Title	Dimension reduction based on canonical correlation
Author(s)	Fung, WK; He, X; Liu, L; Shi, P
Citation	Statistica Sinica, 2002, v. 12 n. 4, p. 1093-1113
Issued Date	2002
URL	http://hdl.handle.net/10722/45357
Rights	Creative Commons: Attribution 3.0 Hong Kong License

Statistica Sinica **12**(2002), 1093-1113

DIMENSION REDUCTION BASED ON CANONICAL CORRELATION

Wing Kam Fung[1], Xuming He[2], Li Liu[1] and Peide Shi[1]

[1] *University of Hong Kong and* [2] *University of Illinois*

Abstract: Dimension reduction is helpful and often necessary in exploring nonlinear or nonparametric regression structures with a large number of predictors. We consider using the canonical variables from the design space whose correlations with a spline basis in the response space are significant. The method can be viewed as a variant of sliced inverse regression (SIR) with simple slicing replaced by B-spline basis functions. The asymptotic distribution theory we develop extends to weakly dependent stationary sequences and enables us to consider asymptotic tests that are useful in determining the number of significant dimensions for modeling. We compare several tests for dimensionality and make specific recommendations for dimension selection based on our theoretical and empirical studies. These tests apply to any form of SIR. The methodology and some of the practical issues are illustrated through a tuition study of American colleges.

Key words and phrases: Asymptotic distribution, canonical correlation, dimension reduction, mixing, sliced inverse regression, splines.

1. Introduction

Consider a regression problem with a response variable y and a predictor vector $x \in R^p$. If the relationship between x and y cannot be easily parameterized, it is often suggested that we turn to the modern arena of nonparametric regression. In recent years, advances in multivariate function estimation (see, e.g., Stone (1994)), in neural nets (see, e.g., Ripley (1994)) and in tree-based regression and classification methods (see, e.g., Breiman, Friedman, Olshen and Stone (1984)) have made it possible to quantify a highly nonlinear predictive relationship from a large number of predictors. However, the so-called curse of dimensionality can only be avoided through a simplification in the model. In the present paper, we consider a sub-dimensional model such that

$$y_i \perp x_i | (x_i^T \beta_1, \ldots, x_i^T \beta_K), \tag{2.1}$$

where the response y is independent of x given a $K \leq p$ dimensional sub-space spanned by $\{\beta_1, \ldots, \beta_K\}$. Special cases of (2.1) include transformed linear regression

$$h(y_i) = x_i^T \beta_1 + e_i, \tag{2.2}$$

or nonparametric additive model

$$y_i = \sum_{j=1}^{K} g_j(x_i^T \beta_j) + e_i, \tag{2.3}$$

where h is a monotone function, g_j are univariate functions, and the e_i represent random noise independent of x_i. If a suitable sub-space is found with a small K, then we would be working with a lower dimensional model favored by both interpretability and statistical efficiency.

Note that the model specification (2.1) does not uniquely determine the vectors $\beta_1, \ldots, \beta_K$. Cook's notion of minimal and central dimension reduction subspaces is designed to address this issue. We refer to Cook (1998) for details. In this paper, we assume that the central dimension reduction subspace exists so K is the smallest possible integer for (2.1) to hold.

A rather innovative tool, sliced inverse regression (SIR), for this dimension reduction problem has been developed by Li (1991) and Duan and Li (1991). Let $\mathbf{\Sigma} = \mathrm{Cov}\,(x)$ and $\mathbf{\Delta} = \mathrm{Cov}\,[E(x|y)]$. The work of SIR is based on a simple idea that under appropriate conditions $\mathbf{\Sigma}^{-1/2}$ times the eigenvectors of the matrix $\mathbf{\Sigma}^{-1/2}\mathbf{\Delta}\mathbf{\Sigma}^{-1/2}$ with nonzero eigenvalues fall into the space spanned by the effective directions β_i.

Several methods have been proposed in the literature to estimate $\mathbf{\Delta}$ or $\mathbf{\Lambda} = \mathbf{\Sigma} - \mathbf{\Delta}$ based on local averages or local covariances computed from points with neighboring y_i; see Aragon and Sarraco (1997) for a recent comparison. An alternative point of view is taken by He and Shen (1997). They consider finding the direction β_1 for model (2.2) that has the maximal correlation with some function of y; see also Chen and Li (1998). In this paper we follow the same approach and consider using all significant canonical directions.

Canonical correlation is a well understood notion in multivariate statistics. Most statistics software includes calculations of canonical variates as a standard procedure. In the special case $K = 1$, it was shown by Shi and Fung (1998) that the canonical correlation approach of He and Shen (1997) can be viewed as an iterative limit of the graphical method of transformation of Cook and Weisberg (1994). We hope that the use of splines and canonical correlation makes dimension reduction easier to understand and to fit into the nonparametric regression framework.

The second part of the paper is to consider tests on the dimensionality K that are asymptotically valid for rather general predictor variables. We compare three tests that are motivated from different aspects of the $\mathbf{\Delta}$ matrix. These tests apply to any form of SIR with the canonical correlation based method (CANCOR) as a leading example.

Most asymptotic studies in the literature on dimension reduction or nonparametric function estimation have assumed independent and identically distributed observations. In this paper, we relax the independence assumption and replace it by the β-mixing condition. This allows some common form of dependence in the data and makes the methodology applicable to some time series data as well.

The rest of the paper is organized as follows. We present the method of dimension reduction by canonical correlation in Section 2 and establish the asymptotic distributions of the estimated correlation in Section 3. The relationship between CANCOR and SIR is also made more explicit in these two sections. We consider the problem of determining the number of effective dimensions in Section 4 with several alternative tests discussed and compared. We find that the usual chi-square test on canonical correlations works well except for skewed or heavy-tailed predictors. In such highly non-Gaussian cases, a matrix rank test is preferable. In Section 5, our proposed method is applied to a dataset collected by U.S. News and World Report in an effort to model the out-of-state tuition of U.S. colleges using twenty available variables. We make some concluding remarks about CANCOR in Section 6 and provide technical proofs in Section 7.

2. Dimension Reduction by CANCOR

Suppose that $\boldsymbol{W}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is a sample of size n, where y_i is real-valued and $x_i \in R^p$. We assume that the response variable y is supported on a finite interval $[a, b]$, see a discussion of this assumption in He and Shen (1997). The basic idea is to build a B-spline basis for the y variable and find their correlations with other variables x. Following Schumaker (1981, p.224) and He and Shi (1994), we consider a partition $a = t_0 < t_1 < \cdots < t_H < t_{H+1} = b$ and let $\pi(y) \in R^{H+m}$ be the set of normalized B-spline basis functions of order m associated with this partition. H will be referred to as the number of internal knots. The common choices of m are 2 for linear splines, 3 for quadratic splines and 4 for cubic splines. In this paper, we use t_i as uniform partitions of $[a, b]$ or as the (i/H)th quantile of the observed y values so they are uniform in percentile ranks. The latter is used in all our empirical investigations reported in Sections 4 and 5. The minimum size of partition H should be chosen such that $H + m \geq K$, where K is the number of effective dimensions being sought. Here, we do not need the exact value K, but a reasonable upper bound will help. Note that our asymptotic analysis allows $H = H_n$ to grow with n.

Let $\boldsymbol{\Pi} = (\pi(y_1), \ldots, \pi(y_n))^T$ and $\boldsymbol{X} = (x_1, \ldots, x_n)^T$. We now consider the canonical correlation between the $H + m$ columns of $\boldsymbol{\Pi}$ and the p columns of $\boldsymbol{X}$. Let $r_{n,l}$ be the l-th canonical correlation coefficient (in decreasing order) and $z_l = x^T \hat{\beta}_l$ the corresponding canonical variate for $\boldsymbol{X}$ $(l = 1, \ldots, \min\{H + m, p\})$. The canonical directions $\hat{\beta}_l$ $(l = 1, \ldots, \hat{K})$ are then taken to be the effective

directions for dimension reduction. We take $\hat{K}$ to be the largest integer such that $r_{n,\hat{K}}$ is significantly different from zero. In Sections 3 and 4, we discuss in more detail the issue of determining the number of effective dimensions.

