantee that the transformed predictors are jointly normal. With the linearity condition reasonably satisfied, we proceeded with the SIR analysis.

The chi-square p-values are shown for $j = 0, 1, 2$ in the fourth column of Table 1. The results indicate that $d = 2$ and thus that the first two SIR predictors are required to characterize the regression. However, we were unable to confirm visually the need for two SIR predictors by using the corresponding visualization tools described in Cook and Weisberg (1994, Chapter 6) and Cook (1998b, Chapter 4) and which indicated instead that only the first SIR predictor is needed. One possible

Table 1: Test results from application of SIR with $H = 8$ to the mussel regression.

j	$\hat{\Lambda}_j$	DF	χ^2 p-value	weighted χ^2 p-value
0	164.9	28	0.000	0.000
1	29.0	18	0.048	0.183
2	11.9	10	0.292	0.290

explanation could involve the assumption of normally distributed predictors that Li (1991) used in the derivation of his chi-square test procedure. We resolved the conflict by using the weighted chi-square test described in this article.

The p-values from the weighted chi-square test applied to the mussel regression are shown in the fifth column of Table 1. The resulting inference, which agrees with our visual assessment, is that $d = 1$, suggesting that the chi-square test may have responded to non-normality in the predictors. The estimated limiting distribution of $\hat{\Lambda}_d$ that we used for this inference should be useful when the number of observations per slice is large enough to allow reasonable estimation of the intra-slice covariance matrices $\mathbf{\Sigma}_{z|s}$. We used $H = 8$ slices leaving about 22 observations per slice to estimate the 4×4 intra-slice covariance matrices.

5 Simulation Results

The power of the chi-square test and the weighted chi-square test will be compared via a simulation study. For all regression models to be considered in this section, three sample sizes are used: $n = 100, 200, 300$. For each sample size and each distribution of the regressor vector $\mathbf{X}$, the p-values corresponding to the test statistics for selected dimensions for both tests were collected over 1,000 replications.

Two models, both of dimension 2, are considered. First, the response Y is generated according to the following model.

$$Y = (4 + X_1)(2 + X_2 + X_3) + 0.5\epsilon \tag{17}$$

where ϵ is a standard normal variate and

$$\begin{aligned} X_1 &= W_1 & X_4 &= V_2 + V_3 \\ X_2 &= V_1 + \frac{W_2}{2} & X_5 &= V_2 - V_3 \\ X_3 &= -V_1 + \frac{W_2}{2} \end{aligned} \tag{18}$$

The only restriction placed on $\mathbf{V}$ and $\mathbf{W}$ is that they be independent. The variables V_1, V_2, V_4 are i.i.d. $t_{(4)}$, $V_3 \sim t_{(3)}$, $V_5 \sim t_{(5)}$, and W_1, W_2 are i.i.d. Gamma(0.25) random variables.

The second model is

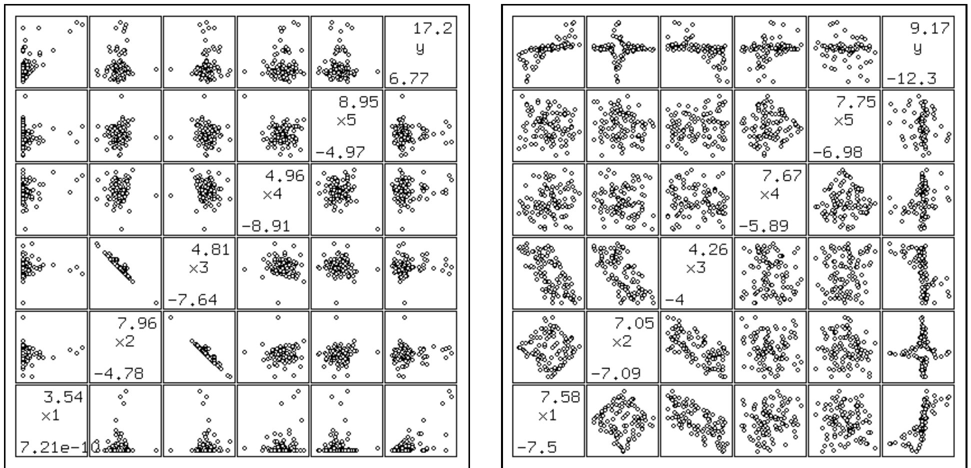
$$Y = \frac{X_1}{0.5 + (X_2 + 1.5)^2} + 0.5\epsilon \tag{19}$$

