

Marginal tests with sliced average variance estimation

BY YONGWU SHAO, R. DENNIS COOK AND SANFORD WEISBERG

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
 ywshao@stat.umn.edu dennis@stat.umn.edu sandy@stat.umn.edu

SUMMARY

We present a new computationally feasible test for the dimension of the central subspace in a regression problem based on sliced average variance estimation. We also provide a marginal coordinate test. Under the null hypothesis, both the test of dimension and the marginal coordinate test involve test statistics that asymptotically have chi-squared distributions given normally distributed predictors, and have a distribution that is a linear combination of chi-squared distributions in general.

Some key words: Marginal coordinate test; Sufficient dimension reduction.

1. INTRODUCTION

Consider a univariate response Y and a vector of continuous predictors $X \in \mathbb{R}^p$. Sufficient dimension reduction seeks to find a subspace given by the column space of a $p \times d$ matrix η with $d \leq p$ such that

$$Y \perp\!\!\!\perp X|\eta^T X, \quad (1)$$

where $\perp\!\!\!\perp$ indicates independence. Under mild conditions the intersection of all dimension reduction subspaces is itself a dimension reduction subspace and then is called the central subspace (Cook, 1996), written as $\mathcal{S}_{Y|X}$. Let $d = \dim(\mathcal{S}_{Y|X})$.

Two early methods that can be used to estimate the central subspace are sliced inverse regression (Li, 1991) and sliced average variance estimation (Cook & Weisberg, 1991). Both of these methods look at the inverse regression problem of $X|Y$ to learn about the central subspace. Sliced inverse regression uses only first moments $E(X|Y)$, while sliced average variance estimation uses first and second moments and, consequently, is more comprehensive than sliced inverse regression (Cook & Lee, 1999). Sliced average variance estimation has been used by Cook & Critchley (2000) to identify outliers and mixtures graphically, by Bura & Pfeiffer (2003) for class prediction with DNA microarray data and by Zhu & Hastie (2003) in feature extraction for nonparametric discriminant analysis.

Hypothesis tests of the form $d = m$ versus $d > m$ are an important component of practical dimension reduction analyses. For sliced inverse regression, Li (1991) provided an easily computed test based on the sum of the smallest eigenvalues of a $p \times p$ matrix. He showed that, when X is normally distributed, the test statistic asymptotically has a χ^2 distribution under the null hypothesis. Bura & Cook (2001) established the asymptotic distribution for general regressors.

For sliced average variance estimation, Cook & Ni (2005) suggested a test for dimension derived using a method similar to that used by Li for sliced inverse regression. If the response is discrete with s distinct values, then the asymptotic null distribution of the test

statistic given by Cook & Ni (2005) is a weighted combination of p^2s independent χ^2 random variables, with weights given by the eigenvalues of a symmetric matrix of order $p^2s \times p^2s$. For example, if $p = 10$ and $s = 10$ then the matrix is of order 1000×1000 , and so huge samples are needed for this test to be useful.

In this article, we provide a new test for the hypothesis $d = m$ based on sliced average variance estimation. Under fairly general conditions, the test statistic converges in distribution to a weighted χ^2 distribution, where the weights can be consistently estimated as the eigenvalues of a symmetric matrix of order $(p - m)(p - m + 1)/2$. If $p = 10$ and $m = 2$ this matrix is 36×36 . If the predictors are normally distributed, the asymptotic null distribution of the test statistic further reduces to a central χ^2 , eliminating the need to estimate the weights.

We assume throughout this article that the scalar response Y and the $p \times 1$ vector of predictors X have a joint distribution, and that the data (y_i, x_i) , $i = 1, \dots, n$, are independent and identically distributed observations on (Y, X) . Subspaces will be denoted by \mathcal{S} , and $\mathcal{S}(\zeta)$ means the subspace spanned by the columns of the $p \times d$ matrix ζ , with $\mathcal{S}(\zeta)^\perp$ as its orthogonal complement. We use P_ζ to denote the projection operator for $\mathcal{S}(\zeta)$ with respect to the identity inner product. If C is a random matrix, we use $\text{var}(C)$ to denote the covariance matrix of $\text{vec}(C)$, where $\text{vec}(\cdot)$ denotes the operator that constructs a vector from a matrix by stacking its columns.

2. SLICED AVERAGE VARIANCE ESTIMATION

Let η be a basis of $\mathcal{S}_{Y|X} = \mathcal{S}(\eta)$ and let $\Sigma = \text{var}(X) > 0$. A standardized predictor with zero mean and identity covariance matrix is $Z = \Sigma^{-1/2}\{X - E(X)\}$. Cook (1998, §6.3) showed that the columns of the matrix $\gamma = \Sigma^{1/2}\eta$ form a basis for $\mathcal{S}_{Y|Z}$, the central subspace for the regression of Y on Z . Thus there is no loss of generality in working on the Z scale. In the sample we replace the x_i by $z_i = \hat{\Sigma}^{-1/2}(x_i - \bar{x})$, where \bar{x} is the average of the x_i and $\hat{\Sigma} = n^{-1}\sum_{i=1}^n(x_i - \bar{x})(x_i - \bar{x})^\top$. The following two standard assumptions are sufficient for sliced average variance estimation to produce consistent estimators of vectors in the central subspace.

Condition 1. The conditional expectation $E(X|\eta^\top X)$ is a linear function of $\eta^\top X$.

Condition 2. The matrix $\text{var}(X|\eta^\top X)$ is nonrandom, i.e. constant.

Condition 1 holds for all subspaces of \mathbb{R}^p if the predictor X has an elliptical distribution (Eaton, 1986), although this scenario is not necessary. Condition 2 will be satisfied if X has a multivariate normal distribution and holds approximately if X is elliptically contoured. The two assumptions apply to the marginal distribution of X and not to the conditional distribution of $Y|X$.