We restrict ourselves to stationary observations $\boldsymbol{W} = \{(x_i, y_i)\}$. The limiting behavior of canonical directions $\hat{\beta}_l$ can be understood through the matrix

$$\boldsymbol{\Delta} = \mathrm{Cov}\,(E(x_1|y_1)) \tag{2.1}$$

in a way similar to the sliced inverse regression (SIR). In fact, CANCOR is just a variation of SIR. The method of SIR estimates $\boldsymbol{\Delta}$ using a step function obtained from slicing. As we show in the proof of Theorem 1, the CANCOR method amounts to estimating $\boldsymbol{\Delta}$ by

$$\boldsymbol{\Delta}_n = n^{-1}\boldsymbol{X}^{*T}\boldsymbol{\Pi}(\boldsymbol{\Pi}^T\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^T\boldsymbol{X}^*, \tag{2.2}$$

whose eigenvalues and eigenvectors will be denoted by $\hat{\lambda}_l$ and $\hat{\eta}_l$, where $\boldsymbol{X}^* = (x_1 - \bar{x}, \ldots, x_n - \bar{x})^T$, and $\bar{x} = n^{-1}\sum_{i=1}^n x_i$. It is then easy to see that the canonical variates for $\boldsymbol{X}$ are $\hat{\beta}_l = (n^{-1}\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1/2}\hat{\eta}_l$. Here we have a spline-based estimate. The idea of using splines to estimate $\boldsymbol{\Delta}$ was mentioned briefly in the discussion of Li (1991) by Kent. The relationship between SIR and canonical correlation was explored by Chen and Li (1998). In the next section we show that CANCOR maintains the same asymptotic properties as SIR.

3. Asymptotic Properties of CANCOR

Let $\boldsymbol{\Delta} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^T$ be the spectral decomposition of $\boldsymbol{\Delta}$ where $\boldsymbol{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Here $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are eigenvalues of $\boldsymbol{\Delta}$. The columns of $\boldsymbol{Q}$, η_l $(l = 1, \ldots, p)$, are the eigenvectors of $\boldsymbol{\Delta}$. We use K^* to denote the number of nonzero eigenvalues of $\boldsymbol{\Delta}$.

The number of CANCOR directions $\hat{K}$ is an estimate of K^*. It was shown in Li (1991) that if, for any given $b \in R^p$, the design variable x satisfies the linearity condition

$$E(x^T b|x^T\beta_1, \ldots, x^T\beta_K) = c_0 + \sum_{i=1}^{K} c_i x^T\beta_i \tag{3.1}$$

for some constants c_i, then $K^* \leq K$ and $\boldsymbol{\Sigma}^{-1/2}$ times the eigenvectors of $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1/2}$ are contained in the space spanned by all the directions $\beta_1, \ldots, \beta_K$. Furthermore if K^*, the rank of $\boldsymbol{\Delta}$, is equal to K and $\boldsymbol{\Sigma} = \boldsymbol{I}$, we have $\beta_i = \eta_i$ $(i = 1, \ldots, K)$. These properties apply equally to any form of SIR (including CANCOR), and we refer the readers to Li (1991) and Cook (1998) for more details. However, the following two remarks are worth making.

Remark 3.1. If the linearity condition (3.1) does not hold, the directions found by CANCOR are not always characterizable by (2.1) but can still be useful in identifying some main features of the regression model. They have their own interpretation even when they are not estimating β_i, that is, they are linear components of x that have a significant correlation with some function of the response. Whether these directions are useful depends on the specific purposes of dimension reduction.

Remark 3.2. Asymptotic studies of SIR (see, e.g., Zhu and Ng (1995)) routinely assume that $\mathbf{\Sigma}$ is known so it becomes $\boldsymbol{I}$ after standardization of x_i. In fact, this is a convenient device for consistently estimating β_j. However, we emphasize that the asymptotic distributions of the direction estimates do not remain the same when $\mathbf{\Sigma}$ is estimated from data. This is further explained at the end of this section. On the other hand, the method of CANCOR requires no standardization, because the canonical variates are automatically scaled (even though the eigenvectors $\hat{\eta}_l$ are not).

We need some assumptions for the asymptotic distributional properties of $\hat{\lambda}_l$ and $\hat{\eta}_l$ from CANCOR, but this section requires neither (3.1) nor standardization of the x_i.

First, we recall that, given a positive integer k and a sequence $\boldsymbol{W} = \{(x_i, y_i)\}$, the β-mixing coefficient is $b_k(\boldsymbol{W}) = \sup_{j\geq 1} E \sup\{|P(B|F_j) - P(B)| : B \in F^{j+k}\}$, where F_j and F^j denote respectively the σ-fields generated by $\{(x_i, y_i) : 1 \leq i \leq j\}$ and $\{(x_i, y_i) : i \geq j\}$. Now the assumptions are as follows.

(A1) There is a positive constant δ such that $E||x_1||^{4+\delta} < \infty$.

(A2) For some $r > 2$, the β-mixing coefficient $b_k(\boldsymbol{W}) = O(k^r)$ as $k \to \infty$.

(A3) Each component of $\zeta(v) = E(x_1|y_1 = v)$ is a function on $[a, b]$ with bounded derivative.

(A4) The marginal density of y_1 is bounded away from 0 and infinity on $[a, b]$.

(A5) For some $\delta_0 > 0$, $n/H_n^4 \to 0$ and $n^{1-\delta_0}/H_n^2 \to \infty$.

Condition (A4) may appear to be a stringent requirement on the distribution of y_1, but note that we have an invariance property with a monotone transformation on the y variable so (A4) always holds if an appropriate transformation (such as one that is close to the c.d.f of y) is used. To understand the invariance property, note that $\lambda_1 = \max_{h,\beta} corr(h(y_1), x_1^T\beta)$ and $\hat{\beta}_1 = argmax_{h,\beta} corr(h(y_1), x_1^T\beta)$ where h is any monotone function on $[a, b]$; see He and Shen (1997) and Chen and Li (1998) for more details. Since a monotone transformation on y does not change the β-mixing coefficient, it is easy to see that, once properly re-written, some common time series models like $AR(p)$ can satisfy our conditions. Further consideration of time series data is not made in this paper.

Condition (A5) dictates that the number of knots used for the spline basis grows with n at a rate between $n^{1/4}$ and $n^{1/2}$. The upper bound on H_n is nearly necessary, but the lower bound can be relaxed for smoother functions of $\zeta(v)$ in (A3). The number of knots here plays a similar role as the number of slices for SIR.

Before we present the result, we need some notation. For any symmetric matrix $\boldsymbol{C}$, let $vec(\boldsymbol{C}) = (c_{11}, \ldots, c_{p1}, c_{12}, \ldots, c_{p2}, \ldots, c_{1p}, \ldots, c_{pp})^T$ be its vector version in p^2 dimensions and $vech(\boldsymbol{C}) = (c_{11}, \ldots, c_{p1}, c_{22}, \ldots, c_{p2}, c_{33}, \ldots, c_{pp})^T$ be a $\{p(p+1)/2\}$-dimensional vector taking only the lower triangular elements of $\boldsymbol{C}$. The relationship between vec and $vech$ is shown as $vec(\boldsymbol{C}) = \boldsymbol{\Phi} vech(\boldsymbol{C})$, where $\boldsymbol{\Phi}$ is a $p^2 \times p(p+1)/2$ matrix with elements

$$[\boldsymbol{\Phi}]_{ij,kl} = \begin{cases} 1 & \text{if } (i,j) = (k,l) \\ 1 & \text{if } (i,j) = (l,k) \\ 0 & \text{otherwise,} \end{cases}$$

with $i = 1, \ldots, p$, $j = 1, \ldots, p$, and $k \geq l = 1, \ldots, p$. For a singular matrix $\boldsymbol{C}$, we use $\boldsymbol{C}^+$ to denote its Moore-Penrose generalized inverse. Readers are referred to Schott (1997) for more details on the matrix operations we use here.

