Statistica Sinica **8**(1998), 289-316

CAN SIR BE AS POPULAR AS MULTIPLE LINEAR REGRESSION?

Chun-Houh Chen and Ker-Chau Li

Academia Sinica and University of California, Los Angeles

Abstract: Despite its limitation in exploring nonlinear structures, multiple linear regression (MLR) still retains its popularity among the practitioners. This is mainly because of the several seemingly irreplaceable features of MLR that users are accustomed to, including : (i) it is easy to implement; (ii) it has a solid theoretical foundation; (iii) diagnostic tools are available for model checking; (iv) standard errors are available for significance assessment; (v) output is easy to interpret.

Whether such advantages can be maintained or not is an important issue in developing new nonlinear methods for high dimension regression. This issue is studied for one of the recently proposed methods, sliced inverse regression (SIR). We show how to enhance the SIR analysis so that these features can be maintained.

Key words and phrases: Dimension reduction, dynamic graphics, inverse regression, projection pursuit, transformation.

1. Introduction

The area of high dimensional regression aims at the exploration of nonlinear data structures that might not be adequately analyzed by standard multiple linear regression. Much of the development in this area has been encouraged by the advent of the modern computer. Methods such as projection pursuit regression (Friedman and Stuetzle (1981), Hall (1989), Chen (1991)), ACE (Breiman and Friedman (1985)), CART (Breiman, Friedman, Olshen and Stone (1984)), MARS (Friedman (1991)), SUPPORT (Chaudhuri, Huang, Loh and Yao (1994)), etc., generally require an extensive search through several well-motivated classes of functions. To reach a good approximation of a general nonlinear regression surface, they often need to compute goodness-of-fit criteria such as R-squared or residual sum of squares in conjunction with data-driven techniques such as cross-validation, GCV, their equivalents or modifications for offsetting model complexity and related problems. Such previously formidable tasks of functional fitting can now be carried out under sophisticated planning.

Instead of massive computation, an alternative strategy is to take good advantage of the modern computer's superior graphical facilities. Our visual talent is exploited for gaining insight about the shape of the true response function and other data structures. Such information can be used to suggest appropriate parametric models, low-dimensional smoothing techniques, or diagnostic tools for more fruitful applications. But all graphs are merely two dimensional objects. Even enhanced with rotation, coloring, and other animation techniques, visualization can still be extremely hard as one goes beyond three or four dimensions. To carry out the graphical analysis in real time, it is necessary to focus only on a selective set of projection directions.

A statistical formulation for addressing this issue is given in Li (1991), using the following dimension reduction model for regression:

$$Y = f(\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x}, \epsilon). \tag{1.1}$$

The output variable Y is related to the p-dimensional regressor $\mathbf{x}$ only through the reduced k dimensional variable $(\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x})'$. No assumption is made about f and the distribution of the error ϵ. We are interested in finding the space spanned by the k unknown β vectors, called the *effective dimension reduction* (e.d.r.) space. Sliced inverse regression (SIR) and principal Hessian directions (pHd) are developed under this dimension reduction framework (see Cook and Weisberg (1991), Carroll and Li (1992), Duan and Li (1991), Hsing and Carroll (1992), Li (1990, 1991, 1992a, 1992b), Zhu and Fang (1994), and Zhu and Ng (1994)). For more discussion on e.d.r. space and related concepts on graphical regression, see Cook (1994) and Cook and Wetzel (1994).

Both computer-intensive methods and graphics-guided methods have extended the scope of multiple linear regression (MLR). However, none of them have yet evolved into a routine practice in regression analysis and MLR still maintains its popularity. MLR has many features that users are accustomed to; for example,

(F.1) The implementation is simple.

(F.2) The statistical theory is solid (given that the model is true).

(F.3) Supplementary graphical and diagnostic tools are available for enhancing the analysis.

(F.4) Standard deviations are available for the estimated parameters.

(F.5). The output is easy to interpret.

Can these features, which are largely pertinent to linear procedures, be inherited by tools for exploring nonlinearity? This worthy goal can be pursued for SIR. In fact, SIR has retained the merits of (F.1) and (F.2). This is reviewed in Section 2. The discussion there also leads to suggestions for (F.3).

The key to (F.4) and (F.5) stems from a natural connection between SIR and MLR, to be established in Section 3. As it turns out, each SIR direction is simply a slope vector of MLR applied to an optimally transformed Y. Each eigenvalue can be interpreted as the R-squared value of the corresponding MLR. The sense of optimality refers to the maximization of the R-squared value.

Another advantage in bringing up the connection with MLR is to obtain standard deviations for the SIR directions. This is derived in Section 4. The formula turns out to be very simple, taking a form almost identical to the familiar one in MLR except for a proportionality constant determined by the eigenvalue in the SIR output. This clarifies the role of the covariance matrix of the regressors in affecting the precision of the SIR estimates. As in MLR, the t-values can also be formed for a quick significance assessment on each regressor variable.

Section 5 illustrates the enhanced SIR analysis. Three simulation examples are studied in Section 5.1. The first model has a three dimensional structure, exhibiting curves from one viewing angle and clustering from another. The second model deals with heteroscedasticity. The third model discusses a nonlinear confounding phenomenon, showing a helix-like pattern in the data. We apply SIR to the Boston housing data (Harrison and Rubinfeld (1978)) in Section 5.2. There we demonstrate how SIR can help summarize the information for studying the relationship between the median house value Y and thirteen regressor variables. A confounding pattern like the third example in Section 5.1 is found from the rotation plot for the house value (the Y variable) against two predictors found by SIR : the number of rooms, and a weighted average of the crime rate and the percentage of the poor. It looks like a helix or a slide. Harrison and Rubinfeld (1978) suggested the logarithmic transformation on Y. But this is not needed in carrying out the SIR analysis, because only the order of Y is needed in slicing. As it turns out, the logarithmic transformation appears unnecessary in our final analysis.

Our findings are summarized in Section 6. We conclude that the enhanced SIR analysis shares the same nice features as MRL. It could take the same role as MRL for routine use in data analysis. Technical details are given in the Appendix.

