Estimating the structural dimension of regressions via parametric inverse regression

Efstathia Bura

The George Washington University, Washington, USA

R. Dennis Cook

University of Minnesota, St. Paul, USA

Summary. A new estimation method for the dimension of a regression at the outset of the analysis is proposed. A linear subspace spanned by projections of the regressor vector $\mathbf{X}$, which contains part or all of the modelling information for the regression of a vector $\mathbf{Y}$ on $\mathbf{X}$, and its dimension are estimated via the means of parametric inverse regression. Smooth parametric curves are fitted to the p inverse regressions via a multivariate linear model. No restrictions are placed on the distribution of the regressors. The estimate of the dimension of the regression is based on optimal estimation procedures. A simulation study shows the method to be more powerful than SIR in some situations.

Keywords: Dimension reduction; Inverse regression; SIR; Parametric inverse regression; Asymptotic test for dimension

1. Introduction

Let $\mathbf{Y} \in \mathbb{R}^m$ and $\mathbf{X} \in \mathbb{R}^p$ with joint cumulative distribution function (c.d.f.) $F(\mathbf{Y}, \mathbf{X})$. Regression analyses typically tend to concentrate on the study of the first two moments of the conditional cumulative distribution function of $\mathbf{Y}$ given $\mathbf{X}$, $F(\mathbf{Y}|\mathbf{X})$. In general, though, the goal of regression is the study of the behaviour

Address for correspondence: Efstathia Bura, Department of Statistics, The George Washington University, 2201 G Street NW, Washington, DC 20052, USA.
E-mail: ebura@ gwu.edu

of $F(\mathbf{Y}|\mathbf{X})$, as the value of $\mathbf{X}$ varies in its marginal sample space. As a means of characterizing the regression structure, consider replacing $\mathbf{X}$ by $k \leq p$ linear combinations of its components, $\boldsymbol{\eta}_1^T\mathbf{X}, \ldots, \boldsymbol{\eta}_k^T\mathbf{X}$, without losing information on $F(\mathbf{Y}|\mathbf{X})$ so that, for all values of $\mathbf{X}$,

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

where $\boldsymbol{\eta}$ is the $p \times k$ matrix with columns $\boldsymbol{\eta}_j$. The notation $U \perp\!\!\!\perp V|W$ in (1) means that U is independent of V given any value for W (Dawid 1979). Expression (1), as a mathematical formulation of the dependence of $\mathbf{Y}$ on $\mathbf{X}$, was introduced by Cook (1994a). It expresses the fact that the conditional c.d.f. of $\mathbf{Y}|\mathbf{X}$ depends on $\mathbf{X}$ only through $\boldsymbol{\eta}^T\mathbf{X}$; the coordinates of a projection of $\mathbf{X}$ onto the k-dimensional linear subspace spanned by the columns of $\boldsymbol{\eta}$. Consequently, $\boldsymbol{\eta}^T\mathbf{X}$ contains equivalent or *sufficient*, in the statistical sense, information for the regression of $\mathbf{Y}$ on $\mathbf{X}$. Most importantly, if $k < p$, then sufficient reduction in the dimension of the regression is achieved, which in turn leads to *sufficient summary plots* of $\mathbf{Y}$ versus $\boldsymbol{\eta}^T\mathbf{X}$ as graphical displays of all the necessary modelling information for the regression of $\mathbf{Y}$ on $\mathbf{X}$. Subsequently, sufficient summary plots can guide the selection of appropriate models for $F(\mathbf{Y}|\mathbf{X})$.

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

Obviously, the smallest dimension-reduction subspace provides the greatest dimension reduction in the predictor vector. There are several ways to define such a subspace. In this article we use the *central dimension-reduction subspace*, denoted by $S_{\mathbf{Y}|\mathbf{X}}$ (Cook 1994b, 1996, 1998a,b). $S_{\mathbf{Y}|\mathbf{X}}$ is the intersection of all dimension-reduction subspaces for $F(\mathbf{Y}|\mathbf{X})$ and is trivially a subspace but is not necessarily a dimension-reduction one. The existence of central subspaces can be assured by placing fairly weak restrictions on aspects of the joint distribution of $\mathbf{Y}$ and $\mathbf{X}$

(Cook, 1994a, 1996). Throughout this article, we focus on regressions for which central dimension-reduction spaces exist.

The subspace $S_{\mathbf{Y}|\mathbf{X}} = S_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{\eta})$ is in effect a "meta-parameter" that is used to index the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$. The columns of the $p \times k$ matrix $\boldsymbol{\eta}$ will denote a basis for the central subspace $S_{\mathbf{Y}|\mathbf{X}}$, and k will be used to denote its dimension or the *structural dimension of the regression of* $\mathbf{Y}$ *on* $\mathbf{X}$ (Cook and Weisberg, 1994). Our main objective is the estimation of $S_{\mathbf{Y}|\mathbf{X}}$.

The paper is organized as follows: Existing dimension estimation methods, with emphasis on SIR (Li, 1991), are reviewed in Section 2. The example in Section 2 serves to illustrate both the application of SIR and its limitations. The proposed estimation method, namely parametric inverse regression or PIR, is introduced and described in Section 3. Section 4 contains its extension to the non-constant variance case. The algorithm describing the PIR dimension reduction procedure is presented in Section 5. In Section 6, PIR is applied to the example of Section 2 and a simulation study to compare the power of the two testing methods for dimension is carried out. A concluding discussion is presented in Section 7. The lengthier proofs are given in the appendix.

2. Background: Inverse Regression and SIR

Methods are available for estimating portions of the central subspace $S_{\mathbf{Y}|\mathbf{X}}$, provided certain conditions are placed on the marginal distribution of the predictors.

Let $S_{\mathrm{E}(\mathbf{X}|\mathbf{Y})}$ denote the subspace spanned by $\{\, \mathrm{E}(\mathbf{X}|\mathbf{Y}) - \mathrm{E}(\mathbf{X}) : \mathbf{Y} \in \Omega_{\mathbf{Y}} \,\}$, where $\Omega_{\mathbf{Y}} \subset \mathbb{R}^m$ is the marginal sample space of $\mathbf{Y}$. Given (1), assume that the marginal distribution of the predictors $\mathbf{X}$ satisfy the following condition, which henceforth will be referred to as the *linearity condition*: for all $\mathbf{b} \in \mathbb{R}^p$, $\mathrm{E}(\mathbf{b}^T\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ is linear in $\boldsymbol{\eta}^T\mathbf{X}$.

Under this linearity condition on the regressor distribution, Li (1991, thm. 3.1) showed that the centered inverse regression curve $\mathrm{E}(\mathbf{X}|\mathbf{Y}) - \mathrm{E}(\mathbf{X})$ satisfies

$$\mathrm{E}(\mathbf{X}|\mathbf{Y}) - \mathrm{E}(\mathbf{X}) \in S(\boldsymbol{\Sigma}_x\boldsymbol{\eta})$$

Equivalently,

$$S_{\mathrm{E}(\mathbf{X}|\mathbf{Y})} \subset S(\mathbf{\Sigma}_x \boldsymbol{\eta}) = \mathbf{\Sigma}_x S_{\mathbf{Y}|\mathbf{X}} \tag{2}$$

where $\mathbf{\Sigma}_x = \mathrm{Cov}(\mathbf{X})$.

The linearity condition on $\mathrm{E}(\mathbf{b}^T\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ is required to hold only for the basis $\boldsymbol{\eta}$ of the central subspace. Since $\boldsymbol{\eta}$ is unknown, in practice we may require that it hold for all possible $\boldsymbol{\eta}$, which is equivalent to elliptical symmetry of the distribution of $\mathbf{X}$ (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; 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.

Suppose that $\mathbf{\Sigma}_x > 0$ and let $\mathbf{Z}$ be the standardised version of $\mathbf{X}$,

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

Obviously, $\mathrm{E}(\mathbf{Z}) = 0$ and $\mathrm{Cov}(\mathbf{Z}) = \mathbf{I}_p$. The observable sample version $\hat{\mathbf{Z}}$ is constructed by replacing $\mathbf{\Sigma}_x$ and $\mathrm{E}(\mathbf{X})$ with their usual moment estimates. Also, since $\mathbf{Z}$ is a 1-1 and onto linear transformation of $\mathbf{X}$, $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^T\mathbf{X}$ if and only if $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}\,|\boldsymbol{\beta}^T\mathbf{Z}$, where $\boldsymbol{\beta} = \mathbf{\Sigma}_x^{1/2}\boldsymbol{\eta}$ or $\boldsymbol{\beta}_i = \mathbf{\Sigma}_x^{1/2}\boldsymbol{\eta}_i$, $i = 1, 2, ..., k$. By (2), we obtain that

$$\mathrm{E}(\mathbf{Z}|\mathbf{Y}) \in S(\mathbf{\Sigma}_x^{1/2}\boldsymbol{\eta}) = S(\boldsymbol{\beta}) = S_{\mathbf{Y}|\mathbf{Z}} \tag{3}$$

The containment relation in (3) readily implies that $\mathrm{E}(\mathbf{Z}|\mathbf{Y}) = \mathbf{P}_{\boldsymbol{\beta}}\mathrm{E}(\mathbf{Z}|\mathbf{Y})$, where $\mathbf{P}_{\boldsymbol{\beta}}$ is the orthogonal projection operator for $S(\boldsymbol{\beta})$ with respect to the usual inner product. It also implies that $S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$ is a subspace of $S_{\mathbf{Y}|\mathbf{Z}}$. This does not guarantee equality between $S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$ and $S_{\mathbf{Y}|\mathbf{Z}}$ and, thus, inference about $S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$ possibly covers only part of $S_{\mathbf{Y}|\mathbf{Z}}$. The missed part of $S_{\mathbf{Y}|\mathbf{Z}}$ might be recovered from higher order moments of the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$ (Cook, 1998b; Cook and Weisberg, 1991; Li, 1991, 1992), but such issues are not addressed in this article.