Theorem 1. *Under the conditions* $(A1) - (A5)$,

$$\sqrt{n}(\hat{\lambda}_l - \lambda_l) \to N(0, \sigma_l^2) \tag{3.2}$$

for $1 \leq l \leq p$. *Furthermore, if* $\lambda_1 > \cdots > \lambda_{K^*} > 0$, $\sqrt{n}(\hat{\eta}_l - \eta_l) \to N(0, \boldsymbol{\Sigma}_l)$ *for* $1 \leq l \leq K^*$, *where* $\sigma_l^2 = Var(\eta_l^T \boldsymbol{N} \eta_l)$, $\boldsymbol{\Sigma}_l = Cov((\boldsymbol{\Delta} - \lambda_l \boldsymbol{I})^+ \boldsymbol{N} \eta_l)$, *and* $\boldsymbol{N} = (x - E(x))(x - E(x))^T - (x - \zeta(y))(x - \zeta(y))^T$ *is a* $p \times p$ *random matrix.*

It is important to note that asymptotic normality holds only when $\sigma_l > 0$. Otherwise, $\hat{\lambda}_l$ converges to 0 faster than the root-n rate. The same phenomenon holds for SIR, even though it has not been pointed out in the existing literature.

For sliced inverse regression, and under the assumption of independence, similar asymptotic results to those of Theorem 1 were given by Hsing and Carroll (1992) and Zhu and Ng (1995). However, the asymptotic variance of $\hat{\beta}_l$ is not provided in such results (including our Theorem 1), unless $\boldsymbol{\Sigma}$ is known. This is because the definition of $\boldsymbol{\Delta}_n$ does not use standardized predictors. The estimated direction $\hat{\beta}_l$ is the eigenvector of $\boldsymbol{\Delta}_n$ only when the predictors x_i are standardized by replacing x_i by $\boldsymbol{\Sigma}_n^{-1/2} x_i$ where $\boldsymbol{\Sigma}_n = n^{-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$. Since $\sqrt{n} vech(\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma})$ is asymptotically normal, it would contribute to the variance-covariance of $\hat{\beta}_l$.

On the other hand, the number of nonzero eigenvalues for $\boldsymbol{\Delta}$ is invariant under any such standardization of the predictor x, so it is simpler to use the asymptotic results for $\hat{\lambda}_l$ without standardization.

4. Determining Dimensionality

One of the important questions in dimension reduction is to determine the number of effective dimensions. Like the traditional SIR, CANCOR does not always recover the central space in its entirety; in other words, the dimension K may be larger than K^* even when (3.1) holds. Other dimension reduction methods such as SAVE of Cook and Weisberg (see discussion of Li (1991)) may play a complementary role in finding directions missed by methods based on the $\mathbf{\Delta}$ matrix.

In this section, we aim to determine K^*, the number of nonzero eigenvalues of the matrix $\mathbf{\Delta}$. This may be done through sequential tests. In particular, we are interested in testing the null hypothesis $H_{0,k}$: $0 = \lambda_{k+1} < \lambda_k$ for some $k = 0, 1, \ldots, p-1$, against the alternative $\lambda_{k+1} > 0$, until the first time $H_{0,k}$ cannot be rejected. A number of authors have considered this problem for sliced inverse regression. Li (1991) used a chi-square approximation for normally distributed predictor x as a conservative guideline. Schott (1994) considered using both the first and second moments of the conditional distribution of x given y, and developed a chi-square test valid for any elliptically symmetric predictor distribution. In a somewhat different setting, with the method of principal Hessian directions (pHd), Cook (1998) constructed a test for the number of effective dimensions whose limiting distribution is a mixture of chi-squares. In this paper, we consider tests with simpler limiting distributions that are applicable to CANCOR or any form of SIR.

The chi-square test for normal predictors used in Li (1991), when adopted for CANCOR and called CHSQ-test here, is to reject $H_{0,k}$ for a given k if

$$-\{n - (p + H + m + 2)/2\} \sum_{i=k}^{p} \log(1 - \hat{\lambda}_i^2) > \chi^2_{(p-k+1)(H+m-k),\alpha},$$

where $\chi^2_{\nu,\alpha}$ is the upper α-th quantile of the chi-square distribution with ν degrees of freedom. This form comes from Anderson (1984, p.498) but is first-order equivalent to the chi-square test used in Li (1991). The test is asymptotically correct if x has a symmetric distribution with finite fourth moment.

It is helpful that we have an asymptotically valid test for dimensionality for more general predictor distributions including those without elliptical symmetry. Here, a natural approach is to use the asymptotic distribution of $\hat{\lambda}_{k+1}$. We reject $H_{0,k}$ if $\hat{\lambda}_{k+1} > z_\alpha \hat{\sigma}_{k+1}/\sqrt{n}$, where $\hat{\sigma}^2_{k+1}$ is taken to be the sample variance of $(\hat{\eta}^T_{k+1}(x_i - \bar{x}))^2 - (\hat{\eta}^T_{k+1}(x_i - m_i))^2$, $(i = 1, \ldots, n)$ with $m_i^T = \pi(y_i)^T(\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T\boldsymbol{X}$. This test, which will be called an ASNM-test in the paper, is valid when $\sigma_{k+1} > 0$. If $\sigma_{k+1} = 0$, the asymptotic level of the ASNM-test will not be α. Unfortunately, this is not rare.

A related test based on $S_n = n\sum_{i\geq k+1}\hat{\lambda}_i$ is considered by Velilla (1998), who considered another form of SIR using a finitely many number of observations per slice (corresponding to H_n in the same order as n). The limiting distribution of S_n was shown to be normal, but Velilla (1998) did not address the problem of whether the asymptotic variance is always positive. Since we do not suggest using a large H_n in CANCOR, we do not include this test in our comparisons.

One can also test dimensionality by considering the null hypothesis that the rank of the matrix $\boldsymbol{\Delta}$ is k. In fact, the problem of assessing the rank of a limit matrix has been studied by Tomišić and Simeon (1993) in chemometric applications and by Biok (1986) for ANOVA models. A rather general approach taken by Gill and Lewbel (1992) and Cragg and Donald (1996) is applicable here with some modifications to suit symmetric matrices. We proceed as follows.

Perform Gaussian elimination for k steps with rows and columns of $\boldsymbol{\Delta}_n$, so that we have

$$\boldsymbol{P}_{n,k}\boldsymbol{\Delta}_n\boldsymbol{P}_{n,k}^T = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{22} \end{pmatrix} \tag{4.1}$$

for some matrix $\boldsymbol{P}_{n,k}$, where $\boldsymbol{\Omega}_{22}$ is a $(p-k)\times(p-k)$ matrix. Each step of Gaussian elimination involves a possible exchange of rows and columns (in search of the largest absolute value among the diagonal elements whose row numbers are no smaller than the current one), and a row and then column operation to make zero all the off-diagonal elements whose row or column numbers are smaller than the current one. Under $H_{0,k}$, the matrix $\boldsymbol{P}_{n,k}$ has a limit $\boldsymbol{P}_k$ as $n\to\infty$, so (4.1) holds in its limit as the k-step Gaussian elimination on $\boldsymbol{\Delta}$.

If no rows or columns need to be exchanged in the process of Gaussian elimination, we have a simple expression for $\boldsymbol{\Omega}_{22} = \boldsymbol{\Delta}_{n,22} - \boldsymbol{\Delta}_{n,21}\boldsymbol{\Delta}_{n,11}^{-1}\boldsymbol{\Delta}_{n,12}$ when $\boldsymbol{\Delta}_n$ is naturally partitioned.