Sliced average variance estimation is based on the conditional variances $\text{var}(Z|Y)$. Assume that Y is discrete, taking values in $\{1, \dots, s\}$, where $s \geq 2$. If Y is not discrete we can always slice it into s nonoverlapping slices to obtain a discrete version. Given Conditions 1 and 2, Cook & Weisberg (1991) showed that the columns of the matrix

$$M = E\{I_p - \text{var}(Z|Y)\}^2 = \sum_{k=1}^s f_k(I_p - \Sigma_k)^2 \quad (2)$$

are contained in the central subspace $\mathcal{S}_{Y|Z}$, where $f_k = \text{pr}(Y = k)$ and $\Sigma_k = \text{var}(Z|Y = k)$. We impose a third condition, called the coverage condition.

Condition 3. For any nonzero $\beta \in \mathcal{S}_{Y|X}$, at least one of $\text{var}\{E(\beta^\top X|Y)\} > 0$ or $\text{var}\{\text{var}(\beta^\top X|Y)\} > 0$ holds.

Condition 3 is weaker than the coverage condition for sliced inverse regression, given by $\text{var}\{E(\beta^\top X|Y)\} > 0$ for any $\beta \in \mathcal{S}_{Y|X}$, $\beta \neq 0$. Sliced average variance estimation is more comprehensive than sliced inverse regression, resulting in a weaker coverage condition.

Under Conditions 1–3, we have $\mathcal{S}(M) = \mathcal{S}_{Y|Z}$. The reason for this is as follows. First $\mathcal{S}(M) \subseteq \mathcal{S}_{Y|Z}$ by Conditions 1 and 2. If we assume that $\mathcal{S}(M) \neq \mathcal{S}_{Y|Z}$, then we can find a $\beta \neq 0$, $\beta \in \mathcal{S}_{Y|Z}$, such that $\beta \in \mathcal{S}(M)^\perp$. For this β we then have $(\Sigma_k - I_p)\beta = 0$, and hence $\beta^\top(\Sigma_k - I_p)\beta = 0$, $\text{var}(\beta^\top Z|Y) = \beta^\top \beta$ and $\text{var}\{E(\beta^\top Z|Y)\} = \text{var}(\beta^\top Z) - E\{\text{var}(\beta^\top Z|Y)\} = 0$. However, since $\beta \in \mathcal{S}_{Y|Z}$, by Condition 3, either $\text{var}\{E(\beta^\top Z|Y)\} > 0$ or $\text{var}\{\text{var}(\beta^\top Z|Y)\} > 0$, and we reach a contradiction. Therefore $\beta \notin \mathcal{S}_{Y|Z}$, and $\mathcal{S}(M) = \mathcal{S}_{Y|Z}$.

Given a random sample (y_i, x_i) , we can estimate the matrix M consistently by

$$\hat{M} = \sum_{k=1}^s \hat{f}_k (I - \hat{\Sigma}_k)^2, \quad (3)$$

where $\hat{\Sigma}_k$ is the sample version of Σ_k based on the n_k observations in slice k , and $\hat{f}_k = n_k/n$ is the sample fraction of observations in slice k . Let $\hat{\beta}_j$ denote the eigenvector corresponding to the j th-largest eigenvalue of \hat{M} . Then the span of $\hat{\Sigma}^{-1/2}(\hat{\beta}_1, \dots, \hat{\beta}_d)$ is a consistent estimator of $\mathcal{S}_{Y|X}$.

Since d is generally unknown, a method is required for estimating it. The typical procedure is based on tests of the marginal dimension hypothesis $d = m$ versus $d > m$. Starting with $m = 0$, test $d = m$. If the hypothesis is rejected, increment m by one and test again, stopping with the first nonsignificant result. The corresponding value of m is the estimate \hat{d} of d .

Marginal dimension tests allow us to reduce dimension by replacing the predictors X by \hat{d} linear combinations $\hat{\beta}_j^\top \hat{\Sigma}^{-1} X$, $j = 1, \dots, \hat{d}$. Common model-based methods of dimension reduction, such as subset selection in linear regression (Weisberg, 2005) or the lasso (Tibshirani, 1996), seek dimension reduction by removing predictors, thereby decreasing p to some smaller number. This type of procedure is also possible with sliced average variance estimation if we use a marginal coordinate hypothesis (Cook, 2004) that requires $\mathcal{S}_{Y|X} \subseteq \mathcal{V}$, where \mathcal{V} is a specified subspace of \mathbb{R}^p . For example, suppose we have three predictors (X_1, X_2, X_3) and we contemplate removing X_3 from our regression by testing the hypothesis $Y \perp\!\!\!\perp X_3 | (X_1, X_2)$. In this case $\mathcal{V} = \mathcal{S}((1, 0, 0)^\top, (0, 1, 0)^\top)$ and our goal is to test $\mathcal{S}_{Y|X} \subseteq \mathcal{V}$. By combining marginal dimension tests and marginal coordinate tests we can achieve dimension reduction both by removing predictors and by replacing the remaining predictors by linear combinations of them, without specifying a parsimonious parametric model.