2. Features (F.1), (F.2) and (F.3)

Sliced inverse regression (SIR) reverses the roles of Y and $\mathbf{x}$. The inverse regression curve, $\eta(y) = E(\mathbf{x}|Y = y)$, and its covariance matrix, $\Sigma_\eta = \text{Cov } \eta(Y)$, are considered. The population version of SIR amounts to the following eigenvalue decomposition:

$$\begin{aligned} &\Sigma_\eta v_i = \lambda_i \Sigma_{\mathbf{x}} v_i \\ &\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \\ &v_i' \Sigma_{\mathbf{x}} v_i = 1, \end{aligned} \tag{2.1}$$

where $\Sigma_{\mathbf{x}}$ denotes the covariance of $\mathbf{x}$. The first few eigenvectors v_i with nonzero eigenvalues are used to project $\mathbf{x}$ for reducing the dimensionality. The linear combinations $v_i'\mathbf{x}$ formed from the SIR directions v_i will be referred to as the SIR variates. How well does SIR keep features (F.1)-(F.3)? This is the focus of this section.

First, regarding the implementation, we may estimate Σ_η by slicing. We divide the range of Y into H slices and compute the mean of $\mathbf{x}$ in each slice, $\bar{\mathbf{x}}_h, h = 1, \ldots, H$. Then we form the weighted covariance matrix of these slice mean vectors,

$$\hat{\Sigma}_\eta = \sum_{h=1}^{H} \hat{p}_h(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})', \tag{2.2}$$

where $\hat{p}_h$ is the proportion of cases falling into slice h and $\bar{\mathbf{x}}$ is the sample mean of $\mathbf{x}$. Finally, we simply replace $\Sigma_{\mathbf{x}}$ by the sample version $\hat{\Sigma}_{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ and proceed with the eigenvalue decomposition :

$$\hat{\Sigma}_\eta \hat{v}_i = \hat{\lambda}_i \hat{\Sigma}_{\mathbf{x}} \hat{v}_i \quad \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p. \tag{2.3}$$

Clearly, the computation of SIR is quite simple and feature (F.1) is retained.

Next, regarding the statistical theory of SIR, Theorem 3.1 in Li (1991) shows that SIR directions fall into the e.d.r. space under the following linearity assumption on the distribution of $\mathbf{x}$:

Linear Design Condition. For any b in R^p, the conditional expectation $E(b'\mathbf{x}|\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x})$ is linear in $\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x}$; that is, for some constants $c_o, c_1, \ldots,$ c_k, $E(b'\mathbf{x}|\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x}) = c_o + c_1\beta_1'\mathbf{x} + \cdots + c_k\beta_k'\mathbf{x}$.

Based on this, SIR estimates are shown to be root-n consistent. They are not sensitive to the number of slices used. Significance tests are also available for determining the dimensionality. Further discussion can be found in Cook and Weisberg (1991), Duan and Li (1991), Hsing and Carroll (1992), Li (1991), Li (1992a), Schott (1994), Zhu and Ng (1995), and Zhu and Fang (1996).

The requirement of the linearity condition on $\mathbf{x}$ appears to be a limitation of SIR. How restrictive is this condition? It depends on our attitude towards data analysis. At first glance, it appears that elliptically contoured distributions are the only ones that may satisfy this condition (see Cook and Weisberg (1991)). But the fact is that this condition is imposed only on the true e.d.r. directions, and it need not hold for other non-e.d.r. directions. From a conservative point of view, this fact does not help much because the e.d.r. directions are unknown and the safe way is to make sure that the condition is satisfied for any directions, even if they may not be in the e.d.r. space.

However, a less conservative point of view can be explored if we are willing to tolerate small bias due to some possible mild violation of the linear design condition. We can assess the size of the set of the β directions for which our linearity condition approximately holds. As proved in Hall and Li (1993), this set typically covers almost all directions in R^p when the dimension of $\mathbf{x}$ gets large. Thus if we adopt a Bayesian argument and consider a flat prior distribution on the unknown β's for example, with a very high probability we may find that the linearity condition is adequate. A simulation study confirming this fact was reported in the Rejoinder in Li (1991), where a non elliptically-contoured distribution, the uniform distribution on $[0,1]^{10}$ for $\mathbf{x}$, was considered. Cook and Weisberg (1991) also reported that SIR is often not overly sensitive to the linear design condition.

It certainly is desirable to avoid the limitation due to the linear design condition entirely. In principle, this can be done in several ways. For example, by density estimation and reweighting, we can force $\mathbf{x}$ to follow elliptical symmetry, an idea similar to importance sampling in the Bayesian computation literature. More resampling/reweighting methods can be found in Brillinger (1991), Li and Duan (1989), and more recently, Cook and Nachtsheim (1994).

Feature (F.3) can be pursued after obtaining the SIR directions. We can plot $b'\mathbf{x}$ against the SIR variates as a diagnostic checking for the linear design condition. A question arises regarding in which direction b is most likely to reveal nonlinearity. This is solved by the helical confounding theory developed in Li (1997). The most nonlinear direction b can be found by applying an eigenvalue decomposition similar to SIR. Instead of Y, we slice on the SIR-variates to obtain $\mathrm{Cov}\,(E(\mathbf{x}|\beta_1'\mathbf{x},\ldots,\beta_k'\mathbf{x}))$ and use this matrix to replace Σ_η in (2.1). The $(k+1)$th eigenvector turns out to be the most nonlinear direction b and the corresponding eigenvalue indicates how serious the nonlinearity is.

3. Transformation, MLR, and SIR

Transformation is a commonly used technique in regression analysis. This can be done either informally by examining plots of residuals for example, or formally by using the power transformation family as suggested in Box and Cox (1964). Difficulties arise in applying these methods to high dimensional data where there are many scatterplots to inspect. Some of them may suggest rather different transformation functions. When this happens, the Box-Cox model is then questionable because the model only allows for one transformation on Y. Monotonicity of the power function family is another constraint in some applications.

The new aspect considered here combines merits from variable transformation and projection pursuit (Huber (1985), Diaconis and Freedman (1984)). For a direction b in R^p, consider the scatterplot of Y against $b'\mathbf{x}$ and allow any (possibly non-monotone) transformation on Y for increasing the linear fit. Define $R^2(b)$ to be the largest R-squared value among all transformations:

$$R^2(b) = \max \rho^2(T(Y), b'\mathbf{x}), \tag{3.1}$$

where $\rho(\cdot,\cdot)$ denotes the correlation coefficient and the maximum is taken over any transformation $T(\cdot)$. A variety of interesting features in the scatterplot, including blurring curves, heteroscedasticity, and clusters may lead to a large value of $R^2(b)$ (see examples in section 5.1 for illustration).