We later show in Lemma 7.1 that

$$\sqrt{n}vec(\boldsymbol{\Delta}_n - \boldsymbol{\Delta}) \to N(0, \mathrm{Cov}\,(vec(\boldsymbol{N}))), \tag{4.2}$$

and by arguments similar to those used in Cragg and Donald (1996) we get

$$\sqrt{n}vec(\boldsymbol{\Omega}_{22}) \to N(0, \boldsymbol{V}) \tag{4.3}$$

under $H_{0,k}$, where $\boldsymbol{V} = \mathrm{Cov}\,(vec(\boldsymbol{Q}))$ and $\boldsymbol{Q}$ is the lower $(p-k)\times(p-k)$ submatrix of $\boldsymbol{P}_k\boldsymbol{N}\boldsymbol{P}_k^T$. Thus

$$\hat{\xi} = n\ vec(\boldsymbol{\Omega}_{22})^T\boldsymbol{V}^+vec(\boldsymbol{\Omega}_{22}) \tag{4.4}$$

converges to the chi-square distribution with ν degrees of freedom, ν being the rank of $\boldsymbol{V}$. We reject the hypothesis $H_{0,k}$ when $\hat{\xi}$ is large and call this the RANK-test.

Although a test based on (4.4) does not require any particular form of the predictor distribution, it is not without caveats. The problem arises when ν equals 0. This can occur when $E(x_i|y) = E(x_i)$ for at least $p - k$ predictors. In our implementation, both $\boldsymbol{V}^+$ and ν are estimated. We estimate ν to be the number of eigenvalues of $\boldsymbol{V}$ exceeding a constant factor c_n times its largest eigenvalue, where c_n is taken to be the smaller of $1/100$ and $n^{-3/4}$. With this strategy, the RANK test is asymptotically correct as long as ν is nonzero. In case $\nu = 0$, the test is carried out assuming at least one degree of freedom so it is too conservative with type I error close to 0.

To investigate the finite-sample performance of our tests, we conducted a Monte Carlo simulation for a number of different models with $p = 5$ and $n = 100$ or 500. We also varied the predictor distribution.

The 1-dimensional model we consider is

$$y = x_1 + x_2 + e, \tag{4.5}$$

and the 2-dimensional model takes one of the two forms

$$y = x_1(1 + x_1 + x_2) + e, \tag{4.6}$$

$$y = x_1/(0.5 + (x_2 + 1.5)^2) + 0.5e, \tag{4.7}$$

where x_i $(i = 1, 2, 3, 4, 5)$ is distributed as F_i but e comes from some distribution G. A total of seven cases are reported in Table 1.

Table 1. Specification of Cases 1-7.

Case	F_1	F_2	F_3	F_4	F_5	G	Model	K	K^*	(3.1)
1	Z	Z	Z	Z	Z	t_5	(4.5)	1	1	Yes
2	Z	B	B	Z	L	0.05Z	(4.6)	2	2	Yes
3	C	-C	C	C	C	Z	(4.5)	1	2	No
4	C	C	C	C	C	C	(4.6)	2	2	Yes
5	t_3	t_3	t_3	t_3	t_3	t_3	(4.5)	1	1	Yes
6	Z	Z	Z	Z	Z	Z	(4.7)	2	2	Yes
7	Z	Z	Z	Z	*	Z	(4.5)	1	1	Yes

Columns 2-6 of Table 1 specifiy the distributions of F_i and G, where Z stands for the standard normal, B for Bernoulli, C for $\chi^2_1 - 1$, L for lognormal, and t_v for Student's distribution with v degrees of freedom. All the x_i and ϵ are independent of one another with the exception * in the table for F_5 of Case 7. In this case, x_5 is taken from the distribution $N(x_1 + x_2, 10^{-6})$. Column 8 specifies the form of the model, and Column 10 indicates whether the linearity condition (3.1) holds.

The central space exists in all the above cases, although the linearity condition fails in Case 3, and linear dependence among predictors is present in Case 7. To see that (3.1) fails in Case 3, let $g(s) = E[x_1|x_1 + x_2 = s]$ where x_1 and $-x_2$ are independent and identically distributed as χ^2_1. Then it is easy to show by symmetry that $g(s) = s + g(-s)$, which cannot hold if $g(s) > 0$ were linear in $s \in R$.

For each case, we draw 300 samples of size $n = 100$ and 300 samples of size $n = 500$. For each sample, we sequentially apply the CHRQ, ASNM and RANK tests to select the dimension. Tables 2−8 contain the frequencies with which the three tests select various dimensions with the two sample sizes.

To understand Tables 2−8 consider, for example, Table 2 with $n = 500$. When the true model is one-dimensional and the predictors are all normal, CHSQ chooses a one-dimensional model 284 out of 300 times. The method of RANK always picks a one-dimensional model in this case, while ASNM makes frequent errors (64+4 times out of 300) in picking one or two extra dimensions.

Table 2. Frequencies of Selected Model Dimensions with Case 1.

Test	n	$K = 0$	$K = 1$	$K = 2$	$K = 3$	$K = 4$
CHSQ	100	0	292	8	0	0
	500	0	284	15	1	0
ASNM	100	0	249	50	1	0
	500	0	232	64	4	0
RANK	100	21	279	0	0	0
	500	0	300	0	0	0

Table 3. Frequencies of Selected Model Dimensions with Case 2.

Test	n	$K = 0$	$K = 1$	$K = 2$	$K = 3$	$K = 4$
CHSQ	100	0	47	246	7	0
	500	0	0	289	12	0
ASNM	100	78	156	54	10	2
	500	0	55	212	32	1
RANK	100	104	80	93	22	1
	500	0	28	254	18	0

Table 4. Frequencies of Selected Model Dimensions with Case 3.

Test	n	$K = 0$	$K = 1$	$K = 2$	$K = 3$	$K = 4$
CHSQ	100	0	1	280	18	1
	500	0	0	279	21	0
ASNM	100	36	31	197	35	1
	500	0	0	248	52	0
RANK	100	0	77	187	36	0
	500	0	0	300	0	0

Table 5. Frequencies of Selected Model Dimensions with Case 4.

Test	n	$K=0$	$K=1$	$K=2$	$K=3$	$K=4$
CHSQ	100	0	2	270	27	1
	500	0	0	280	20	0
ASNM	100	53	36	175	36	0
	500	0	0	246	52	2
RANK	100	101	150	49	0	0
	500	0	3	297	0	0

Table 6. Frequencies of Selected Model Dimensions with Case 5.

Test	n	$K=0$	$K=1$	$K=2$	$K=3$	$K=4$
CHSQ	100	0	181	112	7	0
	500	0	106	192	2	0
ASNM	100	58	190	49	3	0
	500	22	254	21	3	0
RANK	100	66	233	1	0	0
	500	9	290	1	0	0

Table 7. Frequencies of Selected Model Dimensions with Case 6.

Test	n	$K=0$	$K=1$	$K=2$	$K=3$	$K=4$
CHSQ	100	0	101	191	8	0
	500	0	0	291	8	1
ASNM	100	1	58	187	54	0
	500	0	0	231	69	0
RANK	100	86	188	26	0	0
	500	0	0	300	0	0

Table 8. Frequencies of Selected Model Dimensions with Case 7.

Test	n	$K=0$	$K=1$	$K=2$	$K=3$	$K=4$
CHSQ	100	0	289	11	0	0
	500	0	286	12	2	0
ASNM	100	0	287	13	0	0
	500	0	282	18	0	0
RANK	100	0	300	0	0	0
	500	0	300	0	0	0

It is worth noting that Case 3 is a one-dimensional model ($K = 1$) with $K^* = 2$. We find through our investigation that the first CANCOR direction is close to $(1, 1, 0, 0, 0)$, the direction given in model (4.5), but the tests point to $K^* = 2$ dimensions in consistency with theory.

Our study suggests that the CHSQ test is a simple and reliable choice except when the variables are highly skewed or heavy-tailed. In Case 5, CHSQ often picks some extra dimensions even when the right direction has already been well estimated by the first CANCOR direction. The RANK test is more robust but tends to be conservative for small to modest sample sizes. The ASNM test is less predictable. These results are rather consistent with the theoretical aspects we discussed in this section.

There is severe collinearity in the predictors in Case 7. The tests for dimensionality are hardly affected by collinearity, even though the estimated CANCOR direction is unable to choose between $(1, 1, 0, 0, 0)$ and $(0, 0, 0, 0, 1)$.