3. TESTS AND DISTRIBUTIONS

3.1. Marginal coordinate hypotheses

Theory for marginal coordinate tests turns out to be easier than that for the marginal dimensional tests. Define the population quantity $A_k = f_k^{1/2}(\Sigma_k - I_p)$ and its estimator $\hat{A}_k = \hat{f}_k^{1/2}(\hat{\Sigma}_k - I_p)$, and rewrite (2) and (3) as $M = \sum_{k=1}^s A_k^2$ and $\hat{M} = \sum_{k=1}^s \hat{A}_k^2$. The marginal coordinate hypothesis is $\mathcal{S}_{Y|X} \subseteq \mathcal{V}$ versus $\mathcal{S}_{Y|X} \not\subseteq \mathcal{V}$, where $\dim(\mathcal{V}) = m < p$. Let α_x be a user-selected basis for \mathcal{V} expressed as a $p \times m$ matrix of full column rank m . In

Z-scale, population and sample orthonormal bases are given by $\alpha = \Sigma^{1/2} \alpha_x (\alpha_x^T \Sigma \alpha_x)^{-1/2}$ and $\hat{\alpha} = \hat{\Sigma}^{1/2} \alpha_x (\alpha_x^T \hat{\Sigma} \alpha_x)^{-1/2}$, respectively. Finally, let H and \hat{H} be orthonormal bases for the orthogonal complements of $\mathcal{S}(\alpha)$ and $\mathcal{S}(\hat{\alpha})$ respectively, expressed as $p \times (p - m)$ matrices.

Based on sliced inverse regression, Cook (2004, §5) provided a test statistic for the marginal coordinate hypothesis given by $n \text{tr}(\hat{H}^T \hat{M}_{\text{SIR}} \hat{H})$, where \hat{M}_{SIR} is the sliced inverse regression sample matrix. The analogous statistic based on sliced average variance estimation obtained by replacing \hat{M}_{SIR} with \hat{M} seems theoretically intractable, so we use a slightly different test statistic:

$$T_n(\hat{H}) = \frac{n}{2} \sum_{k=1}^s \text{tr}\{(\hat{H}^T \hat{A}_k \hat{H})^2\}. \quad (4)$$

It can be shown that, under Conditions 1–3, $\sum_{k=1}^s \text{tr}\{(H^T A_k H)^2\} = 0$ if and only if $\mathcal{S}(H)$ is contained in the orthogonal complement of $\mathcal{S}_{Y|Z}$, which provides justification for the use of $T_n(\hat{H})$. In particular, if $T_n(\hat{H})$ is large, then H is very likely to have a component in common with $\mathcal{S}_{Y|Z}$. The simpler statistic $n \sum_{k=1}^s \text{tr}(\hat{H}^T \hat{A}_k \hat{H})$ does not have this property, and even if $\sum_{k=1}^s \text{tr}(H^T A_k H)$ is exactly zero we cannot conclude that $\mathcal{S}(H)$ is orthogonal to $\mathcal{S}_{Y|Z}$.

Define $J_k = 1_{Y=k}$, where 1_Ω indicates whether or not an event Ω is true. Define $G = ((J_1 - f_1) f_1^{-1/2}, \dots, (J_s - f_s) f_s^{-1/2})^T$, and let $V = H^T Z$. For the test statistic $T_n(\hat{H})$ given by (4), we have the following theorem, which is proved in the Appendix.

THEOREM 1. *Assume that Conditions 1–3 hold. Then, under the null hypothesis that $\mathcal{S}_{Y|X} \subseteq \mathcal{V}$,*

$$2T_n(\hat{H}) \longrightarrow \sum_j \delta_j \chi_j^2(1),$$

in distribution as $n \rightarrow \infty$, where the δ_j , $j = 1, \dots, s(p - m)(p - m + 1)/2$, are the largest $s(p - m)(p - m + 1)/2$ eigenvalues of the $s(p - m)^2 \times s(p - m)^2$ matrix $\text{var}\{G \otimes (VV^T - I_{p-m})\}$, and $\chi_j^2(1)$ are independent χ^2 random variables with one degree of freedom.

The variance matrix $\text{var}\{G \otimes (VV^T - I_{p-m})\}$ can be consistently estimated by the sample covariance matrix of $h_i \otimes (\hat{H}^T z_i z_i^T \hat{H} - I_{p-m})$, where $h_i = ((J_{1,i} - \hat{f}_1) \hat{f}_1^{-1/2}, \dots, (J_{s,i} - \hat{f}_s) \hat{f}_s^{-1/2})^T$ is a sample version of G , and $J_{k,i} = 1_{y_i=k}$ is a sample version of J_k . The matrix $\text{var}\{G \otimes (VV^T - I_{p-m})\}$ has a size of $s(p - m)^2 \times s(p - m)^2$. If H is a vector, as it may frequently be in practice, then we need to estimate an $s \times s$ covariance matrix, which is usually feasible because in applications s is likely to be at most 10. However, in some applications the dimension of H might be so large that the matrix $\text{var}\{G \otimes (VV^T - I_{p-m})\}$ cannot be estimated accurately, and it is therefore desirable to have an alternative. It turns out that the matrix $\text{var}\{G \otimes (VV^T - I_{p-m})\}$ has a simpler structure under the condition that $\text{var}(\theta^T Z \otimes \theta^T Z | \gamma^T Z)$ is nonrandom; here θ is an orthonormal basis of $\mathcal{S}_{Y|Z}^\perp$.

COROLLARY 1. *Assume that Conditions 1–3 hold, and that $\text{var}(\theta^T Z \otimes \theta^T Z | \gamma^T Z)$ is nonrandom. Then, under the null hypothesis that $\mathcal{S}_{Y|X} \subseteq \mathcal{V}$,*

$$2T_n(\hat{H}) \longrightarrow \sum_j \delta_j \chi_j^2(s - 1), \quad (5)$$

in distribution as $n \rightarrow \infty$, where the δ_j , $j = 1, \dots, (p-m)(p-m+1)/2$, are the largest $(p-m)(p-m+1)/2$ eigenvalues of the $(p-m)^2 \times (p-m)^2$ matrix $\text{var}(H^T Z \otimes H^T Z)$, and $\chi_j^2(s-1)$ are independent χ^2 random variables with $s-1$ degrees of freedom.