Using $R^2(b)$ as the projection index, we may look for a direction b_1 that maximizes $R^2(b)$. After finding b_1, we then turn to those directions uncorrelated to b_1 for the second best direction b_2. This process can be continued in a similar fashion till we find a set of basis vectors, $b_1,\ldots,b_p$, satisfying the conditions

$$\begin{aligned}&\mathrm{Cov}\,(b_i'\mathbf{x}, b_j'\mathbf{x}) = 0,\ \text{ for } i \neq j\\ &R^2(b_i) = \max_b R^2(b),\end{aligned} \tag{3.2}$$

where the maximum in (3.2) is taken over all vectors b satisfying $\mathrm{Cov}\,(b'\mathbf{x}, b_j'\mathbf{x}) = 0$, for $j = 1,\ldots,i-1$.

Unlike other projection pursuit procedures which need extensive searching, we do have a closed form solution for b_i's. In fact, they are just the directions of SIR.

Theorem 3.1. *The SIR direction v_i as defined in* (2.1) *solves the maximization problem* (3.2) *and the maximum values are equal to the eigenvalues of SIR, $R^2(b_i) = \lambda_i$, for $i = 1,\ldots,p$.*

The proof of this Theorem is given in the Appendix A.

We can describe this transformation-based projection pursuit in a conjugate manner which begins with transformations of Y. Consider MLR on the transformed variable $T(Y)$:

$$\min_{a\in R, b\in R^p} E(T(Y) - a - b'\mathbf{x})^2.$$

Denote the least squares solution by $a(T), b(T)$. Recall the *R-squared* value, the proportion of variation in Y explained by the least squares fit:

$$R^2(T) = \frac{\mathrm{Var}\,(a(T) + b(T)'\mathbf{x})}{\mathrm{Var}\,T(Y)} = \rho^2(T(Y), b(T)'\mathbf{x}).$$

The first optimal transformation T_1 is defined to be the one that maximizes the R-squared value: T_1 solves

$$\max_T R^2(T), \tag{3.3}$$

where the maximum is taken over all transformations. To find other optimal transformations, an orthogonality constraint will be imposed. The ith optimal transformation $T_i(Y)$ is one that solves (3.3) with the maximum taken over all $T(Y)$ satisfying the condition that $\mathrm{Cov}\,(T(y), T_j(Y)) = 0$, for $j = 1,\ldots,i-1$.

Closed form solutions are again available. In fact, the following theorem shows that these optimal transformations can be found from the scatterplots of Y and the SIR directions.

Theorem 3.2. *The ith optimal transformation can be written as*

$$T_i(Y) = E(v_i'\mathbf{x}|Y = y), \tag{3.4}$$

where v_i *is the ith SIR direction as defined in* (2.1). *The associated R-squared values are equal to the eigenvalues of SIR:*

$$R^2(T_i) = \lambda_i, \text{ for } i = 1, \dots, p, \tag{3.5}$$

and the slope vectors from MLR applied to the optimal transformations are proportional to the SIR directions:

$$b(T_i) = \lambda_i v_i. \tag{3.6}$$

The key to the proof of this theorem is the observation that (3.2) and (3.3) are two versions of a common double maximization problem :

$$\max_{b,T} \operatorname{Corr}(T(Y), b'\mathbf{x}).$$

(3.2) is obtained by maximizing over T first and then over b, while reversing this order gives (3.3). The rest of the proof follows immediately from Theorem 3.1.

These two theorems depict well the behavior of SIR. It looks for the directions where the regression can be as linear as possible after transforming Y. An eigenvector of SIR corresponds to the regression slope vector in the MLR applied to an optimally transformed Y variable. Eigenvalues of SIR can be interpreted as the associated R-squared values in MLR.

As mentioned earlier, the search of an "optimal" transformation is not restricted to the monotone family. Indeed, no two T_i's can be simultaneously strictly monotone because of the orthogonality condition imposed in solving (3.3).

No single index can reflect all interesting aspects in a scatterdiagram; otherwise we may need only the index, not graphics. Our transformation-based index $R^2(b)$ is no exception. It performs poorly when the scatterdiagram of Y against $b'\mathbf{x}$ contains a pattern of symmetry about some vertical line. The correlation coefficient is zero and we cannot increase it by transforming Y. Thus $R^2(b)$ is always zero no matter how interesting the pattern of symmetry is. This offers an explanation for why SIR cannot recover the e.d.r. direction in a simple quadratic function $Y = (\beta'\mathbf{x})^2$ (see Cook and Weisberg (1991) and the Rejoinder in Li (1991) for related discussion). One remedy is to consider double transformations (Carroll and Ruppert (1988)); namely to allow the transformation on $b'\mathbf{x}$

as well. We may use the maximum correlation between y and $b'\mathbf{x}$ to quantify interestingness in the scatterplot :

$$\max_{T,g} \rho(T(y), g(b'\mathbf{x})),$$

where T, g are any squared integrable functions. How to maximize this index over all possible directions efficiently has not yet been explored.

Nonlinear multivariate analysis techniques (c.f. Gifi (1990)) such as correspondence analysis, optimal scaling, and ACE (Breiman and Friedman (1985), Koyak (1987)) use maximum correlation in statistics in a rather different manner. For example, ACE proposes the model

$$T(Y) = \sum_{i=1}^{p} g_i(x_i) + \epsilon.$$

Only one transformation on Y is allowed. But each regressor is allowed to make transformation, a feature that SIR does not have.

Remark 3.1. The duality relationship displayed in Theorem 3.2 can be put into a more general context in terms of Hilbert spaces. To simplify the notation, assume that $E\mathbf{x} = 0, EY = 0$. Consider an infinite dimensional Hilbert space, $\mathcal{H}_1$, consisting of all squared integrable transformed random variables $T(Y)$ with mean zero. Let $\mathcal{H}_2$ be the $p-$dimensional Hilbert space of $\{b'\mathbf{x} : b \in R^p\}$. These two Hilbert spaces generate a larger Hilbert space, denoted by $\mathcal{H}$. Measure the distance between two elements, w_1, w_2, in $\mathcal{H}$, by the standard deviation of $w_1 - w_2$. For any $T(Y)$ in $\mathcal{H}_1$ the closest element in $\mathcal{H}_2$ is $b(T)'\mathbf{x}$. For any $b'\mathbf{x}$ in $\mathcal{H}_2$, the closest element in $\mathcal{H}_1$ is $E(b'\mathbf{x}|Y)$. Denote $\mathcal{H}_3 = \{E(b'\mathbf{x}|Y) : b \in R^p\}$. The duality relationship in Theorem 3.2 shows the existence of orthogonal basis vectors, e_i and e_i^*, $i = 1, \ldots, p$, for $\mathcal{H}_2$ and $\mathcal{H}_3$ respectively, with the following property:

The element in $\mathcal{H}_1$ closest to e_i is a multiple of e_i^*, and conversely the element in $\mathcal{H}_2$ closest to e_i^* is a multiple of e_i.