We assume throughout that $S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$ is non-trivial, in the sense that it contains non-zero directions, should such exist.

Both (2) and (3) lead to the use of inverse regression as an estimation tool for a fraction of or the entire central dimension-reduction subspace. A popular 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 $\mathbf{Z}$ 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 $\mathbf{Z}$. Then, a very simple nonparametric estimate of the inverse regression curve $\mathrm{E}(\mathbf{Z}|Y)$ serves to estimate the central dimension-reduction subspace via estimating $\mathrm{Cov}(\mathrm{E}(\mathbf{Z}|Y))$. This can be based on the fact that $S\{\mathrm{Cov}(\mathrm{E}(\mathbf{Z}|Y))\} = S_{\mathrm{E}(\mathbf{Z}|Y)}$ except on a set of measure zero (e.g., see Cook, 1998a, Prop. 11.1, Eaton, 1983, Prop. 2.7). The SIR estimate of $\mathrm{Cov}(\mathrm{E}(\mathbf{Z}|Y))$ is given by

$$\widehat{\mathrm{Cov}}(\mathrm{E}(\mathbf{Z}|Y)) = \sum_{h=1}^{H} \hat{p}_h \hat{m}_h \hat{m}_h'$$

where H is the fixed number of slices, $\hat{p}_h = n_h/n$ with n being the total sample size and n_h the number of observations in the hth slice, and $\hat{m}_h$ is the p-vector of the average of $\hat{\mathbf{Z}}$ within slice h. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$ be the ordered eigenvalues of $\widehat{\mathrm{Cov}}(\mathrm{E}(\mathbf{Z}|Y))$. Li (1991) proved that if $d = \dim(S_{\mathrm{E}(\mathbf{X}|Y)})$, the statistic

$$L_d = n \sum_{j=d+1}^{p} \hat{\lambda}_j \tag{4}$$

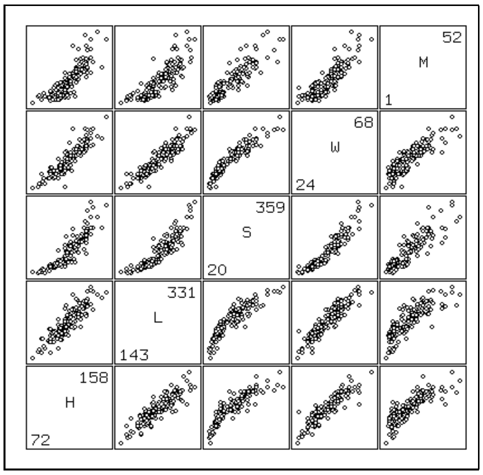
has an asymptotic chi-square distribution with $(p-d)(H-d-1)$ degrees of freedom, provided the regressors are normal. The test statistic can be used to estimate the dimension of $S_{\mathrm{E}(\mathbf{X}|Y)}$ by performing tests of $d = j$ versus $d \geq j+1$, $j = 0, \ldots, p-1$.

Other testing techniques based on inverse regression that use the same simple nonparametric estimation method as Li (1991) have been developed. Schott (1994) proposed a test which requires elliptically symmetric regressors, and for which the tuning constant is the number of observations c per slice as opposed to the number of slices H in Li (1991). In order to obtain the asymptotic distribution of his test statistic, Schott lets c go to infinity. Velilla (1998) introduced another testing method, which does not impose restrictions on the regressor distribution, where c is fixed and the number of slices H varies.

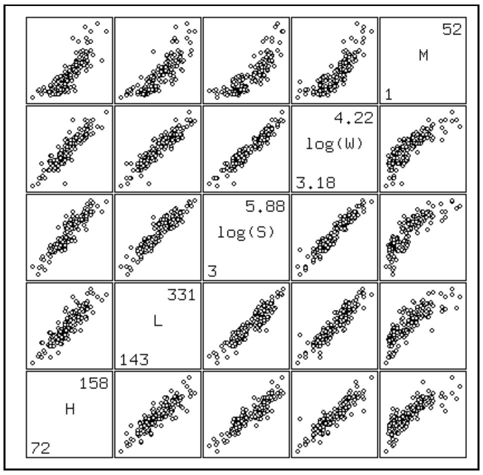
SIR is a simple and useful technique for reducing the dimension in a regression problem, but nevertheless, it has limitations. Normality of the regressor vector $\mathbf{X}$ is required for the chi-square asymptotic test for dimension to apply (Li, 1991). Requiring normality for the predictors was proved not to be necessary for the asymptotic result to hold in Bura and Cook (1999) and Cook (1998a), where it is shown that restrictions should be placed on the conditional covariance structure of the standardised version of $\mathbf{X}$ instead. These restrictions are trivially satisfied if $\mathbf{X}$ has a multivariate normal distribution, but they also contradict Li's (1991) claim that the asymptotic distribution of L_d does not depend on the constant variance assumption of the conditional distribution of $\mathbf{X}$ given Y. Most importantly, SIR can be ambiguous about the estimate of dimension as the latter depends sometimes crucially on the choice of the number of slices. As a result, all methods that depend on a tuning constant related to number of slices choice suffer from the same ambiguity in estimation (Schott, 1994; Velilla, 1998).

To illustrate some of the issues discussed above, we consider the *Horse Mussel Data*: The data consist of a sample of 172 horse mussel measurements collected in the Marlborough Sounds, which are located off the northeast coast of New Zealand's South Island (Camden, 1989; Cook and Weisberg, 1994; Cook, 1998a). The response variable is muscle mass M, the edible portion of the mussel, in g. The quantitative predictors are shell width W in mm, shell height L in mm , shell length L in mm, and shell mass S in g. The actual sampling method is unknown, but we assume that the data are i.i.d. observations from the overall mussel population. The regression software package *Arc* (Cook and Weisberg, 1999) was used for the computations.

A scatterplot matrix of the response, shell height, shell length, shell width and shell mass is presented in Figure 1a. It is evident that the linearity condition needed for SIR to work may be violated. The transformed variables $\log(W)$ and



a. untransformed predictors



b. transformed-predictors

Figure 1: Scatterplots of the Mussel Data

$\log(S)$ will be used in place of W and S, respectively, so that the linearity condition is satisfied by the regressor variables.

The results of applying SIR to the regression of M on H, L, $\log(W)$ and $\log(S)$ are given in Tables 1 and 2; the first contains the results when 6 slices were used and the second when 15 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 $L_0 = 156.68$ with $(p-d)(H-d-1) = (4-0)(6-1) = 20$ degrees of freedom and a p-value of 0.000. As we can see from the two tables SIR gives contradictory results: it estimates the dimension to be 1 or 2, depending on the number of slices used.

Table 1: SIR results for $H = 6$

j	L_j	DF	$p - value$
0	156.68	20	0.000
1	22.992	12	0.028
2	9.0924	6	0.168

Table 2: SIR results for $H = 15$

j	L_j	DF	$p - value$
0	173.55	48	0.000
1	35.477	33	0.352
2	15.614	20	0.740

3. Parametric Inverse Regression

The proposed new dimension reduction method of parametric inverse regression (PIR) fits smooth parametric curves on the p inverse regressions via a multivariate linear model. No distributional restrictions are imposed on the regressor vector. To model the conditional expectation of $\mathbf{Z}$, the standardised version of the regressor vector $\mathbf{X}$, given $\mathbf{Y}$, a multivariate linear model is fitted with $\mathbf{Z} = (z_1, \ldots, z_p)^T$ being the response, and $\mathbf{Y} = (y_1, \ldots, y_m)^T$ the explanatory vector. Let

$$\mathrm{E}\begin{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_p \end{bmatrix} |\mathbf{Y} \end{bmatrix}^T = \begin{bmatrix} f_1(\mathbf{Y}) & \cdots & f_q(\mathbf{Y}) \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix}$$