The constant variance condition, Condition 2, is equivalent to the condition that $E(\theta^T Z \otimes \theta^T Z | \gamma^T Z)$ is nonrandom. Corollary 1 requires that $\text{var}(\theta^T Z \otimes \theta^T Z | \gamma^T Z)$ also be nonrandom. Both conditions are satisfied when X has a multivariate normal distribution, or when $\theta^T Z$ and $\gamma^T Z$ are independent.

The test requires one to estimate the eigenvalues in (5) by the eigenvalues of the sample covariance matrix of the $(\hat{H}^T z_i \otimes \hat{H}^T z_i)$, and then to obtain a p -value from the distribution of a linear combination of χ^2 random variables. The p -value can be approximated by using the modified statistic $2T_n(\hat{H}) (\sum \delta_j / \sum \delta_j^2)$, which is distributed approximately as a χ^2 random variable with $(s-1) (\sum \delta_j)^2 / \sum \delta_j^2$ degrees of freedom (Bentler & Xie, 2000). We use this approximation in §4.

When X is normal, we have $H^T Z \sim N(0, I_{p-m})$, and $\text{var}(H^T Z \otimes H^T Z)/2$ is a projection matrix with rank $(p-m)(p-m+1)/2$ (Schott, 1997, Theorem 7.10), so all its nonzero eigenvalues are equal to 1. The following theorem then holds.

THEOREM 2. *Assume that Condition 3 holds and that X is normally distributed. Then, under the null hypothesis $\mathcal{S}_{Y|X} \subseteq \mathcal{V}$,*

$$T_n(\hat{H}) \rightarrow \chi^2\{(s-1)(p-m)(p-m+1)/2\}, \quad (6)$$

in distribution as $n \rightarrow \infty$.

A common application of the marginal coordinate hypothesis is to examine the possibility of excluding one of the predictors. In this case H is a p -dimensional vector, $H^T Z$ is a random scalar, and $T_n(\hat{H}) \rightarrow \chi^2(s-1)$, in distribution, assuming normality, and $2T_n(\hat{H})/\delta_1 \rightarrow \chi^2(s-1)$, in distribution, in general. Here δ_1 is the variance of $(H^T Z)^2$, which can be consistently estimated by the sample variance of $(\hat{H}^T z_i)^2$. Either version of this test can be used as the basis of a backward elimination procedure, similarly to the use of t tests in linear regression.

3.2. Marginal dimension hypotheses

The test for a marginal dimension hypothesis is essentially the same as that for a marginal coordinate hypothesis, but with the added complication that the space \mathcal{V} is now data-determined rather than fixed. Suppose we want to test $d = m$. Let $\hat{\theta}$ be a $p \times (p-m)$ matrix with columns given by the eigenvectors of \hat{M} corresponding to its smallest $p-m$ eigenvalues. Since $\hat{M} = M + O_p(n^{-1/2})$, by Tyler (1981, Lemma 2.1), $P_{\hat{\theta}} = P_{\theta} + O_p(n^{-1/2})$. Define the following test statistic:

$$T_n(\hat{\theta}) = \frac{n}{2} \sum_{k=1}^s \text{tr}\{(\hat{\theta}^T \hat{A}_k \hat{\theta})^2\}. \quad (7)$$

By Lemma A1 in the Appendix, $T_n(\hat{\theta})$ has the same distribution as $T_n(\theta)$, which can be obtained by Corollary 1 in the last section and Lemma A1. Therefore we have the following theorem.

THEOREM 3. Assume that Conditions 1–3 hold, and assume that $\text{var}(\theta^T Z \otimes \theta^T Z | \gamma^T Z)$ is nonrandom. Then, under the null hypothesis $d = m$,

$$2T_n(\hat{\theta}) \longrightarrow \sum_j \delta_j \chi_j^2(s-1), \quad (8)$$

in distribution as $n \rightarrow \infty$, where the $\delta_j, j = 1, \dots, (p-m)(p-m+1)/2$, are the largest $(p-m)(p-m+1)/2$ eigenvalues of $\text{var}(\theta^T Z \otimes \theta^T Z)$.

If in addition X is normally distributed then, under the null hypothesis $d = m$,

$$T_n(\hat{\theta}) \longrightarrow \chi^2\{(s-1)(p-m)(p-m+1)/2\}, \quad (9)$$

in distribution.

For the special case of testing $Y \perp\!\!\!\perp X$, or $m = 0$, we have $\hat{\theta} = I_p$ and $T_n(I_p) = \text{ntr}(\hat{M})/2 \rightarrow \sum_j \delta_j \chi_j^2(s-1)$, in distribution, where the weights δ_j are the eigenvalues of $\text{var}(\theta^T Z \otimes \theta^T Z) = \text{var}(Z \otimes Z)$. If X is normal, then $T_n(I_p) \rightarrow \chi^2\{(s-1)p(p+1)/2\}$, in distribution.

There is a version of Theorem 3 that corresponds to Theorem 1, but here for brevity we state only the version that corresponds to Corollary 1, because in practice $p - d$ is usually quite large and the result given here will be more useful in practice.

4. SIMULATIONS

A simulation study was conducted to provide support for the asymptotic results. Each simulation run was based on 1000 replications, using the following two models that can illuminate several issues:

$$Y = X_1 + \epsilon \quad (10)$$

$$Y = X_1^2 + X_2 + \epsilon, \quad (11)$$

where $\epsilon \sim N(0, 0.1^2)$. Every simulation run used $s = 5$ slices. For (10), both sliced inverse regression and sliced average variance estimation should be able to find the correct structure. For (11), sliced inverse regression will be unable to detect the central subspace because its coverage condition fails, $E(X_1|Y) = 0$. We used $p = 4$ or 10 , with X having either independent standard normal or independent heavy-tailed t_5 components, and we carried out tests using (4) or (7) as well as the standard test based on sliced inverse regression. For the sliced average variance estimation tests, significance levels were computed in two ways, using (5) or (8), and using (6) or (9), with normality assumed for the predictors.