where ϵ is a standard normal variate and

$$\begin{aligned} X_1 &= V_3 + V_4 + \frac{W}{6} & X_4 &= V_1 + V_2 \\ X_2 &= -V_3 + V_4 + \frac{W}{6} & X_5 &= -V_1 + V_2 \\ X_3 &= -V_4 + \frac{W}{3} \end{aligned} \tag{20}$$

where V_1, V_2, V_3, V_4 are i.i.d. U(-4,4) random variables, and W is a standard normal random variable.

Model (17) is model (24) in Velilla (1998), and model (19) is model (6.3) in Li (1991). The central subspace for both (17) and (19) is two dimensional. The joint regressor distributions (18) and (20) satisfy the linearity condition by construction (Velilla 1998, p. 1092-93), even though $\mathbf{X}$ does not have an elliptically contoured distribution in either case.



a. Model (17) with **X** following (18) b. Model (19) with **X** following (20)

Figure 1: Scatterplots of the Simulated Data

The numerical entries of the rows of Tables 2, 3 and 4 corresponding to the test statistics indexed by 0 and 1 are empirical estimates of the power of the corresponding test. They represent the proportion of times the corresponding null hypothesis $d = 0$ and $d = 1$ is rejected, when the nominal significance level is 0.05 and 0.01 indicated parenthetically. The entries for the test statistics indexed by 2 are empirical estimates of the size of the test. The row entries of the tables throughout this section are to be interpreted in the same or analogous way. Computations were carried out in *Arc* (Cook and Weisberg 1999). The p-values for the weighted chi-square test were calculated with an algorithm by Field (1993).

Figure 1 contains the scatterplots of 100 data points generated according to models (17) and (19). The scatterplots reveal the presence of significant heteroscedasticity in the conditionals of the regressor distribution and the conditional distribution of $\mathbf{X}|Y$. Thus, the main conditions of both Corollary 1 and Theorem 2 appear to be violated. In consequence, the asymptotic distribution of the test statistic $\hat{\Lambda}_d$ is not expected to be chi-square but rather weighted chi-square.

Table 2: Empirical Power and Size for the chi-square and the weighted chi-square tests applied to (17)

X distributed as in (18)						
$n = 100$						
	chi-square			weighted chi-square		
H	5	10	15	5	10	15
$\hat{\Lambda}_0$	1 (1)	1 (1)	1 (1)	1 (.998)	1 (.996)	1 (.967)
$\hat{\Lambda}_1$	.052 (.014)	.07 (.03)	.134 (.05)	.159 (.034)	.225 (.064)	.367 (.12)
$\hat{\Lambda}_2$	.004 (0)	.005 (.001)	.007 (0)	.018 (.005)	.041 (.005)	.078 (.01)
$n = 200$						
	chi-square			weighted chi-square		
H	10	15	20	10	15	20
$\hat{\Lambda}_0$	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1(1)
$\hat{\Lambda}_1$	.114 (.052)	.171 (.068)	.167 (.081)	.258 (.103)	.357 (.152)	.410 (.164)
$\hat{\Lambda}_2$	.038 (.004)	.01 (0)	.003 (0)	.038 (.005)	.077 (.013)	.095 (.015)
$n = 300$						
	chi-square			weighted chi-square		
H	15	20	30	15	20	30
$\hat{\Lambda}_0$	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1(1)
$\hat{\Lambda}_1$	.176 (.091)	.202 (.1)	.199 (.102)	.41 (.172)	.479 (.205)	.549 (.257)
$\hat{\Lambda}_2$	.008 (0)	.007 (.001)	.005 (0)	.061 (.007)	.088 (.011)	.15 (.033)

Tables 2 and 3 indicate that the chi-square test has consistently smaller power than the weighted chi-square test across sample sizes and choice of number of slices. From Table 3, it seems that the power of the chi-square test can decrease as the number of slices increases, whereas the power of the weighted chi-square test increases.

