L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

Dimension reduction via parametric inverse regression

Efstathia Bura

The George Washington University, Washington, USA

Abstract: In this paper, a linear subspace containing part or all of the information for the regression of a m-vector Y on a p-vector X and its dimension are estimated via the means of inverse regression. Smooth parametric curves are fitted to the p inverse regressions through a multivariate linear model, without imposing any strict assumptions on the error distribution. This method is expected to be more powerful in reducing the dimension of a regression problem when compared to SIR, the estimation procedure proposed by Li (1991), that is based on fitting piecewise constant functions to the inverse regression curves.

Key words: Dimension reduction, regression, linear subspace estimation.

AMS subject classification: 62A99, 62H05.

1 Introduction

Let $Y \in \mathrm{R}^m$ and $X \in \mathrm{R}^p$ with joint cumulative distribution function (c.d.f.) $F(Y,X)$. In a regression setting the behavior of the conditional cumulative distribution function of Y given X, $F(Y|X)$, as the value of X varies in its marginal sample space is under study. As a means of characterizing the regression structure, consider replacing X by $k \leq p$ linear combinations of its components, $\eta_1^T X, \ldots, \eta_k^T X$, without losing information on $F(Y|X)$ so that, for all values of X,

$$F(Y|X) = F(Y|\eta_1^T X, \ldots, \eta_k^T X) = F(Y|\eta^T X) \qquad (1)$$

where η is the $p \times k$ matrix with columns η_j, and $F(\cdot|\cdot)$ denotes the conditional c.d.f. of the first argument given the second. Equation (1) holds trivially when $\eta = I_p$, where I_p denotes the identity matrix of dimension

p, and thus it imposes no restrictions on $F(Y|X)$. It can be expressed equivalently as

$$Y \perp\!\!\!\perp X | \eta^T X \tag{2}$$

where the notation $U \perp\!\!\!\perp V|W$ in (2) means that U is independent of V given any value for W (Dawid, 1979). Both (1) and (2) express the fact that the conditional c.d.f. of $Y|X$ depends on X only through $\eta^T X$, the coordinates of a projection of X onto the k-dimensional linear subspace spanned by the columns of η. Consequently, $\eta^T X$ can be used in place of X without loss of information on the regression.

An example where (2) holds is the additive-error regression model

$$Y|X = g(\eta_1^T X, ..., \eta_k^T X) + \epsilon$$

where $\epsilon \perp\!\!\!\perp X$ and $\mathrm{E}(\epsilon) = 0$.

For any vector or matrix α, let $S(\alpha)$ denote its range space and $\dim(S(\alpha))$ denote its dimension. If (1) holds then it also holds with η replaced by any basis for $S(\eta)$. In this sense, (1) and (2) can be regarded as statements about $S(\eta)$ rather than statements about η per se. Thus, when (2) holds we follow Li (1991, 1992) and call $S(\eta)$ a *dimension-reduction subspace* for $F(Y|X)$ or for the regression of Y on X.

Obviously, the smallest dimension-reduction subspace provides the greatest dimension reduction in the predictor vector. Unfortunately, smallest or *minimum* dimension-reduction subspaces (Cook 1994a) are not always unique. To circumvent the latter, Cook (1994b, 1996) introduced the notion of *central* dimension-reduction subspaces:

Definition 1 A subspace S is a central dimension-reduction subspace for the regression of Y on X if (a) S is a dimension-reduction subspace and (b) $S \subset S_{drs}$ for all dimension-reduction subspaces S_{drs}, i.e. $S = \cap S_{drs}$. A central dimension-reduction subspace will be denoted by $S_{Y|X}(\cdot)$.

The intersection of all dimension-reduction subspaces $\cap S_{drs}$ is trivially a subspace but it is not necessarily a dimension-reduction one. Also, it is easy to see that a central dimension-reduction subspace is a minimum dimension-reduction subspace but the converse is not always true. In fact, there are regression problems for which the central dimension-reduction subspace does not exist. A detailed discussion of these issues can be found in Cook (1994b, 1996).

By definition, a central dimension-reduction subspace, being the intersection of all dimension-reduction subspaces, is unique when it exists. The existence of central subspaces can be assured by placing fairly weak restrictions on aspects of the joint distribution of Y and X (Cook 1994a, 1996). In this paper, we concentrate on regressions where central dimension-reduction spaces exist.