Tables 1(a) and 2(a) contain results from marginal coordinate tests for models (10) and (11), respectively. In both models we tested the null hypothesis that $\mathcal{S}_{Y|X} \subseteq \mathcal{V}$ versus the alternative hypothesis that $\mathcal{S}_{Y|X} \not\subseteq \mathcal{V}$, where $\mathcal{V} = \mathcal{S}(e_1, e_2)$ and e_i is the canonical basis vector with a one in the i th place and zero elsewhere. The null hypothesis is true in both models. From the results we can see that the actual level may be adequate for many applications with $n \geq 200$, and in some cases for $n \geq 100$. The number of predictors p did not affect results much. We did not see a clear difference between the results obtained by the three tests, except in the case $X_i \sim t_5$. In this case, the actual level of the sliced average variance estimation test that requires normality is far from the nominal level, as might be expected. The sliced inverse regression test works well here because the linearity condition and the constant variance condition hold (Cook, 1998, §11.3.3).

Table 1. *Estimated levels, as percentages, of nominal 5% tests, based on model (10)*

n	$p = 4, X_i \sim N(0, 1)$			$p = 10, X_i \sim N(0, 1)$		
	S_N	S_G	SIR	S_N	S_G	SIR
(a) <i>Marginal coordinate tests</i>						
100	9.1	8.5	6.1	17.4	10.2	7.4
200	6.6	6.0	5.2	10.6	7.7	6.3
400	5.1	5.1	4.5	8.9	7.1	4.9
800	6.2	5.6	5.5	6.3	5.3	5.4
(b) <i>Marginal dimensional tests</i>						
100	7.6	5.8	6.7	1.7	0.9	5.1
200	6.9	6.6	4.2	10.4	6.5	4.2
400	6.2	5.9	5.5	7.5	5.3	5.4
800	5.9	5.3	4.8	5.7	5.0	5.7

S_N , sliced average variance estimation test assuming normality;

S_G , sliced average variance estimation test without assuming normality;

SIR, test based on sliced inverse regression.

Table 2. *Estimated levels, as percentages, of nominal 5% tests, based on model (11)*

n	$p = 4, X_i \sim N(0, 1)$			$p = 4, X_i \sim t_5$		
	S_N	S_G	SIR	S_N	S_G	SIR
(a) <i>Marginal coordinate tests</i>						
100	6.2	5.5	5.6	72.7	3.1	4.8
200	5.3	5.5	5.3	77.0	3.1	5.7
400	5.0	5.3	4.7	79.1	3.0	4.2
800	5.7	5.2	5.2	85.0	3.2	4.8
(b) <i>Marginal dimensional tests</i>						
100	3.2	3.0	0.3	21.8	0.6	0.0
200	4.8	4.8	0.3	54.8	2.0	0.1
400	4.6	4.2	0.2	77.6	4.3	0.1
800	5.5	5.0	0.2	83.9	3.5	0.2

S_N , sliced average variance estimation test assuming normality;

S_G , sliced average variance estimation test without assuming normality;

SIR, test based on sliced inverse regression.

Tables 1(b) and 2(b) contain results of the marginal dimension test. For model (10), we tested $d = 1$ vs. $d > 1$, while for (11) we tested $d = 2$ vs. $d > 2$. In both tests the null hypothesis is true. As should be expected, the results in Table 1(b) are more favourable to sliced inverse regression. For (11) in Table 2(b), the sliced inverse regression test behaved poorly because its coverage condition is not satisfied. The test based on sliced average variance estimation does not share this problem.

Unlike with the marginal coordinate test, the accuracy of the level of the marginal dimension test decreases as p increases, because the latter test requires estimation of a subspace. If p is large, the estimate of the central subspace degrades, and the performance of the test statistic will suffer. As before, the test based on sliced average variance estimation

Table 3. *Estimated powers, as percentages, at nominal 5% level*

n	$p = 4, X_i \sim N(0, 1)$			$p = 4, X_i \sim t_5$		
	S_N	S_G	SIR	S_N	S_G	SIR
(a) <i>Marginal coordinate tests based on model (10)</i>						
40	15.4	13.1	100.0	21.9	8.1	100.0
60	36.5	35.0	100.0	51.6	19.1	100.0
80	73.6	69.4	100.0	88.5	34.8	100.0
100	97.4	95.5	100.0	98.4	54.4	100.0
(b) <i>Marginal dimension tests based on model (11)</i>						
100	40.8	42.1	5.2	65.0	4.1	5.1
200	94.4	94.3	5.7	96.3	15.3	4.6
400	100.0	100.0	5.0	100.0	59.7	4.8
1000	100.0	100.0	5.0	100.0	96.6	5.6

S_N , sliced average variance estimation test assuming normality;

S_G , sliced average variance estimation test without assuming normality;

SIR, test based on sliced inverse regression.

that assumes normality does not have an actual level close to the nominal level when the predictors have t_5 distributions.

Table 3(a) contains estimated powers of the marginal coordinate test for (10), testing the hypothesis $\mathcal{S}_{Y|X} \subseteq \mathcal{V}$ versus the alternative hypothesis $\mathcal{S}_{Y|X} \not\subseteq \mathcal{V}$, where $\mathcal{V} = \mathcal{S}(e_2)$. The power was computed at a nominal 5% level. We expect that both sliced average variance estimation and sliced inverse regression will perform reasonably well in this model because all required conditions are satisfied. However, sliced inverse regression should have the higher power because it requires estimating fewer objects. We used relatively small sample sizes for Table 3(a) to emphasize this difference. As we increased the sample size, the power for all sliced average variance estimation tests in Table 3(a) approached 100%.