The level of the chi-square test is too small in all cases of Tables 2 and 3, suggesting that it may be conservative when the predictors are not normal. This may partially account for the relatively low power of the chi-square test. The level of the weighted chi-square test increases with the number of slices. This observation supports the idea that the weighted chi-square test is expected to be adversely affected by a small number of observations per slice, due to the variation in the estimated weights $\hat{w}$ in (14).

In addition to the results reported here, a number of other distributions and

Table 3: Empirical Power and Size for the chi-square and the weighted chi-square tests applied to (19)

$\mathbf{X}$ distributed as in (20)						
$n = 100$						
	chi-square			weighted chi-square		
H	5	10	15	5	10	15
$\hat{\Lambda}_0$	1 (1)	1 (.999)	.998 (.974)	1 (1)	1 (1)	.999 (.996)
$\hat{\Lambda}_1$	.182 (.054)	.145 (.033)	.113 (.032)	.255 (.079)	.29 (.085)	.384 (.143)
$\hat{\Lambda}_2$	.006 (0)	.009 (.001)	.007 (0)	.017 (.001)	.035 (.005)	.09 (.013)
$n = 200$						
	chi-square			weighted chi-square		
H	10	15	20	10	15	20
$\hat{\Lambda}_0$	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1(1)
$\hat{\Lambda}_1$	.301 (.099)	.238 (.077)	.196 (.055)	.457 (.183)	.472 (.202)	.519 (.24)
$\hat{\Lambda}_2$	.008 (0)	.019 (.005)	.014 (.001)	.021 (.006)	.064 (.016)	.099 (.022)
$n = 300$						
	chi-square			weighted chi-square		
H	15	20	30	15	20	30
$\hat{\Lambda}_0$	1 (1)	1 (1)	1 (1)	.999 (.999)	1 (1)	1(1)
$\hat{\Lambda}_1$	.379 (.163)	.353 (.133)	.261 (.085)	.604 (.342)	.619 (.356)	.704 (.413)
$\hat{\Lambda}_2$	.013 (.001)	.016 (.002)	.008 (0)	.052 (.01)	.096 (.025)	.173 (.049)

models were investigated. One conclusion from these studies is that the chi-square test is robust when the predictor distribution deviates from normality *provided that there is no significant non-constant variance in the* $\mathbf{X}$*-conditionals* and that the linearity condition holds. The weighted chi-square test also performed very well even in cases where the chi-square test has a clear advantage, namely when the regressor is a normal vector, as exhibited in Table 4 where Y is generated according to model (19), $n = 100$ and the five predictors are independent standard normal random variables. The results of Table 4 suggest that the chi-square test is conservative even when the predictors are normal. In addition, the results of Tables 2, 3 and 4 and other results not reported here suggest that the number of slices for the weighted chi-square test should not be more that 5 to 7 percent of the sample size to avoid test levels being much larger than the nominal level.

Table 4: Empirical Power and Size for the chi-square and the weighted chi-square tests applied to (19)

$\mathbf{X} \sim N_5(0, \mathbf{I}_5)$						
$n = 100$						
	chi-square			weighted chi-square		
H	5	10	15	5	10	15
$\hat{\Lambda}_0$	.996 (.975)	.988 (.933)	.976 (.881)	.997 (.977)	.993 (.952)	.992 (.944)
$\hat{\Lambda}_1$	.435 (.203)	.414 (.174)	.295 (.091)	.52 (.259)	.585 (.293)	.631 (.285)
$\hat{\Lambda}_2$	.016 (.001)	.013 (.003)	.016 (.002)	.032 (.003)	.056 (.009)	.123 (.025)

6 Discussion

SIR is a simple and useful first method for dimension reduction in regression analysis. As such, it seems important to have a well-grounded inference tool for studying the dimension of the inverse regression subspace.

Li (1991) proposed the chi-square test procedure for situations when $\mathbf{X}$ is normal. In the present work, we show that normality of the predictors is not necessary for the asymptotic result to hold, but rather that restrictions should be placed on the conditional covariance structure of the predictors instead. These restrictions are trivially satisfied if $\mathbf{X}$ has a multivariate normal distribution, but they also contradict Li's claim that the asymptotic distribution of $\hat{\Lambda}_d$ does not depend on the constant variance assumption of the conditional distribution of $\mathbf{X}$ given Y.