Our recommendation is to use the simple CHSQ test except for the cases with highly skewed or heavy-tailed predictors. The CHSQ test is especially nonrobust against outliers. In cases where CHSQ tends to fail, we suggest using the RANK test as a conservative way to choose dimensionality. We also suggest examination of the plots of y against the next canonical variate not chosen by RANK to see if one is missing a meaningful dimension. A more effective graphical strategy to approximate the central space "from above" can be found in Cook (1998a) and Chiaromonte and Cook (1997).

5. College Tuition Example

We consider an example of college tuition based on data from the U.S. News & World Report's Guide to America's Best Colleges (1995). The data contains information on tuition, room and board costs, SAT or ACT scores, application/acceptance rates, graduation rate, student/faculty ratio, spending per student, and a number of other variables for over 1300 schools in the U.S. We wish to explore the relationship between tuition and 20 other characteristics variables listed below. The out-of-state tuition is taken to be the response for both public and private schools. For illustration, only a subset of 271 schools without missing values are used in this analysis. The issue of potential bias in this selection is not pursued here. The full data may be found in http://lib.stat.cmu.edu/datasets/colleges/.

The predictor variables used in our analysis are as follows: 1. Public/private indicator (public=1, private=2); 2. Average Combined SAT score; 3. Average ACT score; 4. Number of applications received; 5. Number of applicants accepted; 6 Number of new students enrolled; 7. Percent of new students from top 10% of high school class; 8. Percent of new students from top 25% of high school class; 9. Number of full-time undergraduates; 10. Number of part-time undergraduates; 11. Room and board costs; 12. Additional fees; 13. Estimated book costs; 14. Estimated personal spending; 15. Percent of faculty with Ph.D.'s; 16. Percent of faculty with terminal degree; 17. Student/faculty ratio; 18. Percent of

alumni who donate; 19. Instructional expenditure per student; 20. Graduation rate.

A simple examination of those variables shows that many are heavily skewed and there are outliers in some of them. We take the $logx$ transform of variables 4, 5, 6, 9, 10, 12, 13, 14, 17, 19 and the $log(x/(100-x))$ transform of percentage variables 7, 8, 15, 16 and 18 before using CANCOR. These transformations improve symmetry and normality but a few outliers remain.

We use four internal knots to construct the B-spline basis functions of order three. The results change little when we use one more or one fewer knot and vary the order of splines. All three tests of dimensionality discussed in Section 3 are performed with the resulting p-values in Table 9. All tests agree that the first two CANCOR directions are significant. Only CHSQ leads to significance of the third direction, probably due to the effect of outlying observations.

Table 9. P-values from dimensionality tests for tuition example.

$H_{0,k}$	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
CHSQ-test	0.0000	0.0000	0.0212	0.3137	0.8029
ASNM-test	0.0000	0.0171	0.2525	0.3862	0.2717
RANK-test	0.0000	0.0366	0.7844	0.9502	0.9990

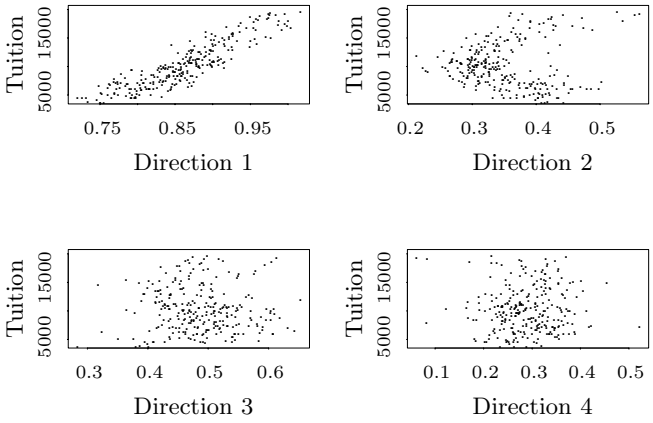

Figure 1. Scatter plots of tuition versus first four CANCOR directions.

Figure 1 shows the scatter plots of tuition against the first four directions found by CANCOR. The first canonical variate z_1 appears linearly related to tuition and the trend is major, the second variate z_2 provides a refinement. It turns out that z_1 is almost the same as a multiple linear regression fit. This

is consistent with a result of Duan and Li (1991) that, under (3.1), the OLS estimates a direction within the central space. A closer inspection shows that z_1 is highly related to quality of the schools. The variable z_2 is highly related to school size. From Figure 1, we see that tuition generally increases with z_2 for larger values of z_1 but decreases with z_2 for smaller values of z_1. There is no visible relationship between tuition and the third variate, which fits well with a 2-dimensional model. Not surprisingly the results are also very similar to those obtained by sliced inverse regression. The most important contribution to z_1 comes from instructional expenditure per student.

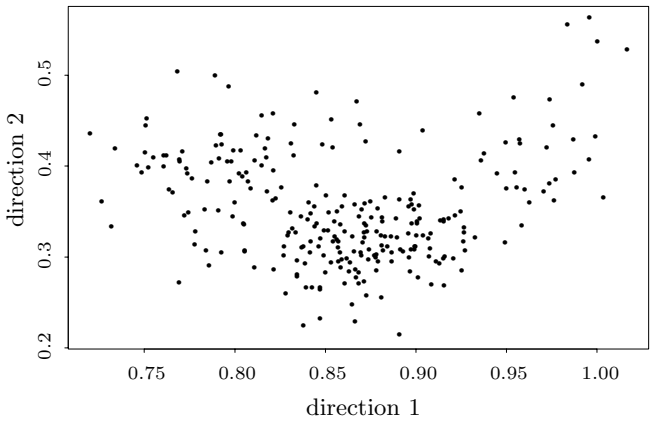

Figure 2. Scatter plots of the first two CANCOR directions.

Figure 2 shows there is a nonlinear relationship between the first and the second CANCOR directions. This "nonlinear confounding" issue, addressed by Li (1997), can make it difficult to analyze data using SIR or CANCOR. We also refer to the Boston Housing data example in Chen and Li (1998) where a similar phenomenon occurs. More discussions may be found in Velilla (1998) and Cook (1998a, Section 13.2). In our example, the first CANCOR direction plays a dominant role, but the quasi-helix plot found in the second direction explains the data from a different angle, and leads us to pay more attention to the school size that might be neglected without CANCOR or a similar procedure. Further examination of the first two variates indicate that they are not dominated by a small number of original variables. This suggests that projection is preferred to variable selection in building a model with a large number of variables.

Neural networks are known to be flexible in approximating functional relationship. We compare our results with those obtained from single hidden layer neural networks. We fit the following form of the feed-forward single layer neural

network

$$f(x) = \alpha_0 + \sum_{l=1}^{K} \alpha_l \psi(\gamma_l^T x)$$

where $\psi(u) = e^u/(1+e^u)$. The function fitted in this way may be compared with CANCOR if we take $\gamma_l^T x$ as reduced variables. Computations are made with the FUNFITS package written at the North Carolina State University, more detail is in Nychka et. al (1998). Figure 4 shows the plots of tuition against the two directions found by FUNFITS. The first one appears to reveal the same, mostly linear, relationship as in Figure 2, but the second direction is less informative. R^2 from this neural network is 83%, as compared to 86% from a tensor-product spline model (see He and Shi, 1996) based on the two canonical variates we found with CANCOR. Also note that the multiple linear regression gave an R^2 of 80%, which is consistent with the dominating linear relationship between our response and the first canonical variate in this example.

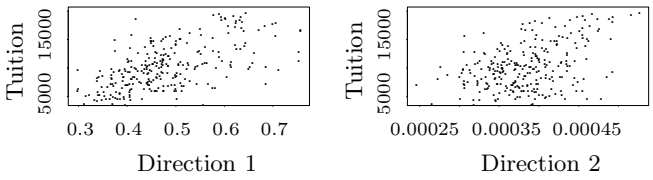

Figure 4. Scatter plots of tuition versus two directions chosen by FUNFITS.

Finally we note that with 20 variables but only 271 observations, we are pushing the limits of the asymptotic theory in Section 3. The example also shows that it is the best to combine test of dimensionality with graphical inspection in data exploration and analysis.