The subspace $S_{Y|X}$ is considered a "super-parameter" that is used to index the conditional distribution of Y given X and its estimation is the main theme of this work. Throughout the rest of this article, the columns of the $p \times k$ matrix η form a basis for the central space $S_{Y|X}$, and k is used to denote its dimension.

1.1 Inverse regression and *SIR*

Methods are available for estimating portions of the central subspace $S_{Y|X}$ if we are willing to place certain conditions on the marginal distribution of the predictors. The method that will be presented in this article is based on inverse regression.

Let $S_{E(X|Y)}$ denote the subspace spanned by $\{\,\mathrm{E}(X|Y) - \mathrm{E}(X) : Y \in \Omega_Y\,\}$, where $\Omega_Y \subset \mathrm{R}^m$ is the marginal sample space of Y. The condition that the marginal distribution of the predictors X must satisfy in order for inverse regression to be useful in estimating a portion of the central subspace is stated in the following theorem. The theorem, as presented by Li (1991), is based on an arbitrary dimension-reduction subspace which need not be central. However, the version here is stated in terms of the central subspace. Throughout this article, boldface capital Latin letters will denote matrices, even though other symbols will also be used for the same purpose provided there is no fear of confusion.

Theorem 1 *Assume that the central subspace $S_{Y|X}(\eta)$ exists for $F(Y|X)$, and that, for all $b \in R^p$, $E(b^T X|\eta^T X)$ is linear in $\eta^T X$. Then the centered inverse regression curve $E(X|Y) - E(X)$ satisfies*

$$E(X|Y) - E(X) \in S(\Sigma_x \eta)$$

Equivalently,

$$S_{E(X|Y)} \subset S(\Sigma_x \eta) = \Sigma_x S_{Y|X}$$

where $\Sigma_x = Cov(X)$.

Proof: Li (1991) proved Theorem 1 for any dimension reduction subspace $S(\eta)$ so that (2) is satisfied. It is obvious that if the theorem holds for an arbitrary dimension reduction subspace, it also holds for the intersectio n of all dimension reduction subspaces; that is, the central dimension reduction subspace $S_{Y|X}$, provided it exists. □

The linearity condition on $\mathrm{E}(b^T X|\eta^T X)$ is required to hold only for the basis η of the central subspace. η being unknown, in practice we may require that it hold for all possible η, which is equivalent to elliptical symmetry of the distribution of X (Eaton, 1986). Li (1991) mentioned that the linearity condition is not a severe restriction, since most low-dimensional projections of a high-dimensional data cloud are close to being normal (Diaconis and Freedman, 1984; and Hall and Li, 1993). In addition, there often exist transformations of the predictors that make them comply with the linearity condition. Cook and Nachtsheim (1994) suggested re-weighting of the predictor vector to make it elliptically contoured.

In the next corollary, which follows directly from Theorem 1, the analogous result is given for a standard random vector. Suppose that $\Sigma_x > 0$ and let Z be the standardized version of X,

$$Z = \Sigma_x^{-1/2}(X - \mathrm{E}(X))$$

Obviously, $\mathrm{E}(Z) = 0$ and $\mathrm{Cov}(Z) = I_p$. Also, since Z is a 1-1 and onto linear transformation of X, $Y \perp\!\!\!\perp X|\eta^T X$ if and only if $Y \perp\!\!\!\perp Z\ |\beta^T Z$, where $\beta = \Sigma_x^{1/2}\eta$ or $\beta_i = \Sigma_x^{1/2}\eta_i$, $i = 1, 2, ..., k$.

Corollary 1

$$E(Z|Y) \in S(\Sigma_x^{1/2}\eta) = S(\beta) = S_{Y|Z}$$

This corollary readily implies that $\mathrm{E}(Z|Y) = \mathbf{P}_\beta \mathrm{E}(Z|Y)$, where $\mathbf{P}_\beta$ is the orthogonal projection operator for $S(\beta)$ with respect to the usual inner product.

Corollary 1 also implies that $S_{E(Z|Y)}$ is a subspace of $S_{Y|Z}$. This does not guarantee equality between $S_{E(Z|Y)}$ and $S_{Y|Z}$, and thus, inference about $S_{E(Z|Y)}$ possibly covers only part of $S_{Y|Z}$. For example, if $Y = Z_1^2$, with Z_1 being the first coordinate variable of Z, and if Z_1 is symmetric about its mean, then $\mathrm{E}(Z|Y) = 0$ even though $S_{Y|Z} = \mathrm{span}((1, 0)^T)$. For a broader discussion of the inability of SIR, and consequently of the method developed in this paper, to diagnose this symmetric dependence see Cook and Weisberg (1991). The missed part of $S_{Y|Z}$ might be recovered from higher order moments of the conditional distribution of Z given Y (Cook and Weisberg 1991; Li 1992), but such issues are not addressed in this article. We assume throughout that $S_{E(Z|Y)}$ is non-trivial, in the sense that it contains non-zero directions, should they exist.