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

$$\mathbf{Z}_n | \mathcal{Y}_n = \mathbf{F}_n \mathbf{B} + \mathbf{E}_n \tag{5}$$

where $\mathbf{Z}_n = (z_{ij}) = (\mathbf{X}_n - \mathrm{E}(\mathbf{X}_n))\mathbf{\Sigma}_x^{-1/2}$, an $n \times p$ random matrix, $\mathbf{F}_n = (\tilde{f}_{il})$, an $n \times q$ fixed matrix with $\tilde{f}_{il} = \tilde{f}_l(\mathbf{Y}_i) = f_{il} - \sum_{i=1}^n f_{il}/n$ being the centered version of $f_{il} = f_l(\mathbf{Y}_i)$, and $\mathbf{B} = (\beta_{lj})$, the $q \times p$ matrix of coefficients. Centering is used so that the model is consistent with the fact that the expectation of the column averages of $\mathbf{Z}_n$ equals 0. The error matrix $\mathbf{E}_n$ satisfies

$$\mathrm{E}(\mathbf{E}_n | \mathcal{Y}_n) = 0 \quad , \quad \mathrm{Cov}(\mathrm{vec}(\mathbf{E}_n) | \mathcal{Y}_n) = \mathbf{\Sigma}_{z|y} \otimes \mathbf{I}_n \tag{6}$$

where $\mathbf{\Sigma}_{z|y}$ is a $p \times p$ positive definite, unknown matrix, that does not depend on $\mathbf{Y}$, and $\mathrm{vec}(\mathbf{E}_n)$ is the vector produced by concatenating the columns of the error matrix $\mathbf{E}_n$. The symbol $\otimes$ denotes the Kronecker product. Clearly, the rank of $\mathbf{F}_n$ is q. We assume that $n \geq \max(p, q)$ in order to avoid trivial cases. No distributional assumptions on the errors are made except that the rows of the error matrix $\mathbf{E}_n$ are conditionally independent with mean 0 and constant covariance matrix $\mathbf{\Sigma}_{z|y}$, given all the responses $\mathcal{Y}_n$.

According to (5), $S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$ is the linear subspace of $S_{\mathbf{Y}|\mathbf{Z}}$ which is spanned by the rows of $\mathbf{F}_n\mathbf{B}$; that is $S(\mathbf{B}^T\mathbf{F}_n^T) = S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$. Therefore, since $\mathrm{rank}(\mathbf{B}^T\mathbf{F}_n^T) = \mathrm{rank}(\mathbf{F}_n\mathbf{B})$, $\mathrm{rank}(\mathbf{F}_n\mathbf{B}) \leq \dim(S_{\mathbf{Y}|\mathbf{Z}})$, yielding that the $\mathrm{rank}(\mathbf{F}_n\mathbf{B})$ is a lower bound on the dimension of the central dimension-reduction subspace.

Observe that $\mathrm{rank}(\mathbf{F}_n\mathbf{B}) = \mathrm{rank}(\mathbf{B}^T\mathbf{F}_n^T\mathbf{F}_n\mathbf{B}) = \mathrm{rank}(\mathbf{B})$ since $\mathbf{F}_n^T\mathbf{F}_n$ is a positive definite matrix (see *A4.4*, Seber (1977)). In consequence, inference on the dimension of $S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$ 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 $S_{\mathbf{Y}|\mathbf{Z}}$.

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

$$\hat{\mathbf{B}}_n = (\mathbf{F}_n^T\mathbf{F}_n)^{-1}\mathbf{F}_n^T\mathbf{Z}_n \tag{7}$$

Since $\mathbf{Z}_n$ is unobservable, $\hat{\mathbf{B}}_n$ is unobservable and thus in practice it is necessary to use the sample version of $\mathbf{Z}_n$ when computing $\hat{\mathbf{B}}_n$. However, it is sufficient to work in terms of $\mathbf{Z}_n$ for the purpose of deriving the asymptotic distribution of the test statistic Λ_d described in Theorem 1 for the rank of $\mathbf{B}$. To find the asymptotic distribution of Λ_d, we first need to find the asymptotic distribution of a standardised version of $\hat{\mathbf{B}}_n$.

Let $\mathbf{H}_n$ denote the covariance matrix of $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{B}}_n - \mathbf{B})$; that is, $\mathbf{H}_n = \mathbf{\Sigma}_{z|y} \otimes (\mathbf{F}_n^T\mathbf{F}_n/n)^{-1}$. If $\mathbf{H}_n$ has a positive definite limit matrix $\mathbf{H}$, then

$$\sqrt{n}\ \mathrm{vec}(\hat{\mathbf{B}}_n - \mathbf{B}) \xrightarrow{\mathcal{D}} N_{qp}(0, \mathbf{H}) \tag{8}$$

provided certain conditions are satisfied (see lemma 1 in the appendix).

Assume there exists a $q \times q$ positive definite matrix $\mathbf{G}$ so that

$$(\mathbf{F}_n^T\mathbf{F}_n/n)^{-1} \xrightarrow[n\to\infty]{} \mathbf{G} \tag{9}$$

Also, assume that a consistent estimate $\hat{\mathbf{\Sigma}}_{z|y}$ is available, as $\mathbf{\Sigma}_{z|y}$ is usually unknown. For example, $\hat{\mathbf{\Sigma}}_{z|y}$ can be taken to be $(n-q)^{-1}(\mathbf{Z}_n - \mathbf{F}_n\hat{\mathbf{B}}_n)^T(\mathbf{Z}_n - \mathbf{F}_n\hat{\mathbf{B}}_n)$, which is also unbiased for $\mathbf{\Sigma}_{z|y}$. Let

$$\hat{\mathbf{H}}_n = \hat{\mathbf{\Sigma}}_{z|y} \otimes (\mathbf{F}_n^T \mathbf{F}_n / n)^{-1} \tag{10}$$

Then, if (9) holds,

$$\hat{\mathbf{H}}_n \xrightarrow[n\to\infty]{} \mathbf{H} \qquad \text{in probability} \tag{11}$$

The convergence in (11) is a direct application of the triangle inequality and the fact that continuous functions of consistent estimates are themselves consistent. The remarks above in conjunction with direct application of the multivariate version of Slutsky's theorem (see [A 4.19] in Bunke and Bunke (1986)), and (8) obtain the following:

$$\sqrt{n}\, \hat{\mathbf{H}}_n^{-1/2} \operatorname{vec}(\hat{\mathbf{B}}_n - \mathbf{B}) \xrightarrow{\mathcal{D}} N(0, \mathbf{I}_{pq}) = N(0, \mathbf{I}_p \otimes \mathbf{I}_q) \tag{12}$$

Let $d = \dim(S_{E(\mathbf{Z}|\mathbf{Y})})$ and $\mathbf{G}_n^{-1} = \mathbf{F}_n^T \mathbf{F}_n / n$. We have shown that $d = \operatorname{rank}(\mathbf{B})$ and thus, since $\operatorname{rank}(\mathbf{B}) = \operatorname{rank}(\mathbf{G}^{-1/2} \mathbf{B} \mathbf{\Sigma}_{z|y}^{-1/2})$, we use the standardised matrix

$$\hat{\mathbf{B}}_{std} = \mathbf{G}_n^{-1/2} \tilde{\mathbf{B}}_n \hat{\mathbf{\Sigma}}_{z|y}^{-1/2}$$

to estimate d where

$$\tilde{\mathbf{B}}_n = (\mathbf{F}_n^T \mathbf{F}_n)^{-1} \mathbf{F}_n^T \hat{\mathbf{Z}}_n \tag{13}$$

is now the estimate of $\mathbf{B}$ based on the sample version $\hat{\mathbf{Z}}_n = (\mathbf{X}_n - \bar{\mathbf{X}}_n) \hat{\mathbf{\Sigma}}_x^{-1/2}$ of the standardised predictor matrix $\mathbf{Z}_n$, where $\bar{\mathbf{X}}_n$ and $\hat{\mathbf{\Sigma}}_x$ are the usual sample moment estimates of $\mathrm{E}(\mathbf{X}_n)$ and $\mathbf{\Sigma}_x$, respectively. We give a test statistic (Λ_d) for $d = \operatorname{rank}(\mathbf{B}) = \dim(S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})})$ in Theorem 1. The proof is given in the appendix.

Theorem 1. Assume that model (5) holds, that $\mathbf{G}_n$ converges pointwise to a positive definite limit, and that $\hat{\mathbf{\Sigma}}_{z|y}$ is a consistent estimate of $\mathbf{\Sigma}_{z|y}$. Let $\hat{\phi}_j$, $j = 1, \ldots, \min(q,p)$, be the singular values of $\hat{\mathbf{B}}_{std}$. Then

$$\Lambda_d = n \sum_{j=d+1}^{\min(q,p)} \hat{\phi}_j^2 \tag{14}$$

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

We use Λ_d as a test statistic for the rank of $\mathbf{B}$. For example, to test the hypothesis that $d = 1$ compare Λ_1 to the percentage points of a chi-squared distribution with $(q-1)(p-1)$ degrees of freedom. As an aside, it is easy to see that the asymptotic test in Theorem 1 coincides with the usual F-test for testing $d = 0$; that is, that all the coefficients are zero, when $p = 1$.

4. The Non-Constant Covariance Case

In Sections 2 and 3 the parametric models that were used assumed that the covariance structure of the error matrix given $\mathbf{Y}$ was constant; that is, independent of $\mathbf{Y}$. Occasionally, this assumption may be seriously violated resulting in dimension estimation errors. In this section the non-constant error covariance structure case is addressed.