This is a special form of the singular value decomposition problem prevalent in the context of correspondence analysis and the related subjects. Using the terminology in multivariate analysis, SIR can indeed be viewed as the canonical analysis between $\mathcal{H}_1$ and $\mathcal{H}_2$. As a generalization, it is possible to enlarge H_2 by including a few second order terms (or B-spline terms) of $\mathbf{x}$.

4. Simple Estimates for the Standard Deviations of the SIR Directions

Outputs from MLR software often attach an estimated standard deviation (i.e. standard error) to each regression coefficient. With that, users can easily form the t-ratio (= the ratio of the coefficient estimate to the standard error) for a quick assessment on the (statistical) significance of each regressor variable. It would be desirable if SIR outputs can provide similar information. But the asymptotics for SIR is more difficult than MLR. The formulae for the covariance matrix of each eigenvector $\hat{v}_i$ can be derived by combining some perturbation results for eigenvalue decomposition with large sample probabilistic arguments. For general cases, they appear complex and hard to interpret. However, the transformation theory in Section 3 offers a clue for simplification in practical use.

As it turns out, our formula is similar to the familiar one in MLR. For the ith SIR direction $\hat{v}_i$, we may associate it with the vector of the square root of the diagonal elements from the matrix

$$\frac{(1-\hat{\lambda}_i)}{\hat{\lambda}_i} \cdot n^{-1}\Sigma_{\mathbf{x}}^{-1}$$

as the estimated standard deviations. This formula brings out three messages useful to bear in mind:

(m.1) The standard errors of a SIR direction are proportional to those for the standard MLR of Y on $\mathbf{x}$.

(m.2) The inaccuracy of a SIR direction gets greater when the corresponding eigenvalue gets smaller.

(m.3) The ratio $\frac{(1-\hat{\lambda}_i)}{\hat{\lambda}_i}$ plays the role of the average of squared residuals in MLR.

To see how transformation theory is used for suggesting our formula, first recall from familiar least squares theory:

$$\mathrm{Cov}\,(\hat{\beta}_{ls}) = \sigma^2 \cdot n^{-1}\Sigma_{\mathbf{x}}^{-1}. \tag{4.1}$$

This formula remains popular for practical use even if MLR is conducted after a transformation of Y, albeit the controversy regarding whether the effect of transformation can be ignored or not (Bickel and Doksum (1981), Box and Cox (1964), Hinkley and Runger (1984)). Since we can interpret the SIR directions as being proportional to the MLR slope estimate after optimal transformation (Theorem 3.2), (m.1) is well-anticipated. It remains to explain (m.3). Suppose the optimal transformation $T_i(Y)$ were given and we conduct the standard MLR for the transformed Y values. Let $\tilde{b}(T_i)$ be the estimate of the slope vector $b(T_i)$. Recall (3.6) : SIR eigenvector v_i can be obtained from $b(T_i)$ after dividing by the constant λ_i. This suggests that the covariance matrix of the SIR estimate $\hat{v}_i$ should be equal to the covariance matrix of $\tilde{b}(T_i)$ divided by λ_i^2. Now apply (4.1) to find out $\mathrm{Cov}\,(\tilde{b}(T_i))$. Since the R-squared value of the regression is λ_i as stated in Theorem 3.2, the residual variance σ^2 in (4.1) must be equal to

$(1-\lambda_i)\mathrm{Var}\,(T_i(y)) = (1-\lambda_i)\lambda_i$. Finally dividing σ^2 by λ_i^2, we are led to the ratio $\frac{(1-\hat{\lambda}_i)}{\hat{\lambda}_i}$ given in (m.3).

Like the t-ratios in MLR, the ratios of the SIR estimates over the respective standard errors provide a convenient way to tell if the corresponding coefficients are statistically significant or not. In Appendix B, rigorous asymptotics will be developed for justifying such applications. More specifically, for the lth regressor variable, we may test the null hypothesis $\mathbf{H}_o$:

$$\mathbf{H}_o : e_l'\beta_i = 0, i = 1, \ldots, k, \tag{4.2}$$

where $e_l = (0, \ldots, 0, 1, \ldots, 0)'$ denotes the lth basis vector. The standard error we obtained is asymptotically valid under the null hypothesis (4.2).

As a cautionary note, our formulae are not valid for constructing confidence intervals. In general, the standard deviations of SIR estimates depend on the true parameters in a rather complex manner. This complexity is largely due to the additional uncertainty caused by approximating the v_i with $\hat{v}_i$ in estimating the transformation $T_i(Y)$ (a phenomenon similar to the problem of Bickel and Doksum (1981)). Thus, it remains unclear how close to the correct ones our simplified standard deviations are.

In deriving the asymptotic distribution, we have also asssumed that the number of slices used in constructing SIR estimate is fixed. Although in theory we can use as many as $H = n/2$ slices (Hsing and Carroll (1992)), practically we find no obvious advantage in using large H.

5. Examples

Three simulations are reported for illustrating the enhanced SIR analysis in Section 5.1. Then we apply it to analyze the Boston housing data.

5.1. Simulations

Example 1. Curves and clusters

Consider the model

$$Y = \mathrm{sign}(\beta_1'\mathbf{x} + \sigma_1\epsilon_1)\log(|\beta_2'\mathbf{x} + \alpha + \sigma_2\epsilon_2|), \tag{5.1}$$

where the function sign$(\cdot)$ takes value 1 or -1 depending on the sign of the argument. All coordinates of $\mathbf{x}$ and ϵ_1, ϵ_2 are independent standard normal random variables. For a clear illustration, we first study the noise-free case, $\sigma_1 = \sigma_2 = 0$. Take the dimension of $\mathbf{x}$ to be $p = 15$ and generate $n = 300$ cases with $\beta_1' = (1,1,1,1,1,1,1,1,1,0,0,0,0,0,0)$, $\beta_2' = (0,0,0,0,0,0,0,0,0,1,1,1,1,1,1)$, $\alpha = 5$. We run SIR with the number of slices equal to 20. Other numbers, 10 and 30, also show similar results. A rotation plot for Y against the first two projections is shown in Figures 5.1(a)-(d). The first eigenvector (Figure 5.1(a)) finds two curves spreading out symmetrically about the horizontal axis and the second one (Figure 5.1(c)) shows a pattern of two clusters. Table 5.1 gives the first two output eigenvectors along with estimated standard deviations and eigenvalues. They are approximately proportional to β_2 and β_1 as desired. From the t-ratios, we see that significant variables are correctly detected.