Theorem 1 and Corollary 1 lead to the use of inverse regression as an estimation means of part or possibly the whole of the central dimension-reduction subspace. One such method is SIR (Sliced Inverse Regression), proposed by Li (1991). In SIR, the range of the one-dimensional variable Y is partitioned into a fixed number of slices and the p components of Z are regressed on $\tilde{Y}$, a discrete version of Y resulting from slicing its range, giving p one-dimensional regression problems, instead of the possibly high-dimensional forward regression of Y on Z. Then, a rather crude nonparametric estimate of the inverse curve $\mathrm{E}(X|Y)$ serves to estimate the central dimension-reduction subspace. *SIR* includes an asymptotic test for inferring about d, a lower bound on k.

But, even though Li (1991) introduced an innovative way of reducing the dimension in a regression problem, *SIR* has limitations of which the most important is that *SIR* can be ambiguous about the estimate of the dimension as the latter depends sometimes crucially on the choice of the number of slices. This can be easily avoided by using standard regression estimation techniques.

In this article, smooth parametric curves are fitted to the p inverse regressions in order to estimate the central subspace $S_{Y|X}(\eta)$, without imposing any restrictions on the dimension of the response vector Y.

2 Parametric inverse regression

For simplicity assume that X is standardized to have 0 mean and the identity covariance matrix. To model the conditional expectation of X given Y, a multivariate linear model is fitted with X being the response, $X^T = (x_1, \ldots, x_p)$, and Y, $Y^T = (y_1, \ldots, y_m)$, the explanatory vector. Let

$$\mathrm{E}\begin{bmatrix} x_1 \\ \vdots \; |Y \\ x_p \end{bmatrix}^T = \begin{bmatrix} f_1(Y) & \cdots & f_q(Y) \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix}$$

where the f_i's are arbitrary, R-valued linearly independent known functions of Y. Suppose that a random sample of size n is available on (Y, X). Then, including a matrix of errors $\mathbf{E}_n$, the model becomes

$$\mathbf{X}_n|Y = \mathbf{Z}_n\mathbf{B} + \mathbf{E}_n \tag{3}$$

where $\mathbf{X}_n = (x_{ij})$, a $n \times p$ random matrix, $\mathbf{Z}_n = (f_{il})$, a $n \times q$ fixed matrix with $f_{il} = f_l(Y_i)$, and $\mathbf{B} = (\beta_{lj})$, the $q \times p$ matrix of coefficients. The error matrix $\mathbf{E}_n$ satisfies

$$\mathrm{E}(\mathbf{E}_n|Y) = 0 \quad \text{and} \quad \mathrm{Cov}(\vec{\mathbf{E}}_n|Y) = \Sigma_{x|y} \otimes I_n$$

where $\Sigma_{x|y}$ is a $p \times p$ positive definite, unknown matrix, that does not depend on Y. $\mathbf{X}_n$, $\mathbf{Z}_n$, and $\mathbf{E}_n$ are indexed by the sample size n to indicate their dependence upon it. The symbol $\otimes$ denotes the Kronecker product. Clearly, the rank of $\mathbf{Z}_n$ is q. We assume that $n \geq p$ in order to avoid trivial cases. No distributional assumptions on the errors are made except that the rows $\epsilon_i^{(n)T}$ of the error matrix $\mathbf{E}_n$ are independent with mean 0 and constant covariance matrix $\Sigma_{x|y}$.

According to (3), $S_{E(X|Y)}$ is the linear subspace of $S_{Y|X}$ which is spanned by the rows of $\mathbf{Z}_n\mathbf{B}$; that is $S(\mathbf{B}^T\mathbf{Z}_n^T) = S_{E(X|Y)}$. Therefore, since $\text{rank}(\mathbf{B}^T\mathbf{Z}_n^T) = \text{rank}(\mathbf{Z}_n\mathbf{B})$, $\text{rank}(\mathbf{Z}_n\mathbf{B}) \leq \dim(S_{Y|X})$, and the $\text{rank}(\mathbf{Z}_n\mathbf{B})$ is a lower bound on the dimension of the central dimension-reduction subspace.