We tested the marginal dimension hypothesis that $d = 1$ versus $d > 1$ using model (11). Table 3(b) contains the estimated powers of our tests. Low power, no greater than the level, is to be expected for sliced inverse regression. The sliced average variance estimation method without assuming normality does well when n is large, but it does not perform so well for small n . Without assuming normality we need a relatively large sample size to estimate accurately the weights δ_j in the limiting distribution of $T_n(\hat{\theta})$. The high powers for the normal version of the sliced average variance estimation test in the fifth column are due in part to the high levels in the corresponding column of Table 2(b). The normal theory test is not recommended if the predictors have very heavy tails. Finally, the results in Table 3(b) for sliced inverse regression reflect a primary motivation for the sliced average variance estimation methodology.

5. SWISS BANKNOTE DATA

In the Swiss Banknote data (Flury & Riedwyl, 1988, p.5), the response Y is a binary variable indicating whether a banknote is counterfeit or not. There are six predictors, all measured in millimetres, namely, Length, Left, Right, Top, Bottom and Diagonal.

Table 4. *Significance levels for marginal dimension test results for the Swiss banknote data*

m	$T_n(\hat{\theta})$	df	p -value given normality	p -value general test
0	148.939	21	0.000	0.000
1	58.281	15	0.000	0.001
2	17.399	10	0.066	0.217
3	4.744	6	0.577	0.704

df, degrees of freedom

Table 5. *Marginal coordinate test results for the Swiss banknote data based on sliced average variance estimation*

Predictor	$T_n(V)$	df	p -value given normality	p -value general test
Length	0.834	1	0.498	0.361
Left	10.069	1	0.029	0.002
Right	0.722	1	0.428	0.396
Top	1.383	1	0.264	0.240
Bottom	29.147	1	0.000	0.000
Diagonal	16.244	1	0.004	0.000

df, degrees of freedom

Since the response is binary, sliced inverse regression can identify only one direction. Cook & Lee (1999) analyzed these data using sliced average variance estimation and concluded that the dimension of the central subspace is likely to be at least two. They also pointed out that there seem to be two types of counterfeit note, based on the plot of the data projected on to the two directions.

The test results from using the proposed marginal dimension test statistic are shown in Table 4. We can see that, if normality of the predictors is assumed, at level 5%, the dimension of the central subspace is inferred to be 2, while at level 10% the dimension is inferred to be 3. Without the assumption of normality, the inferred dimension is 2.

Table 5 contains the results from applications of the marginal coordinate test to each predictor. In the case of Length, the null hypothesis is that Y is independent of Length given all other predictors. A 5% backward elimination procedure suggested that Length, Top, Right and Left could be removed from the regression without much loss of information. This was supported by the test that Y is independent of Length, Top, Right and Left given the remaining two predictors; the p -value of this hypothesis was 0.28 for the general test, and 0.06 given normality. As illustrated in this example, we find it useful to compute both the normal and general tests in most applications.

6. DISCUSSION

Cook & Ni (2005) presented methodology for improved estimation of the central mean subspace based on first moments using minimization of an objective function rather than using the eigenvectors of a matrix in the sliced inverse regression method. They point out that their method could in principle be extended to include second moment information that is used by sliced average variance estimation, but prohibitively large samples would be required. Thus, with the inclusion of the tests proposed in this paper, sliced average

variance estimation appears to be an extremely attractive method for learning about the dimension and structure of a regression problem with minimal assumptions when second moment information is relevant.

ACKNOWLEDGEMENT

We thank the referees and the editor for their useful comments. R. D. Cook and Y. Shao were supported by the U.S. National Science Foundation.

APPENDIX

Technical proofs

Before we go to the proof of Theorem 1, we first prove two lemmas.

LEMMA A1. *Suppose that \tilde{H} is a $p \times (p - m)$ matrix such that $\tilde{H}^T \tilde{H} = I_{p-m}$, $P_{\tilde{H}} = P_H + O_p(n^{-1/2})$. Then $T_n(\tilde{H}) = T_n(H) + o_p(1)$.*

Proof. We have

$$\begin{aligned} P_{\tilde{H}} \hat{A}_k P_{\tilde{H}} &= P_H \hat{A}_k P_H + (P_{\tilde{H}} - P_H) \hat{A}_k P_H + P_H \hat{A}_k (P_{\tilde{H}} - P_H) + o_p(n^{-1/2}) \\ &= P_H \hat{A}_k P_H + (P_{\tilde{H}} - P_H) (\hat{A}_k - A_k) P_H \\ &\quad + P_H (\hat{A}_k - A_k) (P_{\tilde{H}} - P_H) + o_p(n^{-1/2}) \\ &= P_H \hat{A}_k P_H + o_p(n^{-1/2}). \end{aligned}$$

Therefore $P_{\tilde{H}} \hat{A}_k P_{\tilde{H}} = P_H \hat{A}_k P_H + o_p(n^{-1/2})$, and $T_n(\tilde{H}) = T_n(H) + o_p(1)$. \square

Assume without loss of generality that $E(X) = 0$ and $\text{var}(X) = I_p$, so that $Z = X$. For a fixed slice k , let $\mu_k = E(X|Y = k)$ and $\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} J_{k,i} x_i$. Let $\Delta_k = E[J_k \{(X - \mu_k)(X - \mu_k)^T - I_p\}]$, $\hat{\Delta}_k = (1/n) \sum_{i=1}^n [J_{k,i} \{(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T - I_p\}]$, $\xi_i = x_i x_i^T - I_p$, $\bar{\xi} = (1/n) \sum_i \xi_i$, $\phi_{k,i} = J_{k,i} \xi_i - E\{J_k(X X^T - I_p)\}$, $\bar{\phi}_k = (1/n) \sum_i \phi_{k,i}$, $\zeta_{k,i} = J_{k,i} x_i - f_k \mu_k$ and $\bar{\zeta}_k = (1/n) \sum_i \zeta_{k,i}$. The averages $\bar{\xi}$, $\bar{\phi}_k$, $\bar{\zeta}_k$ are $O_p(n^{-1/2})$.