Table 5.1. The first two eigenvectors (with standard deviations and ratios) and eigenvalues of SIR for (5.1) without error terms.

first vector	$(-.05, -.03, -.01, -.03, -.01, -.03, .01, -.03, -.01, .39, .41, .44, .45, .42, .43)$
S.D.	$(.02, .02, .02, .02, .02, .02, .02, .03, .02, .02, .02, .02, .03, .02, .02)$
ratio	$(-2.1, -1.5, -0.5, -1.4, -0.3, -1.6, 0.3, -1.4, -0.3, 18, 18, 19, 18, 20, 18)$
second vector	$(.35, .39, .35, .24, .28, .30, .32, .27, .33, -.00, -.01, .03, -.02, .04, .11)$
S.D.	$(.05, .05, .04, .05, .05, .05, .05, .05, .05, .05, .04, .05, .05, .05, .05)$
ratio	$(7.2, 7.7, 8.0, 4.8, 5.7, 6.5, 6.7, 5.1, 6.6, -0.0, -0.3, 0.6, -0.3, 0.8, 2.2)$
eigenvalues	$(0.88, .61, .16, .13, .12, .08, .07, .05, .04, .02, .02, .01, .01, .00, .00)$

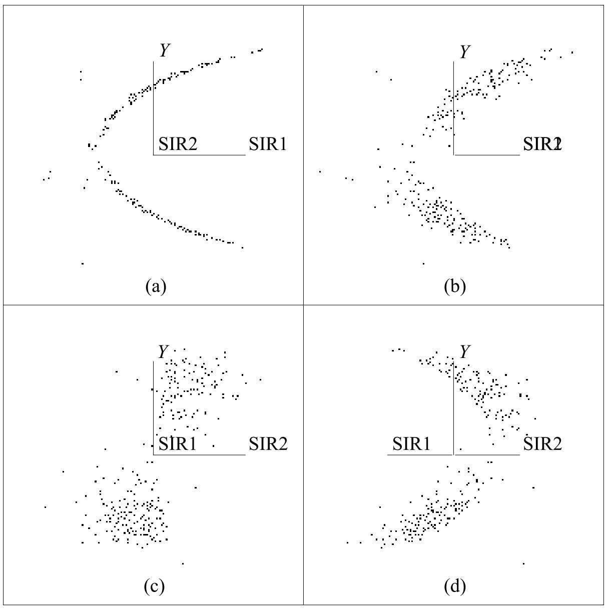

Figure 5.1. SIR's view of data generated from (5.1).

Figure 5.1(a) shows approximately the scatterplot of Y against $\beta_2'\mathbf{x}$. The symmetry about the horizontal axis is due to the sign function which acts on $\beta_1'\mathbf{x}$ behind the screen. This symmetry yields a zero correlation coefficient between Y and $\beta_2'\mathbf{x}$. But it can be increased greatly by folding the picture over along the $x-$axis, which amounts to taking the absolute value transformation $|Y|$. This explains why SIR is capable of finding this direction. According to Theorem

3.2, the optimal transformation is $T_1(Y) = E(\beta_2'\mathbf{x}|Y)$, which should give an even higher correlation coefficient, about $\sqrt{.88} \approx .94$ as estimated by the square root of the first eigenvalue of SIR, than the absolute value transformation.

Figure 5.1(c) shows approximately the scatterplot of Y against $\beta_1'\mathbf{x}$. This is the direction to be found by a linear least squares of Y against $\mathbf{x}$, because Y is uncorrelated with any directions orthogonal to $\beta_1'\mathbf{x}$.

Figures 5.1(b) and 5.1(d) show two additional views of the rotation plot found by SIR. These static views themselves do not offer much additional information; but when we rotate the plot around the vertical axis, the two curves in 5.1(a) are then turned into two thin plates, floating in and out of view.

We also repeat the simulation with the noise level set at $\sigma_1 = \sigma_2 = 1$. The output of SIR is also quite close to the directions of β_2, β_1; see Table 5.2 and Figures 5.2(a)-(b). The curves are now blurred.

Table 5.2. The first two eigenvectors (with standard deviations and ratios) and eigenvalues of SIR for (5.1) with error terms.

first vector	$(-.01, .06, .01, -.01, .02, .01, -.01, -.05, -.01, -.46, -.46, -.44, -.43, -.39, -.38)$
S.D.	$(.03, .03, .03, .03, .03, .03, .03, .04, .03, .03, .03, .03, .04, .03, .03)$
ratio	$(-0.4, 1.9, 0.4, -0.4, 0.5, 0.4, -0.4, -1.5, -0.3, -14, -14, -13, -12, -13, -11)$
second vector	$(.33, .31, .34, .27, .33, .32, .39, .22, .34, -.02, -.17, .09, .06, -.05, .14)$
S.D.	$(.05, .06, .05, .06, .05, .05, .05, .06, .06, .05, .05, .06, .06, .05, .06)$
ratio	$(6.1, 5.5, 7.0, 4.7, 6.0, 6.1, 7.2, 3.7, 6.1 - 0.3, -3.1, 1.5, 1.0, -0.9, 2.5)$
eigenvalues	$(.78, .55, .17, .12, .11, .10, .07, .05, .04, .03, .02, .01, .01, .01, .00)$

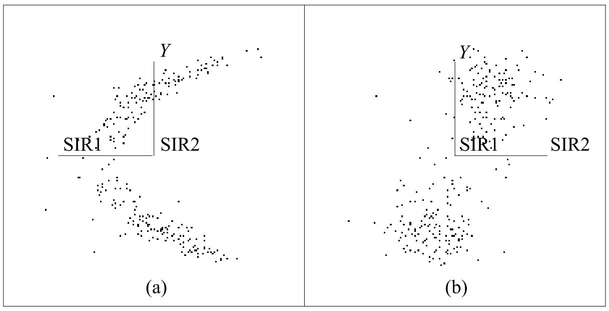

Figure 5.2. SIR's view of data generated from (5.1) with $\sigma_1 = \sigma_2 = 1$, $p = 15$.

Remark 5.1. We also simulated the case with $\alpha = 0$. SIR fails in this case because of the symmetry on the β_2 direction. Second-moment based methods (Cook and Weisberg (1991), the Rejoinder in Li (1991)) and variants of principal Hessian direction (Li (1992b)) are capable of finding the β_2 direction.