But,

$$\text{rank}(\mathbf{Z}_n\mathbf{B}) = \text{rank}(\mathbf{B}^T\mathbf{Z}_n^T\mathbf{Z}_n\mathbf{B}) = \text{rank}(\mathbf{B})$$

since $\mathbf{Z}_n^T\mathbf{Z}_n$ is a positive definite matrix (see [A4.4], Seber, 1977). Thus, the rank of $\mathbf{Z}_n\mathbf{B}$ is actually equal to the rank of $\mathbf{B}$, and hence inference on the dimension of $S_{E(X|Y)}$ can be based solely on $\mathbf{B}$ in the sense that an estimate of the rank of $\mathbf{B}$ constitutes an estimate of a lower bound on the dimension of $S_{Y|X}$.

The estimate of $\mathbf{B}$ to be used for inference on the rank of $\mathbf{B}$ is the ordinary least squares estimate, given by

$$\hat{\mathbf{B}}_n = (\mathbf{Z}_n^T\mathbf{Z}_n)^{-1}\mathbf{Z}_n^T\mathbf{X}_n \tag{4}$$

3 The asymptotic distribution of $\mathbf{B}_n$

Let $e_i^{(n)}$ be the n-vector with 1 in the ith place and zeroes elsewhere. We are interested in the asymptotic distribution of $\sqrt{n}\,(\hat{\mathbf{B}}_n - \mathbf{B})$.

Let $\mathbf{H}_n$ denote the covariance matrix of $\sqrt{n}\,(\hat{\mathbf{B}}_n - \mathbf{B})$,

$$\begin{aligned} \mathbf{H}_n &= \text{Cov}(\overrightarrow{\sqrt{n}\,(\hat{\mathbf{B}}_n - \mathbf{B})}) \\ &= \Sigma_{x|y} \otimes (\mathbf{Z}_n^T\mathbf{Z}_n/n)^{-1} \end{aligned}$$

The notation $\|\cdot\|_{\max}$ identifies the norm on the vector space of matrices defined by

$$\|(a_{ij})\|_{\max} = \max_{i,j} |a_{ij}|$$

for a matrix $= (a_{ij})$. The following lemma about the asymptotic distribution of $\sqrt{n}\,(\hat{\mathbf{B}}_n - \mathbf{B})$ follows readily from Theorem 2.4.3, Bunke and Bunke (1986), and the multivariate version of Slutsky's theorem (see [A 4.19], Bunke and Bunke, 1986).

Lemma 1 *Let* ${}^{>}_{pq}$ *be the space of all* $pq \times pq$ *positive definite matrices and let* $\mathcal{F}$ *be the space of distributions of the errors* $\mathbf{E}_n$*. If*

$$\mathbf{H}_n \longrightarrow \mathbf{H} \in \mathcal{M}^{>}_{pq} \tag{5}$$

then

$$\sqrt{n}\ \overrightarrow{(\hat{\mathbf{B}}_n - \mathbf{B})} \xrightarrow{\mathcal{D}} N_{qp}(0, \mathbf{H}) \tag{6}$$

provided the following three conditions are satisfied

$$\|(\mathbf{Z}_n^T\mathbf{Z}_n)^{-1}\mathbf{Z}_n^T\|_{\max} = o(n^{-1/2}) \qquad (I)$$

$$\sup_{F\in\mathcal{F}} \int_{\|x\|>c} \|x\|^2 dF(x) \longrightarrow 0 \quad as\ c \rightarrow \infty \qquad (II)$$

$$\inf_{\Sigma\in\mathcal{M}(\mathcal{F})} \lambda_{\min}(\Sigma) \geq r > 0 \qquad (III)$$

where

$$\mathcal{M}(\mathcal{F}) = \{ \int_{R^p} xx^T dF(x) : F \in \mathcal{F} \} \subset \mathcal{M}^{>}_{p}$$

The error distributions that are usually considered satisfy Conditions (II) and (III).