6. Concluding Remarks

In this paper, we provide a variant of SIR by using canonical correlation between the predictors and a spline basis in the space of the response variable. Its properties are parallel to those of the traditional SIR. In particular, the CANCOR and SIR directions are estimating the same quantities and are consistent for vectors in the central dimension reduction space under the same conditions on the model. However, the CANCOR directions are more directly interpretable outside the context of a central space. The asymptotic distributions obtained in Section 3 do not assume the linearity condition (3.1) and form the basis for determining the number of nonzero correlations. Unlike earlier work on SIR, we do not assume the predictor variables to be normalized to have unit covariance matrix. This allows some inference on the matrix $\mathbf{\Delta}$ to be carried out without having to incorporate

the sampling distribution of $\hat{\mathbf{\Sigma}}$. In this paper, we also extend the applicability of the dimension reduction method by relaxing the assumption of independent observations.

A comparison of three different types of tests (CHSQ, ASNM, RANK) for dimensionality reveals the strength and weakness of each test. Our findings can help us make good use of these tests in real applications.

7. Proofs

Theorem 1 and the asymptotic results in Section 4 are derived from the following lemma. Without loss of generality, assume $E(x_1) = 0$.

Lemma 7.1. *Under the conditions of Theorem* 1, $\sqrt{n}(\mathbf{\Delta}_n - \mathbf{\Delta}) \longrightarrow \boldsymbol{N}$, *in distribution, where* $\boldsymbol{N}$ *is as defined in Theorem* 1.

Proof of Theorem 1. By Lemma 7.1 above and Theorem 8.5 of Schott (1997, pp. 342-344), we have $\sqrt{n}(\hat{\eta}_l - \eta_l) = -(\mathbf{\Delta} - \lambda_l \boldsymbol{I})^+ \sqrt{n}(\mathbf{\Delta}_n - \mathbf{\Delta})\eta_l + o_P(1)$, and $\sqrt{n}(\hat{\lambda}_l - \lambda_l) = \eta_l^T \sqrt{n}(\mathbf{\Delta}_n - \mathbf{\Delta})\eta_l + o_P(1)$, from which the theorem follows.

To prove Lemma 7.1, it suffices to show that

$$\sqrt{n}\,(\mathbf{\Delta}_n - \mathbf{\Delta}) = \sqrt{n}\left(\tilde{\mathbf{\Delta}}_n - \mathbf{\Delta}\right) + o_P(1), \tag{7.1}$$

where $\tilde{\mathbf{\Delta}}_n = n^{-1}\sum_{k=1}^n x_k x_k^T - n^{-1}\sum_{k=1}^n (x_k - \zeta(y_k))(x_k - \zeta(y_k))^T$. Recall that we have assumed $E(x_k) = 0$.

The rest of this section is devoted to the proof of (7.1). For convenience, we first quote three lemmas that will be used in the proof. Lemma 7.2 can be obtained from Corollary 6.21 of Schumaker (1981, p.227) or Theorem XII.4 of de Boor (1978, p.178) for splines of order $m > 1$. In the case of $m = 1$, it can also be verified directly. Lemma 7.3 can be found in Chen (1991). Lemma 7.4 is a generalization of Lemma 4.4 of Shi (1997) to the sample with β-mixing conditions.

Let $\zeta_j(v)$ be the j-th component of $\zeta(v)$ and $\theta^{(j)}$ be its spline coefficient with $\zeta_j(v) = \pi(v)^T\theta^{(j)} + R_{nvj}$.

Lemma 7.2. *Assume Condition* (A3). *There exists a constant* c *depending only on* m *and* $\zeta(v)$ *such that, for all* $j = 1, \ldots, p$,

$$\sup_v |R_{nvj}| \le cH_n^{-1}. \tag{7.2}$$

Lemma 7.3. *Assume Conditions* (A2), (A4), *and* (A5). *There exist positive constants* c_1 *and* c_2 ($c_2 > c_1$) *such that all eigenvalues of* $\frac{H_n}{n}(\mathbf{\Pi}^T\mathbf{\Pi})$ *lie in* (c_1, c_2) *with probability tending to* 1 *as* $n \to \infty$.

Lemma 7.4. *Under the conditions of Theorem* 1*, we have*

$$\sup_{\alpha^T\alpha=1}\frac{1}{\sqrt{n}}|\sum_{k=1}^{n}(x_{ki}-\zeta_i(y_k))\pi(y_k)^T\alpha| = n^{-1/2}||\sum_{k=1}^{n}(x_{ki}-\zeta_i(y_k))\pi(y_k)|| = O_P(1), \tag{7.3}$$

and

$$n^{-1/2}H_n|\sum_{k=1}^{n}(x_{ki}-\zeta_i(y_k))R_{nkj}| = O_P(1), \tag{7.4}$$

for all $i, j = 1, \ldots, p$.

Proof of (7.1). We define some vectors in bold face that will only be used in this proof. Let x_{ij} be the j-th component of $x_i \in R^p$ $(i = 1, \ldots, n)$ and $\bar{x}_j$ be the average of x_{ij} over i. Similarly, $\bar{\zeta}_j$ is the average of $\zeta_j(y_i)$ over i. Let $\boldsymbol{\xi}_j = (x_{1j} - \bar{x}_j, \ldots, x_{nj} - \bar{x}_j)^T \in R^n$ be the j-th column of $\boldsymbol{X}^*$, $\boldsymbol{u}_j = (u_{1j}, \ldots, u_{nj})^T \in R^n$ with $u_{ij} = x_{ij} - \bar{x}_j - \zeta_j(y_i) + \bar{\zeta}_j$. Finally, let $\boldsymbol{G} = \boldsymbol{\Pi}(\boldsymbol{\Pi}^{\boldsymbol{T}}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^{\boldsymbol{T}}$.

It is easy to see that the ij-th element of $\tilde{\boldsymbol{\Delta}}_n$ is equal to $n^{-1}(\boldsymbol{\xi}_i^T\boldsymbol{\xi}_j - \boldsymbol{u}_i^T\boldsymbol{u}_j) + o_P(n^{-1/2})$. On the other hand, decomposing $\boldsymbol{\xi}_i^T(\boldsymbol{I} - \boldsymbol{G})\boldsymbol{\xi}_j$ yields

$$\boldsymbol{\xi}_i^T\boldsymbol{\xi}_j - \boldsymbol{u}_i^T\boldsymbol{u}_j = \boldsymbol{\xi}_i^T\boldsymbol{G}\boldsymbol{\xi}_j - \boldsymbol{u}_i^T\boldsymbol{G}\boldsymbol{u}_j + 2\boldsymbol{u}_i^T(\boldsymbol{I} - \boldsymbol{G})\boldsymbol{\nu}_j + \boldsymbol{\nu}_i^T(\boldsymbol{I} - \boldsymbol{G})\boldsymbol{\nu}_j. \tag{7.5}$$

So it remains to show that

$$-\boldsymbol{u}_i^T\boldsymbol{G}\boldsymbol{u}_j + 2\boldsymbol{u}_i^T(\boldsymbol{I} - \boldsymbol{G})\boldsymbol{\nu}_j + \boldsymbol{\nu}_i^T(\boldsymbol{I} - \boldsymbol{G})\boldsymbol{\nu}_j = o_P(n^{1/2}). \tag{7.6}$$

We show that each term in (7.6) is of the desired order. Here, it helps to note that $\boldsymbol{G}$ and $\boldsymbol{I} - \boldsymbol{G}$ are idempotent. Since $(\boldsymbol{I} - \boldsymbol{G})\boldsymbol{\nu}_j$ are the residuals of projecting $\boldsymbol{\nu}_j$ onto the space of $\boldsymbol{\Pi}$, we have

$$|\tfrac{1}{n}\boldsymbol{\nu}_i^T(\boldsymbol{I} - \boldsymbol{G})\boldsymbol{\nu}_j| \le (\frac{1}{n}\sum_{k=1}^{n}(R_{nki} - \bar{\zeta}_i)^2)^{1/2}(\frac{1}{n}\sum_{k=1}^{n}(R_{nkj} - \bar{\zeta}_j)^2)^{1/2}. \tag{7.7}$$