We assume that the regression model (5) holds but now $\mathrm{Cov}(\mathbf{Z}|\mathbf{Y})$ is a function of $\mathbf{Y}$:

$$\mathrm{Cov}(\mathbf{Z}|\mathbf{Y}) = \mathbf{\Sigma}_{z|y}(\mathbf{Y}) = (\sigma_{ij}(\mathbf{Y}))_{i,j=1}^{p}$$

In consequence, the covariance structure of the error matrix $\mathbf{E}_n$ can no longer be represented by the Kronecker product of $\mathbf{\Sigma}_{z|y}$ and the identity $\mathbf{I}_n$, for

$$\mathrm{Cov}(Z_{ki}, Z_{kj}|\mathbf{Y} = \mathbf{Y}_k) = \sigma_{ij}(\mathbf{Y}_k)$$

for $k = 1, \ldots, n$, $i, j = 1, \ldots, p$. Hence, the covariance matrix of $\mathrm{vec}(\mathbf{Z}_n|\mathcal{Y}_n)$, and consequently of $\mathrm{vec}(\mathbf{E}_n)$, is a $np \times np$ symmetric matrix consisting of p^2 blocks of order $n \times n$, where the ijth block is the diagonal $n \times n$ matrix with $\sigma_{ij}(\mathbf{Y}_1), \ldots, \sigma_{ij}(\mathbf{Y}_n)$ along its main diagonal for $i, j = 1, \ldots, p$.

In vector form, $\hat{\mathbf{B}}_n$ can be written as $\mathrm{vec}(\hat{\mathbf{B}}_n) = \mathrm{vec}(\mathbf{W}_n\mathbf{Z}_n) = (\mathbf{I}_p \otimes \mathbf{W}_n)\,\mathrm{vec}(\mathbf{Z}_n)$, where $\mathbf{W}_n = (\mathbf{F}_n^T\mathbf{F}_n)^{-1}\mathbf{F}_n^T$ is a $q \times n$ known matrix of weights. The covariance matrix of $\mathrm{vec}(\hat{\mathbf{B}}_n)$ equals $(\mathbf{I}_p \otimes \mathbf{W}_n)\,\mathrm{Cov}(\mathrm{vec}(\mathbf{Z}_n))(\mathbf{I}_p \otimes \mathbf{W}_n^T)$. Thus,

$\text{Cov}(\text{vec}(\hat{\mathbf{B}}_n))$ is a $qp \times qp$ block matrix, whose ijth block is given by

$$\mathbf{W}_n \ \text{diag}(\sigma_{ij}(\mathbf{Y}_1), \dots, \sigma_{ij}(\mathbf{Y}_n)) \ \mathbf{W}_n^T \tag{15}$$

for $i, j = 1, \dots, p$, and $\mathbf{H}_n = \text{Cov}[n^{1/2} \ \text{vec}\,(\mathbf{W}_n \mathbf{Z}_n - \mathbf{B})]$ is a $qp \times qp$ block matrix, whose ijth block is given by (15) multiplied by n.

Assuming that $\mathbf{H}_n$ has a positive definite limit $\mathbf{H}$, we readily obtain

$$\sqrt{n} \ \text{vec}\,(\mathbf{W}_n \mathbf{Z}_n - \mathbf{B}) \xrightarrow{\mathcal{D}} N_{pq}(0, \mathbf{H}) \tag{16}$$

provided conditions (I), (II), and (III) of lemma 1 in the appendix are satisfied. Additionally, (16) also holds if the conditional covariance of X_i and X_j given $\mathbf{Y}$ is asymptotically constant for all $i, j = 1, \dots, p$ (see lemma 2 in the appendix).

Let $\hat{\mathbf{\Sigma}}_{z|y}(\mathbf{Y}) = (\hat{\sigma}_{ij}(\mathbf{Y}))$ be a consistent estimate of $\mathbf{\Sigma}_{z|y}(\mathbf{Y}) = (\sigma_{ij}(\mathbf{Y}))$, for all $i, j = 1, \dots, p$ and all $\mathbf{Y} \in \Omega_{\mathbf{Y}}$. Let $\hat{\mathbf{H}}_n$ be the $qp \times qp$ matrix whose ijth block is given by (15) multiplied by n, with $\hat{\sigma}_{ij}(\mathbf{Y}_k)$ in place of $\sigma_{ij}(\mathbf{Y}_k)$.

Since $\hat{\mathbf{H}}_n$ is nonsingular, if it were also consistent for $\mathbf{H}$, then by the multivariate version of Slutsky's theorem we would obtain

$$\sqrt{n}\, \hat{\mathbf{H}}_n^{-1/2} \ \text{vec}\,(\mathbf{W}_n \mathbf{Z}_n - \mathbf{B}) \xrightarrow{\mathcal{D}} N_{pq}(0, \mathbf{I}_{pq} = \mathbf{I}_p \otimes \mathbf{I}_q)$$

The dimension d is not affected by this nonsingular transformation. Weak consistency, in the sense of $\hat{\mathbf{H}}_n \to \mathbf{H}$ in probability is easy to establish given a weakly consistent estimate of $\mathbf{\Sigma}_{z|y}(\mathbf{Y})$. Moreover, by placing further conditions on the entries of $\hat{\mathbf{\Sigma}}_{z|y}(\mathbf{Y})$ we obtain that $\hat{\mathbf{H}}_n$ is L^2-consistent for $\mathbf{H}$ (see lemma 3 in the appendix).

Let $\text{vec}(\hat{\mathbf{B}}_{\text{std}}) = \hat{\mathbf{H}}_n^{-1/2}\, \text{vec}(\mathbf{W}_n \hat{\mathbf{Z}}_n)$. The $q \times p$ matrix $\hat{\mathbf{B}}_{\text{std}}$ that results from the arrangement of the vector $\text{vec}(\hat{\mathbf{B}}_{\text{std}})$ into p columns, satisfies $\text{rank}(\hat{\mathbf{B}}_{\text{std}}) = \text{rank}(\mathbf{W}_n \hat{\mathbf{Z}}_n)$, since $\hat{\mathbf{H}}_n^{-1/2}$ is nonsingular. Also, let

$$\Lambda_d = n \sum_{j=d+1}^{\min(p,q)} \hat{\phi}_j^2 \tag{17}$$

where $\hat{\phi}_j$, $j = 1, \dots, \min(p, q)$, denote the ordered singular values of $\hat{\mathbf{B}}_{\text{std}}$. The following theorem states the conditions under which the asymptotic distribution of

Λ_d is chi-squared.

Theorem 2. Assume that conditions (I), (II) and (III) of lemma 1 are satisfied and that $\hat{\mathbf{H}}_n$ is consistent or L^2 consistent for a positive definite matrix $\mathbf{H}$. If $d = \text{rank}(\mathbf{B}) = \dim(S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})})$, then Λ_d as defined in (17) is asymptotically distributed as a $\chi^2_{(p-d)(q-d)}$ random variable.

Proof. The proof is analogous to the proof of Theorem 1 in Section 3.

The inferential procedure on d is the same as in the constant covariance case, provided $\mathbf{\Sigma}_{z|y}(\mathbf{Y})$ can be estimated consistently. The computation of a consistent estimate of $\mathbf{\Sigma}_{z|y}$ is presented next.

Assume that $\mathrm{E}(\mathbf{Z}_j^2) < \infty$, for all $j = 1, \ldots, p$. Let

$$\begin{aligned} \hat{\sigma}_{ij}(\mathbf{Y}) &= \widehat{\mathrm{Cov}}_n(\mathbf{Z}_i, \mathbf{Z}_j|\mathbf{Y}) \\ &= \hat{\mathrm{E}}_n(\mathbf{Z}_i\mathbf{Z}_j|\mathbf{Y}) - \hat{\mathrm{E}}_n(\mathbf{Z}_i|\mathbf{Y})\hat{\mathrm{E}}_n(\mathbf{Z}_j|\mathbf{Y}) \end{aligned} \tag{18}$$

for $i,\ j \in \{1, 2, \ldots, p\}$, where $\hat{\mathrm{E}}_n(\cdot|\mathbf{Y})$ denotes the least squares estimate of $\mathrm{E}(\cdot|\mathbf{Y})$ from regressing the argument in $(\cdot)$ on $\mathbf{Y}$. The choice of the regression model to be fitted on the argument in $(\cdot)$ is guided by the data. Under well known conditions, least squares estimates are consistent and thus,

$$\hat{\sigma}_{ij}(\mathbf{Y}) \to \sigma_{ij}(\mathbf{Y}) \tag{19}$$

in probability, for all $i,\ j \in \{1, 2, \ldots, p\}$, and all $\mathbf{Y} \in \Omega_{\mathbf{Y}}$.

Let $\hat{\mathbf{\Sigma}}_{z|y}(\mathbf{Y}_k)$ be the $p \times p$ matrix with entries given by (18), computed at $\mathbf{Y}_k$ for $k = 1, \ldots, n$. Then, (19) implies that $\hat{\mathbf{\Sigma}}_{z|y}(\mathbf{Y}_k)$ is a consistent estimate of $\mathbf{\Sigma}_{z|y}(\mathbf{Y}_k)$, for all $k = 1, \ldots, n$.