Assume that there exists a matrix $\in \mathcal{M}^{>}_{q}$, so that

$$(\mathbf{Z}_n^T\mathbf{Z}_n/n)^{-1} \longrightarrow \mathbf{G} \tag{7}$$

as $n \longrightarrow \infty$. Also, assume that a consistent estimate $\hat{\Sigma}_{x|y}$ is available, as $\Sigma_{x|y}$ is usually unknown. For instance, $\hat{\Sigma}_{x|y}$ can be taken to be the matrix of residuals from the regression of X on Y divided by either n or, $n - \text{rank}(\mathbf{Z}) = n - q$, the denominator choice that makes $\hat{\Sigma}_{x|y}$ unbiased for $\Sigma_{x|y}$ (the proof is omitted). Let

$$\hat{\mathbf{H}}_n = \hat{\Sigma}_{x|y} \otimes (\mathbf{Z}_n^T\mathbf{Z}_n/n)^{-1} \tag{8}$$

Then, if (7) holds,

$$\hat{\mathbf{H}}_n \xrightarrow{P} \mathbf{H} \tag{9}$$

(9) is a direct application of the triangle inequality and the fact that continuous functions of consistent estimates are themselves consistent. These remarks result in the following corollary to Lemma (1).

Corollary 2 *Suppose Conditions (I), (II), and (III) of Lemma 1 hold. Also assume that* $\hat{\Sigma}_{x|y}$ *is a consistent estimate of* $\Sigma_{x|y}$*, and that (7) holds. Then,*

$$\sqrt{n}\ \hat{\mathbf{H}}_n^{-1/2}\overrightarrow{(\hat{\mathbf{B}}_n - \mathbf{B})} \xrightarrow{\mathcal{D}} N(0, I_{pq}) = N(0, I_p \otimes I_q) \tag{10}$$

Proof: Since $\hat{\Sigma}_{x|y}$ is consistent for $\Sigma_{x|y}$ and (7) holds, (10) is the result of a direct application of the multivariate version of Slutsky's theorem (see [A 4.19] in Bunke and Bunke, 1986), and of Lemma 1. □

Let $d = \dim(S_{E(X|Y)})$. We have shown that $d = \operatorname{rank}(\mathbf{B})$ and thus, we can use the least squares estimate of $\mathbf{B}$, to estimate the dimension of $S_{E(X|Y)}$ as follows. Let λ_j, $j = 1, \ldots, \min(q,p)$, be the singular values of $\mathbf{B}$. Then, d is the number of the nonzero singular values of $\mathbf{B}$, and inference about d can be made by testing if

$$\Lambda_d^{(1)} = n \sum_{j=d+1}^{\min(q,p)} \lambda_j^2$$

is equal to zero. We have no direct access to $\Lambda_d^{(1)}$, but by observing that the rank of a matrix is not affected when the matrix is multiplied by a nonsingular matrix, the inference on d can be based on the following test statistic

$$\hat{\Lambda}_d^{(1)} = n \sum_{j=d+1}^{\min(q,p)} \hat{\phi}_j^2 \tag{11}$$

where $\hat{\phi}_j$ are the singular values of

$$(\frac{\mathbf{Z}_n^T\mathbf{Z}_n}{n})^{1/2}\hat{\mathbf{B}}_n\hat{\Sigma}_{x|y}^{-1/2} \tag{12}$$

(12) is used in place of $\hat{\mathbf{B}}_n$ for convenience, as its asymptotic covariance matrix is the identity. Now, the test is based on the asymptotic distribution of $\hat{\Lambda}_d^{(1)}$.

4 The asymptotic distribution of $\hat{\Lambda}_d^{(1)}$

Given the asymptotic normality of the least squares estimate of $\mathbf{B}$, we can obtain the asymptotic distribution of the singular values of a fixed nonsingular transformation of $\mathbf{B}$ based on a result about the asymptotic distribution of the singular values of a matrix by Eaton and Tyler (1994).

Theorem 2 *The asymptotic distribution of $\hat{\Lambda}_d^{(1)}$ defined in (11) and (12) is $\chi^2_{(p-d)\times(q-d)}$.*

Proof: Consider the singular value decomposition of $\mathbf{G}^{-1/2}\mathbf{B}\Sigma_{x|y}^{-1/2}$, where $\mathbf{G}$ is the positive definite limit matrix of $n\ (\mathbf{Z}_n^T\mathbf{Z}_n)^{-1}$,

$$\mathbf{G}^{-1/2}\mathbf{B}\Sigma_{x|y}^{-1/2} = \Gamma_1^T \begin{bmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{bmatrix} \Gamma_2^T$$