Observe, from Lemma 7.2, that

$$\frac{1}{n}\sum_{k=1}^{n}(R_{nkj} - \bar{\zeta}_j)^2 \le \frac{2}{n}\sum_{k=1}^{n}R_{nkj}^2 + 2(\bar{\zeta}_j)^2 = O(H_n^{-2}) + O_P(n^{-1}).$$

Together with (7.7) and Condition (A5), one has

$$\boldsymbol{\nu}_i^T(\boldsymbol{I} - \boldsymbol{G})\boldsymbol{\nu}_j = o_P(n^{1/2}). \tag{7.8}$$

Let $\boldsymbol{u}_i^* = (x_{1i} - \zeta_i(y_1), \ldots, x_{ni} - \zeta_i(y_n))^T \in R^n$, $\kappa_i = ||\sum_{k=1}^{n}(x_{ki} - \zeta_i(y_k))\pi(y_k)^T(\boldsymbol{\Pi}^T\boldsymbol{\Pi})^{-1/2}||$ and $\kappa_i^* = ||\sum_{k=1}^{n}(x_{ki} - \zeta_i(y_k))\boldsymbol{\pi}(y_k)||(H_n/n)^{1/2}$. From Lemma 7.3 we have $(\frac{H_n}{n}\boldsymbol{\Pi}^T\boldsymbol{\Pi})^{-1} = O_P(1)$ and

$$|\boldsymbol{u}_i^{*T}\boldsymbol{G}\boldsymbol{u}_j^*| \le \kappa_i\kappa_j \le O_P(\kappa_i^*\kappa_j^*) \tag{7.9}$$

for $i, j = 1, \ldots, p$. Note that $\pi_l(\cdot)$ is locally supported, $\sup_t \pi(t)^T\pi(t) \le m+3$ and $E(\pi(y_1)) = O(H_n^{-1})$. By Lemma 7.4 we have $\kappa_i^* = O_P(H_n)$, and therefore

$$|\boldsymbol{u}_i^{*T}\boldsymbol{G}\boldsymbol{u}_j^*| = O_P(H_n) \quad \text{for all} \quad 1 \le i,\ j \le p. \tag{7.10}$$

Let $\mathbf{1}$ be the n-vector of ones. We have $\boldsymbol{u}_i = \boldsymbol{u}_i^* + \mathbf{1}(\bar{\zeta}_i - \bar{x}_i)$, and

$$\boldsymbol{u}_i^T\boldsymbol{G}\boldsymbol{u}_j - \boldsymbol{u}_i^{*T}\boldsymbol{G}\boldsymbol{u}_j^* = (\bar{\zeta}_i - \bar{x}_i)(\bar{\zeta}_j - \bar{x}_j)\mathbf{1}^T\boldsymbol{G}\mathbf{1} + \boldsymbol{u}_i^{*T}\boldsymbol{G}\mathbf{1}(\bar{\zeta}_j - \bar{x}_j) + (\bar{\zeta}_i - \bar{x}_i)\mathbf{1}^T\boldsymbol{G}\boldsymbol{u}_j^*. \tag{7.11}$$

By the Cauchy-Schwartz inequality and the fact that $\mathbf{1}^T\boldsymbol{G}\mathbf{1} \le n$ (due to the eigenvalue of $\boldsymbol{G}$ being at most 1), we see that each term on the right hand side of (7.11) is $O_P(H_n)$. Therefore, we have

$$\boldsymbol{u}_i^T\boldsymbol{G}\boldsymbol{u}_j = O_P(H_n) = o_P(n^{1/2}) \quad \text{for all} \quad 1 \le i,\ j \le p. \tag{7.12}$$

Similar arguments show that

$$\boldsymbol{u}_i^T(\boldsymbol{I}-\boldsymbol{G})\boldsymbol{\nu}_j = \boldsymbol{u}_i^T\boldsymbol{R}_{nj} - \boldsymbol{u}_i^T\boldsymbol{G}\boldsymbol{R}_{nj} + \boldsymbol{u}_i^T(\boldsymbol{I}-\boldsymbol{G})\mathbf{1}(\mu_j - \bar{\zeta}_j) = o_P(n^{1/2}). \tag{7.13}$$

The proof of (7.1) is then complete.

Finally, we prove (7.3) in Lemma 7.4. The proof for (7.4) is similar and thus omitted.

For an integer pair (υ_n, τ_n) with $\tau_n = [n/(2\upsilon_n)]$, we divide the strictly stationary n-sequence $\boldsymbol{W}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\})$ into $2\tau_n$ blocks of length υ_n and the remainder block of length $n - 2\upsilon_n\tau_n$ ($\le \upsilon_n$). Denote the index sets of the odd blocks and the even blocks by O's and E's respectively, and denote the index set of the remainder block by R_e. That is, $R_e = \{i : 2\upsilon_n\tau_n + 1 \le i \le n\}$, $O_j = \{i : 2(j-1)\upsilon_n + 1 \le i \le (2j-1)\upsilon_n\}$ and $E_j = \{i : (2j-1)\upsilon_n + 1 \le i \le 2j\upsilon_n\}$ for $j = 1, \ldots, \tau_n$. Let $\boldsymbol{W}(O_j) = \{(X_i, Y_i) : i \in O_j\}$ for $j = 1, \ldots, \tau_n$.

Under the mixing conditions, we can choose υ_n large enough so that the dependence between the odd υ_n-blocks is weak and therefore the odd υ_n-blocks can be approximated by a sequence of independent blocks with the same within-block structure. On the other hand υ_n is not so large and the odd υ_n-blocks together behave similarly to the original mixing sequence. Specifically, we construct an independent sequence of blocks $\boldsymbol{W}^*(O_j) = \{(X_i^*, Y_i^*) : i \in O_j\}$ such that $\boldsymbol{W}^*(O_j)$ and $\boldsymbol{W}(O_j)$ have the same distribution and $\boldsymbol{W}^* = \{\boldsymbol{W}^*(O_j) : j = 1, \ldots, \tau_n\}$ is independent of $\boldsymbol{W}_n$. $\boldsymbol{W}^*$ is called an independent block υ_n-sequence (IB sequence), which is connected implicitly with a pair of integers. The following lemma is used by Yu (1994).

Lemma A.1. *Let the distributions of $\boldsymbol{W}_n$ and $\boldsymbol{W}^*$ be Q and Q^* respectively. For any measurable function h on $R^{\tau_n\upsilon_n}$ with bound C, we have $|E_Q h(\boldsymbol{W}_n) - E_{Q^*}h(\boldsymbol{W}^*)| \le C(\tau_n - 1)b_{\upsilon_n}(\boldsymbol{W})$.*

For simplicity, we assume uniform partitions for our B-spline knot sequences: $t_k = k/H_n$, $k = 1, \ldots, H_n$ and $N = H_n + m - 1$. Only nonessential modifications are needed to deal with the percentile-based partitions described in Section 2. The letter M is used in the rest of the proof to denote a generic constant whose value may vary from line to line.

To prove (7.3) for any $l = 1, \ldots, p$, just write $\Psi_k \hat{=} x_{kl} - \zeta_l(y_k)$ for $k = 1, \ldots, n$. We have suppressed the index l here. It has been assumed that $E(\Psi_1^{2+c_0}) < \infty$ with constant $c_0 = 2 + \delta$. Let $\upsilon_n = n^b$ with $b = 1/(r - c_1 + 1)$, where $0 < c_1 < \min\{r - 1, [(r-1)(2+c_0) - 2]/(2+c_0)\}$, and r is a constant given in Condition (A2). Then, we have

$$2b \in (0,1) \qquad \text{and} \qquad 2b/(1-2b) = 2/(r - c_1 - 1) < 2 + c_0. \tag{7.14}$$

Lemma A.2. *Under the conditions of Theorem 1, for $q = 2 + c_0$, $s^{-1} + 2q^{-1} = 1$, we have*

$$E\left(\sum_{i \in O_j} \sum_{k \in O_j} \pi(y_i)^T \pi(y_k) \Psi_i \Psi_k\right) \le \sum_{i \in O_j} E\left(|\pi(y_i)|^2 \Psi_i^2\right) + M\, \upsilon_n^{1-\delta_0},$$

where $\delta_0 = rc_0/(2+c_0) - 1 = r(2+\delta)/(4+\delta) - 1 > 0$, and M is some constant.