Remark 5.2. It is relatively easy to find the transformation functions once the SIR directions are obtained. For example, $T_1(Y)$ can be found by first rotating the scatterplot (a) in Figure 5.1 by 90^o so that the Y becomes the horizontal axis and SIR1 becomes the vertical axis. After that, simply smooth the data by applying any nonparametric curve fitting techniques. But we do not need these transformations for further analysis. In this example, it is better to fit a two dimensional regression surface, Y on the first two SIR variates. We could easily reach more than 95% of the R-squared value of the fit, as compared with about 60% if MLR is applied to the same data. In fact, as suggested from the rotation plot, we can further partition the data into two regions along the second SIR variate, according to $\hat{v}_2'\mathbf{x} > 0$ or not, so that after partition in each region we only need one-dimensional smoothing of Y on $\hat{v}_1'\mathbf{x}$.

In the literature, Box and Cox (1964) claimed that their power transformation theory intends to accomplish : (1) linearizing the regression, (2) stabilizing the variance, (3) achieving normality. Whether such goals can be met or not depends very much on the data themselves. This raises an important issue : how does one know such a transformation does not exist? An answer to this question could save us a lot of time by avoiding fruitless attempts with various transformation functions. Thus if two or more significant SIR directions are found, it is a good indication that Box and Cox's goals might not be met.

In general, the transformation theory of SIR leads to the conclusion that transformation on Y may be recommended only when the data show one significant SIR direction. In the Rejoinder of Li (1991), such an example, the worsted yarn data from Box and Cox (1964), is mentioned , where a logarithmic transformation curve is visible from the SIR plot.

Example 2. Heteroscedasticity

A popular model for studying heteroscedasticity is

$$y = \beta_1'\mathbf{x} + \epsilon g(\alpha + \beta_2'\mathbf{x}), \tag{5.2}$$

where g is often conveniently taken to be a power transformation function (see, e.g., Carroll, Wu, and Ruppert (1988)).

To see how SIR helps the residual analysis, we take $g(x) = .2x$ and generate 100 cases for $p = 6$ with $\beta_1' = (1,1,1,1,0,0)$, $\beta_2' = (0,0,0,0,1,1)$, $\alpha = 3$, $\epsilon \sim N(0,1)$. Fit the data by the usual linear least squares and find the residual r. Since $\beta_1'\mathbf{x}$ is uncorrelated with $\beta_2'\mathbf{x}$, the heteroscedasticity occurs along a direction orthogonal to the direction of the best linear fit. Thus we do not anticipate to find any pattern by examining the usual residual plot, the plot of r against predicted values (see Figure 5.3(a)).

Table 5.3. The first eigenvector (with standard deviations and ratios) and eigenvalues of SIR for residuals of (5.2).

first eigenvector	$(-.05, -.04, .18, .06, -.71, -.79)$
S.D.	$(.13, .17, .13, .14, .13, .14)$
ratio	$(-0.4, -0.2, 1.4, 0.5, -5.4, -5.5)$
eigenvalues	$(.37, .23, .13, .07, .03, .01)$

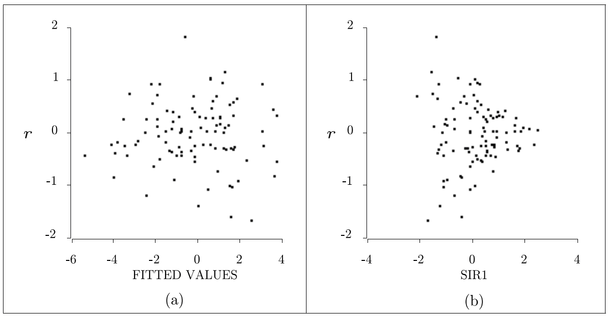

Figure 5.3. Residuals against linear squares fit(a) and SIR1 direction(b) of model (5.2).

Now we run SIR on r. Figure 5.3(b) gives the plot of r against the first direction found by SIR. It does reveal the heteroscedasticity pattern well.

The reason why SIR can help in residual analysis is easy to understand. Although r is, by definition, uncorrelated with $\mathbf{x}$, we can apply a transformation on r to increase the correlation and SIR does that in an "optimal" way. There is no need to take the absolute value transformation on r before applying SIR. The flexibility in allowing for non-monotone transformation is the key to the success.

Example 3. Horseshoe and helix

A five-dimensional input variable $\mathbf{x} = (x_1, \ldots, x_5)'$ is obtained by first generating 1000 cases for $\mathbf{x}$ from the standard normal distribution and then retaining only those cases that satisfy the constraint:

$$x_1^2 - 0.5 < x_2 < x_1^2 + 0.5. \tag{5.3}$$

This reduces the sample size to 288. Now a linear model is used to generate Y

$$Y = x_1 + 0.5\epsilon, \ \epsilon \sim N(0, 1). \tag{5.4}$$

The output of SIR shows two large eigenvalues (see Table 5.4). Figures 5.4(a)-(d) are some static pictures of the rotational plot found by SIR. By rotating the plot about the vertical axis, we find data points spinning like a helix or a slide.

Table 5.4. The first two eigenvectors (with standard deviations and ratios) and eigenvalues of SIR for (5.3), (5.4).

first eigenvector	$(-1.64, .10, .02, -.03, .01)$
S.D.	$(.07, .09, 0.04, 0.04, .04)$
ratio	$(-23, 1.1, 0.5, -0.6, 0.3)$
second eigenvector	$(.16, 2.1, .06, -.01, -.02)$
S.D.	$(.15, .19, .09, 0.09, 0.09)$
ratio	$(1.0110.6 - 0.1 - 0.2)$
eigenvalues	$(.66, .30, .056, .023, .01)$

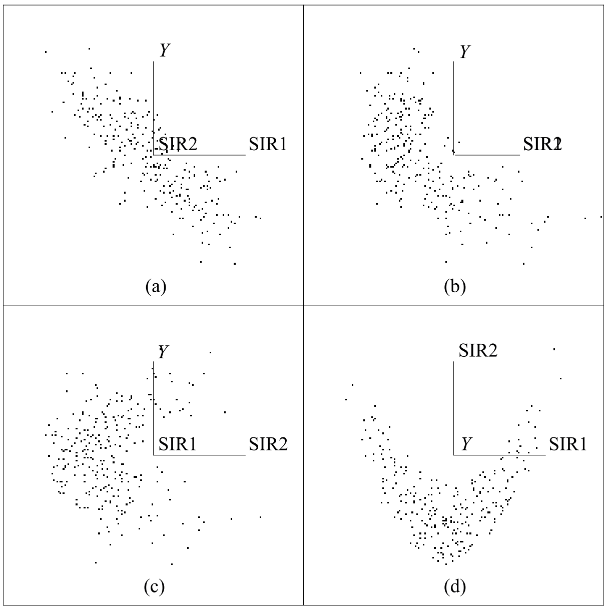

Figure 5.4. SIR's view of data generated from (5.3)-(5.4).