$\mathbf{D}$ is a $d\times d$ diagonal matrix with the positive singular values of $\mathbf{G}^{-1/2}\mathbf{B}\Sigma_{x|y}^{-1/2}$ along its diagonal. Partition $\Gamma_1^T = (\Gamma_{11},\Gamma_{12}) \; : q\times q$, $\Gamma_{11} \; : q\times d$, $\Gamma_{12} \; : q\times(q-d)$,

$$\Gamma_2^T = \begin{bmatrix}\Gamma_{21}^T\\ \Gamma_{22}^T\end{bmatrix} \quad : p\times p$$

where $\Gamma_{21}^T \; : \; d\times p$, $\Gamma_{22}^T \; : \; (p-d)\times p$. By the Eaton-Tyler result, the limiting distribution of the smallest $q-d$ singular values of

$$\sqrt{n}\;((\frac{\mathbf{Z}_n^T\mathbf{Z}_n}{n})^{1/2}\hat{\mathbf{B}}_n\hat{\Sigma}_{x|y}^{-1/2})$$

is the same as the limiting distribution of the singular values of the $(q-d)\times(p-d)$ matrix

$$\sqrt{n}\;\mathbf{B}_n = \sqrt{n}\;(\Gamma_{12}^T(\frac{\mathbf{Z}_n^T\mathbf{Z}_n}{n})^{1/2}\hat{\mathbf{B}}_n\hat{\Sigma}_{x|y}^{-1/2}\Gamma_{22})$$

By (10), the asymptotic distribution of $\sqrt{n}\;\vec{\mathbf{B}}_n$ is

$$\sqrt{n}\;\overrightarrow{(\Gamma_{12}^T(\frac{\mathbf{Z}_n^T\mathbf{Z}_n}{n})^{1/2}\hat{\mathbf{B}}_n\hat{\Sigma}_{x|y}^{-1/2}\Gamma_{22})} \xrightarrow{\mathcal{D}} N_{(p-d)(q-d)}(0, I_{p-d}\otimes I_{q-d}) \tag{13}$$

But then, $\hat{\Lambda}_d^{(1)}$ has the same asymptotic distribution as the sum of the squares of the singular values of $\sqrt{n}\;(\Gamma_{12}^T(\frac{\mathbf{Z}_n^T\mathbf{Z}_n}{n})^{1/2}\hat{\mathbf{B}}_n\hat{\Sigma}_{x|y}^{-1/2}\Gamma_{22})$ which is $\chi^2_{(p-d)\times(q-d)}$ by (13). □

Note that the asymptotic test derived above is equivalent to the usual F-test for testing $d=0$ when $p=1$; that is, when we fit q functions of Y on the one-dimensional X and we test the overall validity of the model by testing the hypothesis $\beta_{11}=\beta_{21}=\ldots=\beta_{q1}=0$.

4.1 A summarizing theorem

All the key results discussed and proved in the previous sections are summarized in the following theorem.

Theorem 3 *Assume that* $\mathbf{X}_n|Y = \mathbf{Z}_n\mathbf{B}+\mathbf{E}_n$, *with* $E(\mathbf{E}_n)=0$, $Cov(\vec{\mathbf{E}}_n) = \Sigma_{x|y}\otimes I_n$, *where* $\mathbf{X}_n \; : n\times p$, $\mathbf{Z}_n \; : n\times q$, $\mathbf{B} \; : q\times p$, *with* $rank(\mathbf{Z}_n) = q$. *Let* $\hat{\mathbf{B}}_n = (\mathbf{Z}_n^T\mathbf{Z}_n)^{-1}\mathbf{Z}_n^T\mathbf{X}_n$ *be the ordinary least squares estimate of* $\mathbf{B}$. *Let* $\hat{\Sigma}_{x|y}$ *be a consistent estimate of* $\Sigma_{x|y}$ *and* $\mathbf{G}_n^{-1} = \mathbf{Z}_n^T\mathbf{Z}_n/n$. *Assume that* $\mathbf{G}_n \longrightarrow \mathbf{G}$ *pointwise, where* $\mathbf{G}$ *is a* $q\times q$ *positive definite matrix. Then,*

$$\sqrt{n}\;(\hat{\Sigma}_{x|y}^{-1/2}\otimes\mathbf{G}_n^{-1/2})\overrightarrow{(\hat{\mathbf{B}}_n-\mathbf{B})} \xrightarrow{\mathcal{D}} N_{pq}(0, I_p\otimes I_q)$$

Let $d = rank(\mathbf{B})$. Also, let $\hat{\phi}_j$, $j = 1, \ldots, \min(q,p)$, be the singular values of $\mathbf{G}_n^{-1/2}\hat{\mathbf{B}}_n\hat{\Sigma}_{x|y}^{-1/2}$. Then

$$\hat{\Lambda}_d^{(1)} = n \sum_{j=d+1}^{\min(q,p)} \hat{\phi}_j^2$$

is asymptotically distributed as a $\chi^2_{(q-d)(p-d)}$ random variable.