To obtain L^2 consistency for the covariance matrix estimate $\hat{\mathbf{H}}_n$, we should also require that $\hat{\sigma}_{ij}(\mathbf{Y}_k)$ be L^2 consistent for $\sigma_{ij}(\mathbf{Y}_k)$, for all $i, j = 1, \ldots, p$, $k = 1, \ldots, n$, and that $\mathrm{Cov}(\hat{\sigma}_{ij}(\mathbf{Y}_k), \hat{\sigma}_{ij}(\mathbf{Y}_l)) \xrightarrow[n\to\infty]{} 0$, for all $k, l = 1, \ldots, n$, $k \neq l$ (see lemma 3).

5. Parametric Inverse Regression Algorithm

We can use the asymptotic distribution of Λ_d to estimate the rank d of $\mathbf{B}$, or equivalently the dimension of the subspace $S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})} \subset S_{\mathbf{Y}|\mathbf{Z}}$, as follows: Start with $j = 0$. To test the hypothesis that $d = j$,

1. By visual inspection of the scatterplot matrix, decide what function(s) of $\mathbf{Y}$ fit the data the best. The choice of function(s) can be facilitated by dynamically overlaying curves to the scatterplots such as, for example, polynomials of different degrees. The latter can be easily implemented in software that supports dynamic graphics. Form the centered incidence matrix $\mathbf{F}_n$ for the p inverse regressions, and compute its rank q.

2. Standardise the regressor vector by letting $\hat{\mathbf{Z}}_n = (\mathbf{X}_n - \bar{\mathbf{X}}_n)\hat{\mathbf{\Sigma}}_x^{-1/2}$, where $\bar{\mathbf{X}}_n = \mathbf{1}_n \bar{\mathbf{X}}^T$ and $\hat{\mathbf{\Sigma}}_x$ is the moment estimate of the covariance matrix of $\mathbf{X}$.

3. Obtain the least squares matrix of estimated coefficients $\tilde{\mathbf{B}}_n$ from the p inverse regressions of the $\hat{Z}_i$'s, $i = 1, 2, \ldots, p$, on $\mathbf{Y}$.

4. By visual inspection of the scatterplot matrix, decide whether the constant covariance assumption holds. If it does,

   a. Let $\hat{\mathbf{\Sigma}}_{z|y}$ be the matrix of residuals from the regression of $\mathbf{Z}$ on $\mathbf{Y}$ divided by $n - q$, and $\mathbf{G}_n = (\mathbf{F}_n^T \mathbf{F}_n / n)^{-1}$.

   b. Compute the standardised least squares matrix of coefficients: $\hat{\mathbf{B}}_{\mathrm{std}} = \mathbf{G}_n^{-1/2} \tilde{\mathbf{B}}_n \hat{\mathbf{\Sigma}}_{z|y}^{-1/2}$.

   If it does not,

   a. Use least squares to estimate $\hat{\sigma}_{ij}(\mathbf{Y}_k)$ as described in (18), for $k = 1, \ldots, n$.

   b. Let $\mathbf{W}_n = (\mathbf{F}_n^T \mathbf{F}_n)^{-1} \mathbf{F}_n^T$. Form the covariance matrix $\mathbf{H}_n$, as a $qp \times qp$ matrix, whose ij th block is given by (15) multiplied by n.

   c. Compute the standardised matrix of coefficients $\hat{\mathbf{B}}_{\mathrm{std}}$ as described in the paragraph preceding (17).

5. Compute the ordered singular values $\hat{\phi}_j$ of $\hat{\mathbf{B}}_{std}$ and construct the test statistic Λ_d.

6. Compare Λ_j to the quantiles of a $\chi^2_{(p-j)\times(q-j)}$; if it is smaller conclude that $d = j$; if it is bigger, conclude that $d > j$, set $j = j + 1$ and repeat the procedure.

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

6. Applications and Simulations

6.1 The Horse Mussel Data Revisited

To illustrate the PIR dimension reduction method, we reconsider the Horse Mussel data. The scatterplot matrix in Figure 1b visually suggests fitting quadratic curves on all three inverse regression plots. The quadratic fit is also supported by overlaying polynomials of degree 2 on the scatterplots of the regressors versus the response. The results of the analysis are given in Table 3.

Table 3: Results for the Mussel Data

j	Λ_j	DF	$p-value$
0	701.82	8	0.000
1	5.8895	3	0.117

The test indicates a one-dimensional structure. Thus, one linear combination of the regressors is sufficient to characterize the behavior of the conditional c.d.f. of M given H, L, $\log(S)$, $\log(W)$:

$$0.028H - 0.029L - 0.593\log(S) + 0.804\log(W) \tag{20}$$

The coefficients of the transformed regressors in (20) are the elements of the eigenvector corresponding to the largest eigenvalue from the eigenanalysis in step 6 of the PIR algorithm of Section 5.

6.2 Power Comparison of SIR and PIR

This section contains a simulation based comparison of the power of the two dimension estimation methods SIR and PIR. For all regression models to be considered in this section, three sample sizes are used: $n = 50, 100, 250$. For each sample size and each distribution of the regressor vector $\mathbf{X}$, the p-values corresponding to the test statistics for selected dimensions for both tests were collected over 1,000 replications. For the PIR method, only polynomials in Y were fitted to the inverse regressions. The degree of the polynomial was decided upon by visual inspection of the scatterplots of a few initial simulated data, in conjunction with the interactive superimposition of curves of different degrees. The latter can be easily carried out in *Arc* (Cook and Weisberg, 1999).

The first model to be studied has structural dimension 1. The response Y is generated according to the following model.

$$y = x_1 + x_2 + x_4 + 0.5\epsilon \tag{21}$$

where ϵ is a standard normal variate. Two distributions are considered for the regressor vector $\mathbf{X}$: (a) $\mathbf{X} \sim N(0, \mathbf{I}_4)$, and (b) $\mathbf{X} \sim$ Pearson type II with parameters $m = -0.5$ and $\mathbf{\Sigma} = \mathbf{I}_4$ (Johnson, 1987). Pearson type II belongs to the family of elliptically contoured distributions and therefore satisfies the linearity condition necessary for the theory to apply.

The numerical entries of the rows of Tables 4 and 5 corresponding to the test statistics indexed by 0 are empirical estimates of the power of the corresponding test. They represent the proportion of times the null of dimension 0 is rejected, when the nominal significance level is 0.05. The numbers in parentheses are the analogous proportions for a significance level of 0.01. The entries of the rows of the same tables corresponding to test statistics indexed by 1 are empirical estimates of

Table 4: Empirical Power and Size of SIR applied to (21)

$n = 50$

	Normal $\mathbf{X}$			Pearson II $\mathbf{X}$		
H	5	10	15	5	10	15
L_0	1.0 (1.0)	1.0 (.975)	.964 (.640)	.993 (.968)	.941 (.687)	.756 (.350)
L_1	.053 (.009)	.036 (.008)	.025 (.003)	.063 (.010)	.041 (.006)	.031 (.004)

$n = 100$

	Normal $\mathbf{X}$			Pearson II $\mathbf{X}$		
H	5	10	15	5	10	15
L_0	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.999 (.999)	.997 (.996)	1.0 (.993)
L_1	.053 (.009)	.053 (.008)	.036 (.007)	.068 (.013)	.052 (.005)	.045 (.009)

$n = 250$

	Normal $\mathbf{X}$			Pearson II $\mathbf{X}$		
H	5	15	30	5	15	30
L_0	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (.999)	.999 (.999)
L_1	.053 (.011)	.042 (.005)	.047 (.008)	.064 (.015)	.053 (.008)	.055 (.014)

the size of the test. The missing entries in Table 5 correspond to fitting a first degree polynomial, thus allowing only the test of $d = 0$ versus $d = 1$ because $q = 1$. The row entries of the tables throughout this section are to be interpreted in the same or analogous way. The symbol H stands for number of slices, and degree stands for the degree of the fitted polynomial.

Apparently, the SIR and PIR tests are roughly equally powerful for both normal and Pearson II distributed X's, with PIR performing uniformly slightly better. The estimated power entries also seem to indicate that the performance of SIR deteriorates as the number of slices increases, especially when $\mathbf{X}$ is Pearson II and the sample size is small. PIR's performance, on the other hand, is consistently more powerful even when we overfit the inverse mean functions as exhibited in Table 5 where polynomials of degree 2, 3 and 4 were fitted (for normal $\mathbf{X}$).

Next, we consider the following model of structural dimension 2:

$$y = x_1(x_2 + x_4 + 1) + 0.5\epsilon \tag{22}$$

Table 5: Empirical Power and Size of PIR for model (21)

$n = 50$						
	Normal **X**				Pearson II **X**	
degree	1	2	3	4	1	2
Λ_0	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (.999)	.999 (.999)
Λ_1		.087 (.022)	.025 (.007)	.061 (.022)		.037 (.013)
$n = 100$						
	Normal **X**				Pearson II **X**	
degree	1	2	3	4	1	2
Λ_0	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
Λ_1		.068 (.017)	.025 (.005)	.024 (.003)		.033 (0.003)
$n = 250$						
	Normal **X**				Pearson II **X**	
degree	1	2	3	4	1	2
Λ_0	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
Λ_1		.052 (.005)	.059 (.014)	0.061 (.016)		.02 (0)

where ϵ is a standard normal variate. The regressor vector $\mathbf{X}$ is assumed to have a 4-variate standard normal distribution.