LEMMA A2. *It holds that*

$$\hat{\Delta}_k = \Delta_k + \bar{\phi}_k - \mu_k \bar{\zeta}_k^T - \bar{\zeta}_k \mu_k^T + (\hat{f}_k - f_k) \mu_k \mu_k^T + o_p(n^{-1/2})$$

Proof. We have

$$\begin{aligned} \hat{\Delta}_k &= \frac{1}{n} \sum_{i=1}^n [J_{k,i} \{(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T - I_p\}] \\ &= \frac{1}{n} \sum_{i=1}^n [J_{k,i} \{(x_i - \mu_k)(x_i - \mu_k)^T - I_p\}] + \hat{f}_k (\hat{\mu}_k - \mu_k) (\hat{\mu}_k - \mu_k)^T \\ &= \frac{1}{n} \sum_{i=1}^n \{J_{k,i} (x_i x_i^T - I_p)\} + \hat{f}_k \mu_k \mu_k^T \\ &\quad - \frac{1}{n} \left(\sum_{i=1}^n J_{k,i} x_i \right) \mu_k^T - \mu_k \left(\frac{1}{n} \sum_{i=1}^n J_{k,i} x_i^T \right) + o_p(n^{-1/2}) \\ &= \Delta_k + \bar{\phi}_k - \mu_k \bar{\zeta}_k^T - \bar{\zeta}_k \mu_k^T + (\hat{f}_k - f_k) \mu_k \mu_k^T + o_p(n^{-1/2}) \end{aligned}$$

The second equation holds because $\hat{f}_k \hat{\mu}_k = (1/n) \sum_{i=1}^n J_{k,i} x_i$, and hence $\sum_{i=1}^n J_{k,i} (x_i - \hat{\mu}_k) = 0$. The third equation holds because $(\hat{\mu}_k - \mu_k)(\hat{\mu}_k - \mu_k)^\top$ is $o_p(n^{-1/2})$. \square

Proof of Theorem 1. Since $P_{\hat{H}} = P_H + O_p(n^{-1/2})$, by Lemma A1, $T_n(\hat{H}) = T_n(H) + o_p(1)$. Therefore, we only need to derive the distribution of $T_n(H)$. We can write $A_k = f_k^{-1/2} \Delta_k$ and $\hat{A}_k = \hat{f}_k^{-1/2} \hat{\Sigma}^{-1/2} \{\hat{\Delta}_k - \hat{f}_k(\hat{\Sigma} - I_p)\} \hat{\Sigma}^{-1/2}$. According to Cook (1998, §12.3.1) we have $\hat{\Sigma} - I_p = \bar{\xi} + o_p(n^{-1/2})$ and $\hat{\Sigma}^{-1/2} = I_p - (\bar{\xi}/2) + o_p(n^{-1/2})$. Therefore, $\hat{\Sigma}^{-1/2} \bar{\xi} \hat{\Sigma}^{-1/2} = \bar{\xi} + o_p(n^{-1/2})$. Applying Lemma A2, we obtain

$$\hat{\Sigma}^{-1/2} \hat{\Delta}_k \hat{\Sigma}^{-1/2} = \Delta_k + \Delta_k \bar{\xi} + \bar{\xi}^\top \Delta_k + \bar{\phi}_k - \mu_k \bar{\xi}^\top - \bar{\xi} \mu_k^\top + (\hat{f}_k - f_k) \mu_k \mu_k^\top + o_p(n^{-1/2}).$$

According to the null hypothesis, $S_{Y|X} \subseteq \mathcal{V}$, we have $\Delta_k H = 0$. Since H is orthogonal to the central subspace, we have $\mu_k^\top H = 0$, and therefore $H^\top \hat{\Sigma}^{-1/2} \hat{\Delta}_k \hat{\Sigma}^{-1/2} H = H^\top \bar{\phi}_k H + o_p(n^{-1/2})$. Let $B_k = H^\top \hat{A}_k H$ and $B = (B_1, \dots, B_s)^\top$. By $\hat{A}_k = \hat{f}_k^{-1/2} \hat{\Sigma}^{-1/2} (\hat{\Delta}_k - \hat{f}_k \bar{\xi}) \hat{\Sigma}^{-1/2} + o_p(n^{-1/2})$ and $\hat{f}_k = f_k + O_p(n^{-1/2})$ we have

$$\begin{aligned} B_k &= f_k^{-1/2} H^\top (\bar{\phi}_k - \bar{\xi}) H + o_p(n^{-1/2}) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{J_{k,i} - f_k}{f_k^{1/2}} (V_i V_i^\top - I_r) - E \left\{ \frac{J_k - f_k}{f_k^{1/2}} (V V^\top - I_r) \right\} + o_p(n^{-1/2}), \end{aligned}$$

where $V_i = H^\top x_i$ and $r = p - m$. Let $g_i = ((J_{1,i} - f_1) f_1^{-1/2}, \dots, (J_{s,i} - f_s) f_s^{-1/2})^\top$. Then we have

$$B = \frac{1}{n} \sum_{i=1}^n \{g_i \otimes (V_i V_i^\top - I_r)\} - E\{G \otimes (V V^\top - I_r)\} + o_p(n^{-1/2}).$$