Proof: Both results are immediate consequences of Corollary 2 and Theorem 2. □

We can use the asymptotic distribution of $\hat{\Lambda}_d^{(1)}$ to estimate the rank d of $\mathbf{B}$, or equivalently the dimension of the subspace $S_{E(X|Y)} \subset S_{Y|X}$, as follows: Fix j with $0 \leq j \leq q$. Compare $\hat{\Lambda}_j^{(1)}$ to the quantiles of a $\chi^2_{(p-j)\times(q-j)}$; if it is bigger, conclude that $d > j$; if not, conclude that $d \leq j$, and repeat the procedure.

5 An example

To illustrate the method, we consider the *Horse Mussel Data*: The data consist of a sample of 201 horse mussel measurements collected in the Marlborough Sounds, which are located off the northeast coast of New Zealand's South Island (Camden, 1989). The response variable is muscle mass M, the edible portion of the mussel, in grams. The quantitative predictors are shell width W , shell length L, in *mm*, and shell mass S in grams. The actual sampling method is unknown, but we assume that the data are i.i.d. observations from the overall mussel population. The $R-$ *code* (Cook and Weisberg, 1994) was used for the computations.

In Figure 1a a scatterplot matrix of the response, shell length, shell width and shell mass is presented. It is evident that the linearity condition needed for SIR to work may be violated. The transformed variables $W^{1/2}$ and $S^{1/4}$ will be used in place of W and S, respectively, so that the linearity condition is satisfied by the regressor variables.

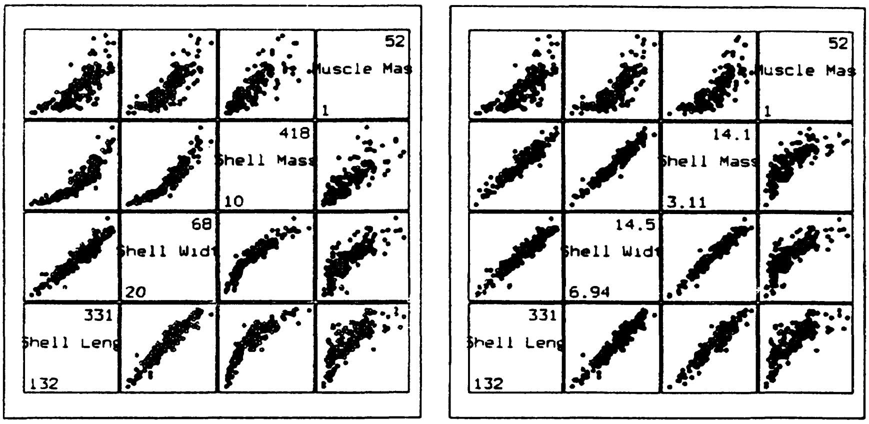
Theorem 1 applies to the transformed data and SIR can be used to estimate the central dimension-reduction subspace. The results of applying SIR to the regression of M on L, $W^{1/2}$ and $S^{1/4}$ are given in Tables 1 and 2; Table 1 contains the results when 5 slices were used and Table 2 when 20 slices were used. The rows of both tables summarize hypothesis tests of the form $d = j$ versus $d > j$. For example, the first row gives the statistic $\hat{\Lambda}_0 = 154.7$ with $(p-d)(H-d-1) = (3-0)(5-1) = 12$ degrees of freedom and a p-value of 0.000. As it can be seen from the two tables, SIR gives contradictory results: it estimates the dimension to be 1 or 2, depending on the number of slices used.

Now, consider fitting smooth parametric curves. The scatterplot matrix in Figure 1b suggests fitting quadratic curves on all three inverse regression plots. The results of the analysis are given in Table 3.

The test indicates a one-dimensional structure supporting that one linear combination of the regressors can be sufficient to characterize the behavior of the conditional c.d.f. of M given L, $W^{1/2}$, $L^{1/4}$:

$$0.0235275L + 0.003176W^{1/2} + 0.00541707S^{1/4} \tag{14}$$

The same conclusion of one-dimensional structure is also reached using regression graphical techniques to estimate the structural dimension of the this regression problem (see Cook and Weisberg, 1994).



a. untransformed predictors b. transformed-predictors

Figure 1: Scatterplots of the Mussel Data

Table 1: SIR results for $H = 5$

j	$\hat{\Lambda}_j$	DF	$p - value$
0	154.7	12	0.000
1	14.81	6	0.022
2	4.973	2	0.083