From Table 6 we can see that SIR would fail to detect dimension 2 with probability as high as 0.94 (when $n = 50$ and $H = 15$). For SIR to give correct results, the sample size would have to be significantly increased. In our simulations, SIR performs well when $n = 250$ and the number of slices is small. On the other hand, PIR is significantly more powerful for all sample sizes, even for a sample size of 50, and both degrees 3 and 4.

The last model we study is based on model C in Velilla (1998, p. 1094). The regressor vector $\mathbf{X}^T = (X_1, \ldots, X_5)$ was generated so that $X_1 = W_1$, $X_2 = V_1 + W_2/2$, $X_3 = -V_1 + W_2/2$, $X_4 = V_2 + V_3$, and $X_5 = V_2 - V_3$. The only restriction placed on $\mathbf{V}$ and $\mathbf{W}$ is that they be independent. Here, V_1, V_2, V_4 are i.i.d. $t_{(4)}$ random variables, $V_3 \sim t_{(3)}$, $V_5 \sim t_{(5)}$, and W_1, W_2 are i.i.d. Gamma(0.25)

Table 6: Empirical Power and Size for SIR applied to (22)

	$n=50$			$n=100$		
H	5	10	15	5	10	15
L_0	.677 (.486)	.589 (.319)	.403 (.161)	.951 (.885)	.919 (.819)	.879 (.743)
L_1	.163 (.049)	.123 (.028)	.062 (.009)	.444 (.244)	.357 (.164)	.305 (.121)
L_2	.005 (.001)	.010 (0)	.005 (0)	.018 (.004)	.023 (.004)	.013 (.003)

	$n=250$			
H	5	10	15	30
L_0	1.0 (.999)	1.0 (1.0)	1.0 (.998)	.998 (.99)
L_1	.934 (.832)	.922 (.821)	.844 (.697)	.705 (.477)
L_2	.038 (.004)	.036 (.001)	.038 (.005)	.026 (.002)

Table 7: Empirical Power and Size of PIR applied to model (22)

	$n=50$		$n=100$		$n=250$	
degree	3	4	3	4	3	4
Λ_0	.972 (.935)	.974 (.938)	.998 (.994)	.998 (.993)	1.0 (1.0)	1.0 (1.0)
Λ_1	.575 (.388)	.577 (.392)	.810 (.668)	.797 (.643)	.988 (.969)	.985 (.966)
Λ_2	.032 (.006)	.024 (.007)	.047 (.008)	.046 (.009)	0.057 (.008)	.049 (.007)

random variables. The dependent variable y was generated according to the model:

$$y = (4 + x_1)(2 + x_2 + x_3) + 0.5\epsilon \tag{23}$$

where ϵ is a standard normal variate.

Model (23) has structural dimension 2. The regressor distribution satisfies the linearity condition by construction (see Velilla (1998), p. 1092-93), even though it does not have an elliptically contoured distribution.

Table 8 indicates that SIR estimates the structural dimension to be 1, across sample sizes and choice of number of slices. On the other hand, from Table 9 we can see that PIR is performing much better with power ranging from 56.2%, when $n = 50$, the level is 0.01 and a cubic is fitted, to 90.4% when $n = 250$, the level is 0.05 and a fourth degree polynomial is used.

Table 8: Empirical Power and Size of SIR applied to (23)

	$n = 50$			$n = 100$		
H	5	10	15	5	10	15
L_0	.996 (.996)	.951 (.718)	.831 (.5)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
L_1	.039 (.009)	.075 (.014)	.085 (.019)	.032 (.008)	.079 (.026)	.119 (.035)
L_2	.001 (0)	.002 (0)	.007 (.001)	.001 (0)	0 (0)	.007 (0)

	$n = 250$			
H	5	10	15	30
L_0	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
L_1	.049 (.014)	.13 (.064)	.164 (.071)	.183 (.9)
L_2	.002 (0)	0.004 (0)	.007 (0)	.011 (.003)

Table 9: Empirical Power and Size of PIR applied to (23)

	$n = 50$		$n = 100$		$n = 250$	
degree	3	4	3	4	3	4
Λ_0	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
Λ_1	.726 (.562)	.808 (.668)	.786 (.664)	.84 (.732)	.871 (.78)	.904 (.836)
Λ_2	.113 (.021)	.182 (.053)	.087 (.02)	.149 (.051)	.092 (.027)	.137 (.048)

All power calculations performed for the models presented in this section, as well as a number of power calculations for other models and/or other regressor distributions not reported in this article, indicate that PIR is at least as powerful as SIR for models with normal or normal-like regressor distributions. More importantly, the simulation study shows that PIR, along with its non-constant variance version, is significantly more powerful when the models are more "complex", and/or the regressor distributions are far from normal, as in the case of model (23). Nevertheless, our simulations also tend to indicate that the non-constant PIR method should be employed only when the violation of the constant variance assumption is serious as it requires yet an additional estimation step.

7. Discussion

We presented the new method of parametric inverse regression (PIR) for dimension reduction in a regression context. PIR fits linear models to the inverse regressions of $\mathbf{X}$ on $\mathbf{Y}$. An asymptotic χ^2 test for the dimension d of $S_{\mathrm{E}(\mathbf{X}|\mathbf{Y})}$ is obtained as a result of the asymptotic normality of the least squares estimate of $\mathbf{B}$. The estimated dimension is an estimate of a lower bound for the dimension of $S_{\mathbf{Y}|\mathbf{X}}$.

The results were also extended to the non-constant covariance structure case, under certain conditions. The technique of PIR does not impose any restrictions on the distribution of $\mathbf{X}$. Even though it requires choosing the parametric function to be fitted, this is guided by and based upon the observed data. In contrast, SIR and its variants (Schott, 1994; Velilla, 1998) require the subjective a priori choice of a tuning constant, such as number of slices, which is not suggested by the observed data but is rather arbitrarily selected. In consequence, PIR does not suffer from the ambiguity of the dimension estimation of the aforementioned methods.

The power of the PIR chi-square test is shown to be higher than the SIR chi-square test's via a simulation study. This is not surprising as (a) we are fitting continuous curves based on all the data, instead of grouping the data in slices, and (b) the fitting method is least squares; a method yielding estimates with optimal properties. As we confined the study to only fitting polynomials, we would expect even higher power gain by including other parametric curves that appear to reflect the data pattern more accurately. Nevertheless, PIR is desirable when marginal plots of the predictors versus the response give a good idea about a functional form for the inverse mean functions. When there is doubt about these mean functions, SIR may be the better choice.

PIR can be easily implemented as the computer code is distributed freely. The code can be downloaded from *http://gwis2.circ.gwu.edu/˜ebura/publications.html.* It requires obtaining *Arc* first (Cook and Weisberg, 1999) from *www.stat.umn.edu/arc.*

Acknowledgements

The authors wish to thank the joint editor, the associate editor and two referees for their careful and constructive comments. Their suggestions helped improve the article.

Appendix

Lemma 1. Let $\mathcal{M}_{pq}^{>}$ 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

$$\mathbf{H}_n \xrightarrow[n\to\infty]{} \mathbf{H} \in \mathcal{M}_{pq}^{>} \tag{24}$$

then

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

provided the following three conditions are satisfied

$$\|(\mathbf{F}_n^T\mathbf{F}_n)^{-1}\mathbf{F}_n^T\|_{\max} = o(n^{-1/2}) \qquad (I)$$
$$\sup_{F\in\mathcal{F}} \int_{\|x\|>c} \|x\|^2 dF(x) \longrightarrow 0 \quad as\ \ c \to \infty \qquad (II)$$
$$\inf_{\mathbf{\Sigma}\in\mathcal{M}(\mathcal{F})} \lambda_{\min}(\mathbf{\Sigma}) \geq r > 0 \qquad (III)$$

where $\mathcal{M}(\mathcal{F}) = \{\int_{R^p} xx^T dF(x) : F \in \mathcal{F}\} \subset \mathcal{M}_p^{>}$. The 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 $\mathbf{A} = (a_{ij})$. Also, $\lambda_{\min}(\mathbf{\Sigma})$ is the smallest eigenvalue of $\mathbf{\Sigma}$ and r is some positive real. The error distributions that are usually considered satisfy Conditions (II) and (III).

Proof: The lemma 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)).