Since the $g_i \otimes (V_i V_i^\top - I_r)$ are independent and identically distributed with mean $E\{G \otimes (V V^\top - I_r)\}$, by the Central Limit Theorem, we can conclude that $n^{1/2} \text{vec}(B)$ converges in distribution to $N[0, \text{var}\{G \otimes (V V^\top - I_r)\}]$. The test statistics can be written as $T_n(\hat{H}) = (n/2) \|\text{vec}(B)\|^2$, and therefore $2T_n(\hat{H})$ converges in distribution to a weighted chi-squared distribution, with the weights given by the eigenvalues of $\text{var}\{G \otimes (V V^\top - I_r)\}$, which has at most $s(p - m)(p - m + 1)/2$ nonzero eigenvalues because $V V^\top - I_r$ is symmetric. \square

Proof of Corollary 1. Define $W = V V^\top - I_r$. Since $\text{var}(V) = E\{\text{var}(V|\gamma^\top X) + \text{var}\{E(V|\gamma^\top X)\}\}$, $\text{var}(V|\gamma^\top X)$ is nonrandom and $E(V|\gamma^\top X) = 0$, we have $\text{var}(V|\gamma^\top X) = \text{var}(V) = I_r$, and $E(W|\gamma^\top X) = 0$. Then

$$\begin{aligned} \text{var}(G \otimes W) &= \text{var}\{G \otimes \text{vec}(W)\} \\ &= E[\text{var}\{G \otimes \text{vec}(W)|\gamma^\top X\}] \\ &= E[E\{G G^\top \otimes \text{vec}(W) \text{vec}(W)^\top | \gamma^\top X\}] \\ &= E[E\{G G^\top | \gamma^\top X\} \otimes E\{\text{vec}(W) \text{vec}(W)^\top | \gamma^\top X\}]] \\ &= E\{E\{G G^\top | \gamma^\top X\} \otimes \text{var}(W) | \gamma^\top X\}. \end{aligned}$$

In the third and subsequent equalities we used the conditional independence of G and W given $\gamma^\top X$, and the fact that $E(W|\gamma^\top X) = 0$.

Assuming that $\text{var}(W|\gamma^\top X)$ is nonrandom, since $E(W|\gamma^\top X) = 0$, we have $\text{var}(W) = \text{var}(W|\gamma^\top X)$. Hence,

$$\text{var}(G^\top \otimes W) = \text{var}(G) \otimes \text{var}(W) = \text{var}(G) \otimes \text{var}(V V^\top).$$

Since $\text{var}(G) = I_s - \tau \tau^\top$, where $\tau = (f_1^{1/2}, \dots, f_s^{1/2})^\top$, $\text{var}(G)$ is a projection matrix with rank $s - 1$. The eigenvalues of $\text{var}(G) \otimes \text{var}(V V^\top)$ are the eigenvalues of the $r^2 \times r^2$ matrix $\text{var}(V V^\top)$, each with multiplicity $s - 1$. Since $V V^\top$ is symmetric, $\text{var}(V V^\top)$ has at most $r(r + 1)/2$ nonzero

eigenvalues. Let δ_j , $j = 1, \dots, r(r+1)/2$, be the largest $r(r+1)/2$ eigenvalues of $\text{var}(VV^T)$. Then

$$2T_n(\hat{H}) \longrightarrow \sum_j \delta_j \chi_j^2(s-1),$$

in distribution. □

REFERENCES

- BENTLER, P. & XIE, J. (2000). Corrections to test statistics in principal Hessian directions. *Statist. Prob. Lett.* **47**, 381–89.
- BURA, E. & COOK, R. D. (2001). Extended sliced inverse regression: the weighted chi-squared test. *J. Am. Statist. Assoc.* **96**, 996–1003.
- BURA, E. & PFEIFFER, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics* **19**, 1252–8.
- COOK, R. D. (1996). Graphics for regressions with a binary response. *J. Am. Statist. Assoc.* **91**, 983–92.
- COOK, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. New York: Wiley.
- COOK, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32**, 1062–92.
- COOK, R. D. & CRITCHLEY, F. (2000). Identifying regression outliers and mixtures graphically. *J. Am. Statist. Assoc.* **95**, 781–94.
- COOK, R. D. & LEE, H. (1999). Dimension reduction in regressions with a binary response. *J. Am. Statist. Assoc.* **94**, 1187–200.
- COOK, R. D. & NI, L. (2005). Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Am. Statist. Assoc.* **100**, 410–29.
- COOK, R. D. & WEISBERG, S. (1991). Discussion of a paper by K. C. Li. *J. Am. Statist. Assoc.* **86**, 328–32.
- EATON, M. L. (1986). A characterization of spherical distributions. *J. Mult. Anal.* **20**, 272–6.
- FLURY, B. & RIEDWYL, H. (1988). *Multivariate Statistics, A Practical Approach*. New York: John Wiley and Sons.
- LI, K. C. (1991). Sliced inverse regression for dimension reduction (with Discussion). *J. Am. Statist. Assoc.* **86**, 316–42.
- SCHOTT, J. R. (1997). *Matrix Analysis for Statistics*. New York: Wiley.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–88.
- TYLER, D. (1981). Asymptotic inference for eigenvectors. *Ann. Statist.* **9**, 725–36.
- WEISBERG, S. (2005). *Applied Linear Regression*, 3rd ed. New York: Wiley.
- ZHU, M. & HASTIE, T. J. (2003). Feature extraction for nonparametric discriminate analysis. *J. Comp. Graph. Statist.* **12**, 101–20.

[Received December 2005. Revised August 2006]