Table 2: SIR results for $H = 20$

j	$\hat{\Lambda}_j$	DF	$p - value$
0	177.2	45	0.000
1	31.55	28	0.293
2	9.877	13	0.704

6 Discussion

In order to estimate a lower bound on the dimension of the central dimension-reduction subspace $S_{Y|X}$, the conditional expectation of the standardized

j	$\hat{\Lambda}_j^{(1)}$	DF	$p - value$
0	502.9	9	0.000
1	6.998	4	0.136
2	3.8210E-27	1	1.000

Table 3: Parametric results for the Mussel data

X given Y was modeled according to the linear model (3) placing relaxed conditions on the error distribution, namely zero mean and constant covariance structure. The decision on what model to fit is based on data inspection. The select ed model should be a sufficiently complex model that accommodates the data. For example, if polynomials are fitted, the degree should be a number that provides a good fit to all inverse regression curves.

An asymptotic χ^2 test for the dimension d of $S_{E(X|Y)}$ was obtained as a result of the asymptotic normality of the least squares estimate of $\mathbf{B}$. The estimated dimension is in fact an estimate of a lower bound for the dimension of $S_{Y|X}$.

The d eigenvectors of the least squares estimate of $\mathbf{B}$, that correspond to its d largest eigenvalues, multiplied by $\mathbf{Z}_n$ yield estimates of d of the basis vectors of $S_{Y|X}$. They, in turn, can be scaled back to estimates of basis vectors of the central dimension-reduction subspace for the non-standardized X, by multiplication with $\hat{\Sigma}_x^{-1/2}$, where $\hat{\Sigma}_x$ is the moment estimate of Σ_x.

These results can be extended to the non-constant covariance structure model, under certain conditions. In addition, a similar test has been developed for the case where the inverse regression curves are not all of the same shape. This test does not have an asymptotic distribution with quantiles as easy to compute as these of a χ^2. All of the above developments can be found in Bura (1996).

The technique of this article does not suffer from most of the shortcomings of SIR and requires neither the marginal distribution of X to be normal nor Y to be one-dimensional. Further research is needed to assess the sensitivity of the method to ou tliers. The power of the test is also expected to be higher due to the fitting method.

As an aside, it is worthwhile to comment that even though the estimation procedure developed in this paper was motivated by the use of inverse regression as a means to reduce the dimension of a forward regression problem, it is also a method of estimating the linear subspace spanned by a regression curve. In this context, if the linear subspace is estimated to be $\{0\}$ and this cannot be attributed to symmetric dependence (see Cook and Weisberg, 1991), we can possibly infer that the regression curve is intrinsically nonlinear.

References

[1] Bunke, H. and Bunke, O. (Editors) (1986). *Statistical Inference in Linear Models. Statistical Methods of Model Building,* Vol. 1. Chichester: Wiley.

[2] Bura, E. (1996). Dimension reduction via inverse regression. Ph.D. Dissertation, School of Statistics, University of Minnesota.

[3] Cook, R. D. (1994a). On the interpretation of regression plots. *J. Am. Statist. Assoc.* **89**, 177-189.

[4] Cook, R. D. (1994b). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *1994 Proc. of the Section on Physical and Engineering Sciences*, Alexandria, VA: American Statistical Association, pp. 18–25.

[5] Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Am. Statist. Assoc.* **91**, 983–992.

[6] Cook, R. D. and Nachtsheim, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *J. Am. Statist. Assoc.*To appear.

[7] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression" by K. C. Li. *J. Am. Statist. Assoc.* **86**, 328–332.

[8] Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics.* New York: Wiley.

[9] Dawid, A. P. (1979). Conditional independence in statistical theory. *J. R. Statist. Soc.* B **41**, 1–31.

[10] Eaton, Morris L.(1986). A characterization of spherical distributions. *J. Multiv. Anal.* **20**, 272–276.

[11] Eaton, Morris L. and Tyler, D. E. (1994). The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis. *J. Mult. Anal.* **34**, 439–446.

[12] Diaconis, Persi and Freedman, David (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793–815.

[13] Hall, Peter and Li, Ker-Chau (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.* **21**, 867–889.

[14] Li, Ker-Chau (1991). Sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.* **86**, 316–342.

[15] Li, Ker-Chau (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Statist. Assoc.* **87**, 1025–1039.

[16] Seber, G. A. F. (1977). *Linear Regression Analysis.* New York: Wiley.