The first direction shows a linear pattern (Figure 5.4(a)) and the second direction finds a curve (Figure 5.4(c)). They correspond to x_1 and x_2 approximately. The scatterdiagram of these two SIR directions, Figure 5.4(d), shows a horseshoe pattern, exhibiting the quadratic constraint (5.3).

In this example, x_2 is nonlinearly correlated with x_1, a situation where the linear design condition is severely violated. SIR picks up this additional direction because Y can be transformed to retain a significant correlation with x_2.

A data set with a pattern like the one just observed here creates some difficulties in modeling which have not received proper attention in the literature. First of all, we may not be able to tell if the number of the components is one or two. For example, data generated by a two-components model of the form $Y = \text{sign}(x_1)\sqrt{|x_2|} + .5\epsilon$ presents little visual difference from the one we just found. In addition, even if a one-component model is assumed, we may not have much information to estimate the correct direction well without knowing the correct functional form.

Perhaps exhibiting this low dimensional nonlinear confounding patterns is scientifically more important than attempting to resolving this issue statistically. Graphics gives scientists something to focus on. It helps stimulate relevant knowledge. Furthermore, such limitation due to confounding is shown to be pertinent to every estimation procedure (see Li (1997)).

5.2. The Boston housing data cloud: a helix or slide

We now apply SIR to the Boston housing data (Harrison and Rubinfeld (1978)). The dependent variable Y is taken to be the logarithm of a variable of special interest, the median value of owner occupied homes in each of the 506 census tracts in Boston Standard Metropolitan Statistical Areas. There are 13 regressor variables (see Table 5.5). The sample size is $n = 506$, each case representing a census tract.

Table 5.5. Variables in Boston housing data.

Y	logarithm of the median value of owner-occupied home
x_1	crime rate by town
x_2	proportion of town' residential land zoned for lots greater than 25,000 square feet
x_3	proportion of nonretail business acres per town
x_4	= 1 if tract bounds Charles River, =0 otherwise
x_5	nitrogen oxide concentration in pphm
x_6	average number of rooms in owner units
x_7	proportion of owner units built prior to 1940
x_8	weighted distances to five employment centers in the Boston region
x_9	index of accessibility to radial highways
x_{10}	full property tax rate
x_{11}	pupil-teacher ratio by town school district
x_{12}	black proportion of the population
x_{13}	proportion of population that is in the lower status

Table 5.6. The first three eigenvectors (with standard deviations and ratios) and eigenvalues of SIR for Boston housing data.

first vector	$(-.022, .002, -.001, .27, -27, 0.25, -.004, -.77, .28, -.001, -.10, 1.3, -7.6)$
S.D.	$(.003, .001, .006, .087, .991, .042, .001, .089, .050, .0003, .013, .271, .512)$
ratio	$(-6.8, 1.6, -0.2, 3.1, -6.7, 6.0, -2.7, -8.7, 5.5, -4.0, -7.7, 4.6, -15.0)$
second vector	$(.05, .01, -.05, .07, -10.8, 1.04, -.007, -1.47, .05, .0003, -.05, -1.3, 5.9)$
S.D.	$(.007, .003, .014, .190, 8.70, .091, .003, .193, .110, .0007, .028, .592, 1.12)$
ratio	$(6.6, 4.3, -3.8, 0.4, -1.2, 11, -2.3, -7.6, 0.5, 0.4, -1.9, -2.2, 5.3)$
third vector	$(-.09, .01, -.009, -.12, 29.0.55, .03, .18, -.09, .0005, .08, -.36, 1.58)$
S.D.	$(.013, .006, .026, .364, 16.7, .174, .006, .370, .211, .001, .054, 1.13, 2.14)$
ratio	$(-6.8, 2.0, -0.4, -0.3, 1.7, 3.1, 5.3, 0.5, -0.5, 0.4, 1.5, -0.3, 0.8)$
eigenvalues	$(.82, .48, .20, .08, .05, .04, .03, .02, .01, .00, .00, .00, .00)$

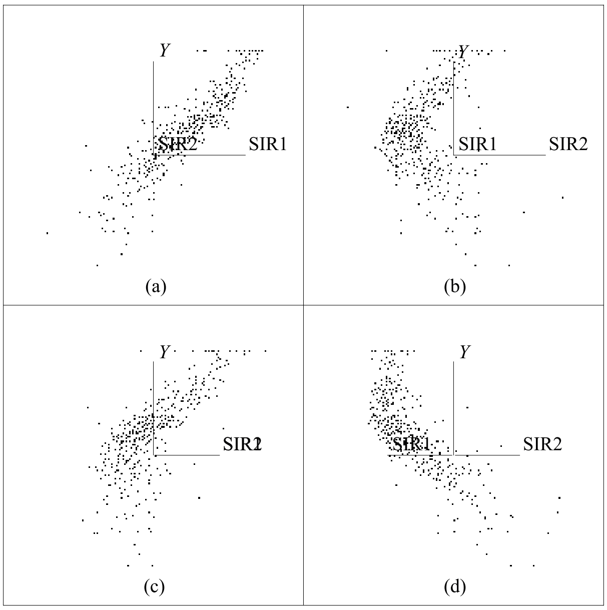

Figure 5.5. SIR's view on Boston housing data. Y is the logrithm of the median value of owner-occupied house.

We run SIR with H, the number of slices, ranging from 10 to 30. The results are almost the same. The one for $H = 15$ is reported in Figures 5.5(a)-(d). By rotating the plot, a helix or slide is found, resembling what we have seen in Figures 5.4(a)-(d).

We would like to find a smaller number variables that contribute the most to the variation of each projection. For simplicity, we use a forward stepwise regression instead of all subset selections. For the first projection, the main contributor is x_{13}, proportion of poor, with x_6, average number of rooms in houses, as a close runner-up. The second projection is analyzed similarly but during selection the significant variables identified in the first direction have to be kept in the regression. As it turns out, the top candidate x_1, crime rate, leads seven other competing variables only marginally. We fail to find a small number of regressors to approximate the second projection.