The proof of Lemma A.2 is based on a result of Dehling and Philipp (1982, p. 692), but the details are omitted here.

Lemma A.3. *Under the conditions of Theorem* 1,

$$\lim_{n \to \infty} \mathcal{P}\left(n^{-1/2} \Big| \sum_{i = 2\tau_n \upsilon_n + 1}^{n} \Psi_i \pi(y_i)\Big| \ge L\right) = 0 \quad \text{for any } \ L > 0.$$

Lemma A.3 can be verified using the Tchebychev inequality. We now prove (7.3). Let $S_i^* = (X_i^*, y_i^*)$, $i \ge 1$, be the IB sequence corresponding to $\boldsymbol{W}$. Since $\boldsymbol{W}$ is strictly stationary we have, from Lemmas A.1 and A.3, that for large L

$$\begin{aligned} &\mathcal{P}\{n^{-1/2} \sup_{|\alpha|=1} |\sum_{i=1}^{n} \Psi_i \pi(y_i)^T \alpha| > L\} \\ &\le 2\mathcal{P}\{n^{-1/2} \sup_{|\alpha|=1} |\sum_{j=1}^{\tau_n} \sum_{i \in O_j} \Psi_i^* \pi(y_i^*)^T \alpha| > L/4\} + 2(\tau_n - 1) b_{\upsilon_n}(\boldsymbol{W}) + o(1). \end{aligned} \tag{7.15}$$

Note that $\{(X_i, Y_i)\ i \in O_j\}$ and $\{(X_i^*, y_i^*)\ i \in O_j\}$ have the same distribution. Thus, $E(\sum_{i \in O_j} \Psi_i^* \pi(y_i^*)^T \alpha) = E(\sum_{i \in O_j} \Psi_i \pi(y_i)^T \alpha) = E(\sum_{i \in O_j} \pi(y_i)^T \alpha\, E(\Psi_i \mid y_i)) = 0$. Together with (7.15), Lemma A.2, the Tchebychev inequality and the

independence between the blocks of the IB sequence,

$$\mathcal{P}\{n^{-1/2}\sup_{|\alpha|=1}|\sum_{j=1}^{\tau_n}\sum_{i\in O_j}\Psi_i^*\pi(y_i^*)^T\alpha|>L/4\}\le\mathcal{P}\{n^{-1/2}|\sum_{j=1}^{\tau_n}\sum_{i\in O_j}\Psi_i^*\pi(y_i^*)|>L/4\}$$
$$\le\frac{16}{L^2n}\sum_{j=1}^{\tau_n}\text{trace}\left(E\sum_{i\in O_j}\sum_{k\in O_j}\Psi_i\pi(y_i)\Psi_k\pi(y_k)^T\right)\le\frac{16}{L^2}(4(m+3)E\Psi_1^2+M\,\upsilon_n^{-(r-1)}). \tag{7.16}$$

Since $\upsilon_n=n^{-b}$ and $r>2$, the desired result follows from (7.15) and (7.16).

Acknowledgement

Research of Wing Kam Fung and Peide Shi is partially supported by an RGC Grant from the University of Hong Kong. Research of Xuming He and Li Liu is supported in part by NSF Grants SBR 96-17278 and DMS 0102411. Part of the paper drew material from an earlier technical report of Shi and Fung, and part of the paper is included in the Ph.D. dissertation of Liu. The authors benefited from unusually insightful comments and suggestions of a referee on an earlier draft of the paper.

References

Anderson, T. W. (1984). *Multivariate Analysis*. Wiley, New York.

Aragon, Y. and Saracco, J. (1997). Sliced Inversed Regression (SIR): an appraisal of small sample alternatives to slicing. *Comput. Statist.* **12**, 109-130.

Boik, R. J. (1986). Testing the rank of a matrix with applications to the analysis of interaction in ANOVA. *J. Amer. Statist. Assoc.* **81**, 243-248.

Brieman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth Internatational Group, Belmont, California.

Chen, C. H. and Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* **8**, 289-316.

Chen, H. (1991). Polynomial splines and nonparametric regression. *J. Nonparametr. Statist.*, **1**, 143-156.

Chiaromonte, F. and Cook, R. D. (1997). On foundations of regression graphics. Technical Report, School of Statistics, University of Minnesota.

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

Cook, R. D. (1998a). *Regression Graphics: Ideas for Studying Regression through Graphics*. Wiley, New York.

Cook, R. D. and Weisberg, S. (1994). Transforming a response variable for linearity. *Biometrika* **81**, 731-737.

Cragg, J. G. and Donald, S. G. (1996). On the asymptotic properties of LDU-based tests of the rank of a matrix. *J. Amer. Statist. Assoc.* **91**, 1301-1309.

Dehling, H. and Philipp, W. (1982). Almost sure invariance principles for weakly dependent vector-valued random variables. *Ann. Probab.* **10**, 689-701.

Duan, N. and Li, K. C. (1991). Slicing regression: A link-free regression method. *Ann. Statist.* **19**, 505-530.

Gill, L. and Lewbel, A. (1992). Testing the rank and definiteness of estimated matrices with applications to factor, state-space and ARMA models. *J. Amer. Statist. Assoc.* **87**, 766-776.

He, X. and Shen, L. J. (1997). Linear regression after spline transformation. *Biometrika* **84**, 474-481.

He, X. and Shi, P. D. (1994). Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *J. Nonparametr. Statist.* **3**, 299-308.

He, X. and Shi, P. D. (1996). Bivariate tensor-product splines in a partly linear model. *J. Multivariate Anal.* **58**, 162-181.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20**, 1040-1061.

Ibragimov, I. A. and Solev, V. N. (1969). A condition for regularity of Gaussian stationary processes. *Soviet Math. Dokl.* **10** 371-375.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, K. C. (1997). Nonlinear confounding in high-dimensional regression. *Ann. Statist.* **25**, 577-612.

Nychka, D., Bailey, B., Ellner, S., Haaland, P. and O'Connell M. (1998). FUNFITS: data analysis and statistical tools for estimating functions. Technical Report, Department of Statistics, North Carolina State University.

Ripley, B. D. (1994). *Pattern Recognition and Neural Networks.* Cambridge University Press.

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

Schott, J. R. (1997). *Matrix Analysis for Statistics.* John Wiley, New-York.

Schumaker, L. L. (1981). *Spline Functions.* John Wiley, New York.

Shi, P. D. (1997). M-Type Regression Splines Involving Time Series, *J. Statist. Plann. Inference* **58**, 17-37.

Shi, P. D. and Fung, W. K. (1998). A note on transforming a response variable for linearity. *Biometrika* **85**, 749-754.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22**, 118-171.

Tomišić, V. and Simeon, V. (1993). Assessment of the effective rank of a (co)variance matrix: a non-parametric goodness-of-fit test. *J. Chemometrics* **7**, 381-392.

Velilla, S. (1998). Assessing the number of linear components in a general regression problem. *J. Amer. Statist. Assoc.* **94**, 1088-1098.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences, *Ann. Probab.* **22**, 94-116.

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

University of Hong Kong, Pokfulam Road, Hong Kong.

E-mail: hrntfwk@hkucc.hku.hk

Department of Statistics, University of Illinois, 725 S. Wright, Champaign, IL 61820, U.S.A.

E-mail: x-he@uiuc.edu

Department of Statistics, University of Illinois, 725 S. Wright, Champaign, IL 61820, U.S.A.

Imaging Research Inc., St. Catharines, Ontario, Canada 12S 3A1.

(Received October 2000; accepted April 2002)