Proof of theorem 1

Consider the singular value decomposition of $\mathbf{G}^{-1/2}\mathbf{B}\mathbf{\Sigma}_{z|y}^{-1/2}$, where $\mathbf{G}$ is the positive definite limit matrix of $n\ (\mathbf{F}_n^T\mathbf{F}_n)^{-1}$,

$$\mathbf{G}^{-1/2}\mathbf{B}\mathbf{\Sigma}_{z|y}^{-1/2} = \mathbf{\Gamma}_1^T \begin{bmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{\Gamma}_2^T$$

$\mathbf{D}$ is a $d \times d$ diagonal matrix with the positive singular values of $\mathbf{G}^{-1/2}\mathbf{B}\mathbf{\Sigma}_{z|y}^{-1/2}$ along its diagonal. Partition $\mathbf{\Gamma}_1^T = (\mathbf{\Gamma}_{11}, \mathbf{\Gamma}_{12})\ : q \times q$, $\mathbf{\Gamma}_{11}\ : q \times d$, $\mathbf{\Gamma}_{12}\ : q \times (q-d)$, $\mathbf{\Gamma}_2^T = (\mathbf{\Gamma}_{21}, \mathbf{\Gamma}_{22})^T\ : p \times p$, where $\mathbf{\Gamma}_{21}^T\ :\ d \times p$, $\mathbf{\Gamma}_{22}^T\ :\ (p-d) \times p$. By the Eaton-Tyler result (Eaton and Tyler, 1994) about the asymptotic distribution of the singular values of a matrix, the limiting distribution of the smallest $\min(q-d, p-d)$ singular values of

$$\sqrt{n}\ (\mathbf{G}_n^{-1/2}\hat{\mathbf{B}}_n\hat{\boldsymbol{\Sigma}}_{z|y}^{-1/2})$$

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

$$\sqrt{n}\ \mathbf{B}_n = \sqrt{n}\ (\boldsymbol{\Gamma}_{12}^T\mathbf{G}_n^{-1/2}\hat{\mathbf{B}}_n\hat{\boldsymbol{\Sigma}}_{z|y}^{-1/2}\boldsymbol{\Gamma}_{22})$$

By (12), we have

$$\sqrt{n}\ \ \mathrm{vec}\,(\boldsymbol{\Gamma}_{12}^T\mathbf{G}_n^{-1/2}\hat{\mathbf{B}}_n\hat{\boldsymbol{\Sigma}}_{z|y}^{-1/2}\boldsymbol{\Gamma}_{22}) \xrightarrow{\mathcal{D}} N_{(p-d)(q-d)}(0, \mathbf{I}_{p-d}\otimes\mathbf{I}_{q-d}) \tag{25}$$

where $\hat{\mathbf{B}}_n$ is the OLSE of $\mathbf{B}$ given in (7). Observe that since $\mathbf{F}_n^T\mathbf{1}_n = 0$, $\tilde{\mathbf{B}}_n = \hat{\mathbf{B}}_n\boldsymbol{\Sigma}_x^{1/2}\hat{\boldsymbol{\Sigma}}_x^{-1/2}$ and thus

$$\sqrt{n}(\tilde{\mathbf{B}}_n - \mathbf{B}) = \sqrt{n}(\hat{\mathbf{B}}_n - \mathbf{B})\boldsymbol{\Sigma}_x^{1/2}\hat{\boldsymbol{\Sigma}}_x^{-1/2} + \sqrt{n}\mathbf{B}(\boldsymbol{\Sigma}_x^{1/2}\hat{\boldsymbol{\Sigma}}_x^{-1/2} - \mathbf{I}) \tag{26}$$

Then,

$$\begin{aligned}
\sqrt{n}\,\mathrm{vec}(\boldsymbol{\Gamma}_{12}^T\mathbf{G}_n^{-1/2}\tilde{\mathbf{B}}_n\hat{\boldsymbol{\Sigma}}_{z|y}^{-1/2}\boldsymbol{\Gamma}_{22}) &= \sqrt{n}\,\mathrm{vec}(\boldsymbol{\Gamma}_{12}^T\mathbf{G}_n^{-1/2}(\tilde{\mathbf{B}}_n - \mathbf{B})\hat{\boldsymbol{\Sigma}}_{z|y}^{-1/2}\boldsymbol{\Gamma}_{22}) + o_p(1)\\
&= \sqrt{n}\,\mathrm{vec}(\boldsymbol{\Gamma}_{12}^T\mathbf{G}_n^{-1/2}(\hat{\mathbf{B}}_n - \mathbf{B})\boldsymbol{\Sigma}_x^{-1/2}\hat{\boldsymbol{\Sigma}}_x^{-1/2}\hat{\boldsymbol{\Sigma}}_{z|y}^{-1/2}\boldsymbol{\Gamma}_{22})\\
&\quad +\sqrt{n}\,\mathrm{vec}(\boldsymbol{\Gamma}_{12}^T\mathbf{G}_n^{-1/2}\mathbf{B}(\Sigma_x^{-1/2}\hat{\boldsymbol{\Sigma}}_x^{-1/2} - \mathbf{I})\hat{\boldsymbol{\Sigma}}_{z|y}^{-1/2}\boldsymbol{\Gamma}_{22})\\
&\quad +o_p(1)
\end{aligned}$$

The first equality follows from substituting $(\tilde{\mathbf{B}}_n - \mathbf{B} + \mathbf{B})$ for $\tilde{\mathbf{B}}_n$, expanding and using the fact that $\mathbf{B}\boldsymbol{\Sigma}_{z|y}^{-1/2}\boldsymbol{\Gamma}_{22} = 0$. The second equality follows by substituting (26). Then the first term on the right of the second equality is going to the distribution indicated in (25). The second term is going to 0 by Slutsky's theorem because $\boldsymbol{\Gamma}_{12}^T\mathbf{G}_n^{-1/2}\mathbf{B} \to \boldsymbol{\Gamma}_{12}^T\mathbf{G}^{-1/2}\mathbf{B} = 0$. Consequently, Λ_d has the same asymptotic distribution as the sum of the squares of the singular values of $\sqrt{n}\ (\boldsymbol{\Gamma}_{12}^T\mathbf{G}_n^{-1/2}\hat{\mathbf{B}}_n\hat{\boldsymbol{\Sigma}}_{z|y}^{-1/2}\boldsymbol{\Gamma}_{22})$ which is $\chi^2_{(p-d)\times(q-d)}$ by (25).

Lemma 2. Suppose that $\boldsymbol{\Sigma}(\mathbf{Y}_n) \to \boldsymbol{\Sigma}$, as $n\to\infty$, for all $\mathbf{Y}_n$ in the $\mathbf{Y}$-sample space, where $\boldsymbol{\Sigma}$ is non-singular. If $n\mathbf{W}_n\mathbf{W}_n^T = (\mathbf{F}_n^T\mathbf{F}_n/n)^{-1}$ has a positive definite limit matrix $\mathbf{G} = (g_{lm})$, then $\mathbf{H}_n$ has a positive definite limit matrix $\mathbf{H}$, and

$$\sqrt{n}\ \ \mathrm{vec}\,(\mathbf{W}_n\mathbf{Z}_n - \mathbf{B}) \xrightarrow{\mathcal{D}} N_{pq}(0, \mathbf{H})$$

provided

$$\|\mathbf{W}_n\|_{\max} = \|(\mathbf{F}_n^T\mathbf{F}_n)^{-1}\mathbf{F}_n^T\|_{\max} = o(n^{-1/2}) \tag{27}$$

and Conditions (II) and (III) of Lemma 1 hold.

Proof: Since $\sigma_{ij}(\mathbf{Y}_k) \to \sigma_{ij}$ as $k \to \infty$, we have that for all $\epsilon > 0$, there exists k_0 such that for all $k \geq k_0$,

$$\sigma_{ij} - \epsilon < \sigma_{ij}(\mathbf{Y}_k) < \sigma_{ij} + \epsilon$$

Without loss of generality, we can assume that $\sum_{k=1}^{\infty} nW_{lk}W_{mk} = g_{lm} > 0$ (the development is analogous for the case $g_{lm} < 0$. When $g_{lm} = 0$, or when $\sigma_{ij} = 0$ the limit is 0). Then, for a sufficiently large $k_1 \geq k_0$ we have

$$(\sigma_{ij} - \epsilon) \sum_{k \geq k_1} nW_{lk}W_{mk} < \sum_{k \geq k_1} \sigma_{ij}(\mathbf{Y}_k) nW_{lk}W_{mk} < (\sigma_{ij} + \epsilon) \sum_{k \geq k_1} nW_{lk}W_{mk}$$