Theorem 1 is the basis for the weighted chi-square test which applies regardless of the predictor distribution. SIR would now seem to be a more widely applicable dimension reduction technique, requiring only the linearity condition.

The general applicability of the weighted chi-square test does not seem to be overwhelmed by the variability induced by estimating the weights required for its computation, provided the number of slices does not exceed 5 to 7 percent of the sample size. As shown in the simulation study, the weighted chi-square test performs noticeably better than the chi-square test when the regressors are far from normal.

In addition, it also compares well to the chi-square test in cases where the regressors are normal.

The testing procedure for dimension can be viewed as a sequence of at most p, the number of predictors, tests. How this multiple testing affects the overall size of the testing procedure has not been investigated. Approaches similar to those in multiple testing might be appropriate. Additionally, it should be mentioned that the detection of a direction by a testing procedure does not guarantee that it will be close enough to the central subspace to be useful. Directions detected by the weighted chi-square test with a marginal p-value may be farther from the central subspace than directions with a relatively small p-value. Because the chi-square test is conservative, the directions it does detect might be relatively close to the central subspace.

7 Acknowledgements

The authors thank the editor, associate editor, and a referee for their careful reading and comments that have helped improve this article.

References

Bura, E. (1996), *Dimension Reduction via Inverse Regression.* Ph.D. Dissertation, School of Statistics, University of Minnesota.

Cambanis, S., Huang, S., and Simons, G. (1981), "On the theory of elliptically contoured distributions," *Journal of Multivariate Analysis*, 7, 368–385.

Cook, R. D. (1998a), "Principal Hessian Directions Revisited (with discussion)," *Journal of the American Statistical Association*, 91, 983–992.

——— (1998b), Regression Graphics. New York: Wiley.

——— (1996), "Graphics for regressions with a binary response," *Journal of the American Statistical Association*, 91, 983–992.

——— (1994a), "On the interpretation of regression plots," *Journal of the American Statistical Association*, 89, 177–189.

——— (1994b), "Using dimension-reduction subspaces to identify important inputs in models of physical systems," in *1994 Proceedings of the Section on Physical and Engineering Sciences*, Alexandria, VA: American Statistical Association, pp. 18–25.

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

——— (1994), *An Introduction to Regression Graphics.* New York: Wiley.

——— (1983), "Diagnostics for heteroscedasticity in regression," *Biometrika*, 70, 1–10.

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

Eaton, M. L. and Tyler, D. E. (1994), "The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis," *Journal of Multivariate Analysis*, 34, 439–446.

Ferré, L. (1998), "Determining the Dimension in Sliced Inverse Regression and Related Methods," *Journal of the American Statistical Association*, 93, 132–140.

Field, C. (1993), "Tail areas of linear combinations of chi-squares and non-central chi-squares," *Journal of Statistical Computation and Simulation*, 45, 243–248.

Diaconis, P. and Freedman, D. (1984), "Asymptotics of graphical projection pursuit," *The Annals of Statistics*, 12, 793–815.

Hall, P. and Li, K. C. (1993), "On almost linearity of low dimensional projections from high dimensional data," *The Annals of Statistics*, 21, 867–889.

Hsing, T. and Carroll, R. J. (1992), "An asymptotic theory for sliced inverse regression," *The Annals of Statistics*, 20, 1040–1061.

Johnson, M. E., Kotz, S. and Balakrishnan, N. (1994), *Continuous Univariate Distributions. Vol. 1, 2.* Wiley: New York.

Kelker, D. (1970), "Distribution theory of spherical distributions and a location-scale parameter generalization," *Sankhya, Series A, Indian Journal of Statistics*, 32, 412–430.

Li, K. C. (1991), "Sliced inverse regression for dimension reduction," *J. Am. Statist. Assoc.*, 86, 316–342.

——— (1992), "On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma," *J. Am. Statist. Assoc.*, 87, 1025–1039.

Zhu, L. X. and Ng, K. W. (1995), "Asymptotics of sliced inverse regression," *Statistica Sinica*, 5, 727–736.