There are two cases: (i) $\sigma_{ij} > 0$ and (ii) $\sigma_{ij} < 0$. Case (i) is equivalent to $\sigma_{ij} - \epsilon > 0$ and similarly (ii) is equivalent to $\sigma_{ij} + \epsilon < 0$, for sufficiently small ϵ. Therefore, for case (i) and for all $\epsilon > 0$ with $\min(g_{lm}, \sigma_{ij}) > \epsilon$ there exists $k_2 \geq k_1$ such that

$$(\sigma_{ij} - \epsilon)(g_{lm} - \epsilon) \leq \sum_{k \geq k_2} \sigma_{ij}(\mathbf{Y}_k) nW_{lk}W_{mk} \leq (\sigma_{ij} + \epsilon)(g_{lm} + \epsilon) \tag{28}$$

or, equivalently (since all products of ϵ are negligible) $\sum_k^n \sigma_{ij}(\mathbf{Y}_k) nW_{lk}W_{mk} \longrightarrow \sigma_{ij} g_{lm}$, as $n \to \infty$. Case (ii) is analogous to case (i) with the inequalities in (28) reversed. Therefore, the ijth block of $\mathbf{H}_n$ satisfies $n\mathbf{W}_n \mathrm{diag}(\sigma_{ij}(\mathbf{Y}_1), \ldots, \sigma_{ij}(\mathbf{Y}_n))\mathbf{W}_n^T \to \sigma_{ij}\mathbf{G}$, which yields that $\mathbf{H}_n \to \mathbf{\Sigma} \otimes \mathbf{G} = \mathbf{H}$. Since the Kronecker product of two positive definite matrices is positive definite (see Harville 1997, ex. 7, p. 369), $\mathbf{H}$ is positive definite. By a direct application of Slutsky's theorem ([A 4.19], Bunke and Bunke 1986), we obtain that

$$\mathbf{H}_n^{-1/2} \ \mathrm{vec}\,(\mathbf{W}_n\mathbf{Z}_n - \mathbf{B}) \xrightarrow{\mathcal{D}} N_{pq}(0, \mathbf{I}_{pq} = \mathbf{I}_p \otimes \mathbf{I}_q)$$

if and only if

$$n^{1/2} \ \mathrm{vec}\,(\mathbf{W}_n\mathbf{Z}_n - \mathbf{B}) \xrightarrow{\mathcal{D}} N_{pq}(0, \mathbf{H}) \tag{29}$$

Now, a sufficient condition for (29) is (27) (see Bura 1996, Chapter 5).

Lemma 3. Suppose that

$$n\,\mathbf{W}_n\mathbf{W}_n^T \xrightarrow[n\to\infty]{} \mathbf{G} \in \mathcal{M}_q^{>} \tag{30}$$

Suppose that $\hat{\sigma}_{ij}(\mathbf{Y})$ converges to $\sigma_{ij}(\mathbf{Y})$ in quadratic mean, for all $i, j = 1, \ldots, p$, and all $\mathbf{Y}$ in the relevant sample space. Also, suppose that $\mathrm{Cov}(\hat{\sigma}_{ij}(\mathbf{Y}_k), \hat{\sigma}_{ij}(\mathbf{Y}_l)) \longrightarrow 0$ as $n \to \infty$, for $k, l = 1, \ldots, n$, $k \neq l$. Then, $\hat{\mathbf{H}}_n$ is a L^2-consistent estimate of $\mathbf{H}$.

Proof: Since $\mathbf{H}$ is the limit matrix of $\mathbf{H}_n$, it suffices to show that

$$\hat{\mathbf{H}}_n - \mathbf{H}_n \xrightarrow[n\to\infty]{} 0 \quad \text{in } L^2 \tag{31}$$

Then, from the triangle inequality, it follows that $\hat{\mathbf{H}}_n$ is a L^2-consistent estimate of $\mathbf{H}$. Consider the ijth blocks of $\hat{\mathbf{H}}_n$ and $\mathbf{H}_n$. The lmth entry of the ijth block of $\mathbf{H}_n$ is

$$\sum_k^n \sigma_{ij}(\mathbf{Y}_k)(nW_{lk}W_{mk}) \tag{32}$$

and of $\hat{\mathbf{H}}_n$ is given by (32) with $\hat{\sigma}_{ij}$ in place of σ_{ij}, for all $m, n = 1, \ldots, q$, and all $i, j = 1, \ldots, p$. Then, (31) is true if and only if

$$\sum_k^n \hat{\sigma}_{ij}(\mathbf{Y}_k)(nW_{lk}W_{mk}) - \sum_k^n \sigma_{ij}(\mathbf{Y}_k)(nW_{lk}W_{mk}) \xrightarrow[n\to\infty]{} 0$$

in L^2, for all $m, n = 1, \ldots, q$, and $i, j = 1, \ldots, p$. Now,

$$\begin{aligned}
&\mathrm{E}[\sum_{k=1}^n \hat{\sigma}_{ij}(\mathbf{Y}_k)(nW_{lk}W_{mk}) - \sum_{k=1}^n \sigma_{ij}(\mathbf{Y}_k)(nW_{lk}W_{mk})]^2 \\
&\quad = \mathrm{E}[\sum_{k=1}^n (\hat{\sigma}_{ij}(\mathbf{Y}_k) - \sigma_{ij}(\mathbf{Y}_k))^2 (nW_{lk}W_{mk})^2] \\
&\quad\quad + \mathrm{E}[\sum_{k=1}^n \sum_{\substack{r=1 \\ r\neq k}}^n (\hat{\sigma}_{ij}(\mathbf{Y}_k) - \sigma_{ij}(\mathbf{Y}_k))(\hat{\sigma}_{ij}(\mathbf{Y}_r) - \sigma_{ij}(\mathbf{Y}_r)) \\
&\quad\quad\quad \times (nW_{lk}W_{mk})(nW_{lr}W_{mr})] \\
&\quad = \sum_{k=1}^n \mathrm{E}[(\hat{\sigma}_{ij}(\mathbf{Y}_k) - \sigma_{ij}(\mathbf{Y}_k))^2](nW_{lk}W_{mk})^2] \qquad (33) \\
&\quad\quad + \sum_{k=1}^n \sum_{\substack{r=1 \\ r\neq k}}^n \mathrm{E}[(\hat{\sigma}_{ij}(\mathbf{Y}_k) - \sigma_{ij}(\mathbf{Y}_k))(\hat{\sigma}_{ij}(\mathbf{Y}_r) - \sigma_{ij}(\mathbf{Y}_r)) \\
&\quad\quad\quad \times (nW_{lk}W_{mk})(nW_{lr}W_{mr})] \qquad (34)
\end{aligned}$$

The integration can be brought inside the sum by the bounded convergence theorem (see Billingsley 1986, p. 214), since (30) holds by assumption and $\hat{\sigma}_{ij}(Y_k)$ is consistent in quadratic mean for $\sigma_{ij}(\mathbf{Y}_k)$, therefore $\hat{\sigma}_{ij}(\mathbf{Y}_k) - \sigma_{ij}(\mathbf{Y}_k)$ is L^2 bounded, for all $k = 1, \ldots, n$. But then, we also have that (33) vanishes by the L^2 consistency of

$\hat{\sigma}_{ij}(\mathbf{Y}_k)$. Furthermore, (34) goes to zero by assumption. Hence, $\mathrm{E}[\sum_{k=1}^{n}(\hat{\sigma}_{ij}(\mathbf{Y}_k) - \sigma_{ij}(\mathbf{Y}_k))(nW_{lk}W_{mk})]^2 \xrightarrow[n\to\infty]{} 0$ for all $l, m = 1, \ldots, q$, $i, j = 1, \ldots, p$.

References

Billingsley, P. (1986) *Probability and Measure.* New York: Wiley.

Bunke, H. and Bunke, O. (editors) (1986) *Statistical Inference in Linear Models. Statistical Methods of Model Building, Vol I.* Wiley, Chichester.

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

Bura, E. and Cook, R. D. (1999) Extending SIR: The Weighted Chi-Square Test. Unpublished manuscript submitted to the *J. Am. Statist. Ass.*.

Camden, M. (1989) *The Data Bundle.* Wellington, New Zealand: New Zealand Statistical Association.

Cook, R. D. (1994a) On the interpretation of regression plots. *J. Am. Statist. Ass.*, **89**, 177–189.

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

——— (1996) Graphics for regressions with a binary response. *J. Am. Statist. Ass.*, **91**, 983–992.

——— (1998a) *Regression Graphics: Ideas for Studying Regressions through Graphics.* New York: Wiley.

——— (1998b) Principal Hessian Directions Revisited (with discussion). *J. Am. Statist. Ass.*, **93**, 84–100.

Cook, R. D. and Nachtsheim, C. J. (1994) Re-weighting to achieve elliptically contoured covariates in regression. *J. Am. Statist. Ass.*, **89**, 592–599.

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

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

——— (1999) *Applied Regression including Computing and Graphics.* New York: Wiley.

Dawid, A. P. (1979) Conditional Independence in Statistical Theory. *J. R. Statist. Soc. B*, **41**, 1–31.

Diaconis, P. and Freedman, D. (1984) Asymptotics of graphical projection pursuit. *Ann. Statist.*, **12**, 793–815.

Eaton, M. L.(1986) A Characterization of Spherical Distributions. *J. Mult. Anal.*, **20**, 272–276.

Eaton, M. 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.

Hall, P. and Li, K.-C. (1993) On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.*, **21**, 867–889.

Harville, D. A. (1997) *Matrix Algebra From a Statistician's Perspective.* Springer-Verlag: New York.

Johnson, M. E. (1987) *Multivariate Statistical Simulation.* Wiley: New York.

Li, K.-C. (1991) Sliced Inverse Regression for Dimension Reduction *J. Am. Statist. Ass.*, **86**, 316–342.

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

Schott, J. R. (1994) Determining the dimensionality in sliced inverse regression. *J. Am. Statist. Ass.*, **89**, 141–148.

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

Tierney, L. (1990) *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics.* New York: Wiley.

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