A closer look at the relationship of crime rate with other variables is taken by inspecting scatterplots. A group of tracts with high crime rate is easily identified. The values of the variables x_2, x_3, x_9, x_{10}, and x_{11} turn out to be the same for all members in this group.

Excluding this high crime rate group, we run SIR on the remaining 374 cases. The pictures are similar to but sharper than the ones obtained from the whole sample. This time we succeed in reducing the number of variables contributing to the first two projections to just three: x_6, x_1 and x_{13}.

We finally run SIR again, with x_1, x_6, x_{13} as the regressor variables (see Table 5.7). The first projection is x_6. The second projection is tailored from a linear combination of the crime rate and the proportion of poor, $x_1 + 30x_{13}$, (numbers rounded), for maintaining orthogonality to x_6.

Table 5.7. The first two eigenvectors (with standard deviations and ratios) and eigenvalues of SIR for Y on x_1, x_6, x_{13}.

first vector	$(.08, 1.30, -3.14)$
S.D.	$(.045, .053, .633)$
ratio	$(1.72, 24.6, -4.95)$
second vector	$(.64, 1.38, 18.3)$
S.D.	$(.111, .131, 1.57)$
ratio	$(5.81, 10.6, 11.7)$
eigenvalues	$(.80, .40, .03)$

We again find the helix-type of data pattern. Figure 5.6(a) shows a linear relationship between Y and x_6. A different angle of the same rotation plot, Figure 5.6(b), suggests a nonlinear association between Y and another variable, $x_1 + 30x_{13}$. The data do not seem to contain enough information to tell which factor is more important. Figure 5.7 reveals that these two factors are indeed nonlinearly associated as well, reminiscent of the confounding issue discussed in Example 3 of Section 5.1.

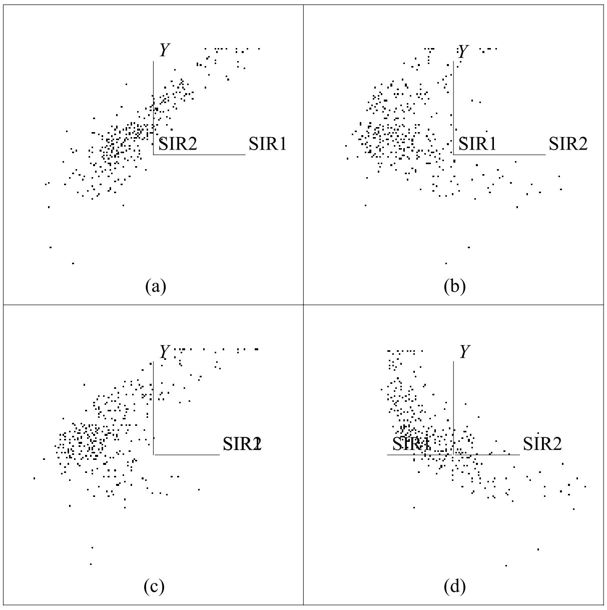

Figure 5.6. SIR's view on the subsample that excludes a high crime rate group with only x_1, x_6, x_{13} as the regressor variables. Y is the logrithm of the median value of owner-occupied house.

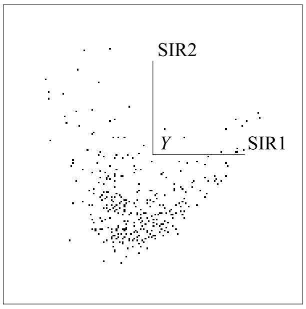

Figure 5.7. The scatterplot of the first two components in Figures 5.5 (a)-(d).

The logarithmic transformation used to obtain Y is borrowed from Harrison and Rubinfeld (1978); but because SIR is invariant under the monotone transformation of Y, SIR would still find the same projections if the original scale were used. In Figures 5.8(a)-(b), the original scale of house price is plotted. Compared with Figures 5.6(a)-(b), we don't find it compelling to use a logarithmic transformation (a point also concluded in the ACE analysis of Breiman and Friedman(1985)).

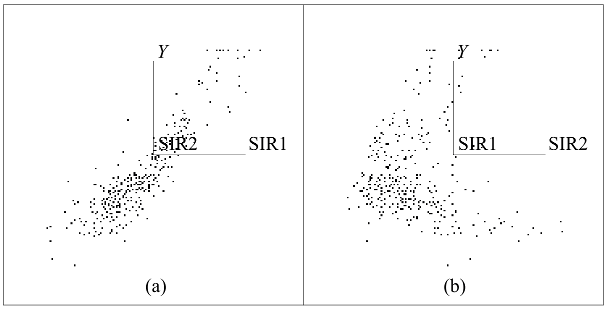

Figure 5.8. SIR's view on the subsample that excludes a high crime rate group with only x_1, x_6, x_{13} as the regressor variables. Y is the original scale, median value of owner-occupied house.

Unlike other previous analyses focusing on prediction equations, SIR identifies two key factors of different nature and provides a graphical summary. The variable x_6, average number of rooms, is a physical factor, which may reflect the construction cost and the practical utility of a house to some degree. It affects the structure value of a house. The other variable $x_1 + 30x_{13}$, the crime rate and percentage of the poor, is a socio-economic factor. It reflects the desirability of the house's neighborhood which in turn affects the area's land value. SIR reveals the nonlinear association between these two factors. The importance of the physical factor is also confirmed by other methods; for example, the straightforward linear regression, the more complicated model fitting of Harrison and

Rubinfeld, and ACE of Breiman and Friedman. In fact, in each of these studies, the physical variable is the leading factor which accounts for the highest percentage of variation in the prediction equation. One might then conclude that this is the dominant factor. However, the helix type of nonlinear confounding pattern exhibited by the three-dimensional SIR plot offers a challenge to such a statement.

It is usually hard to draw any decisive conclusion from a single study. If the same helix shape of distribution also exists in data from other cities, for example, then the finding would be much more noteworthy. The graphics found here, however, is not available from linear regression or other methods. The exposure of the helix type data cloud offers a warning diagnosis for methods based on approximating the regression surface, which are sensitive to nonlinear confounding.

Remark 5.3. The logarithmic transformation is sometimes recommended to help stabilize variances. This is a very subtle issue. Because we did not find heterogeneous patterns like Example 2 in Section 4.2, our analysis suggests that this is at most a secondary issue for this data.

6. Conclusion

Linear regression by ordinary least squares is often used in pratice. It has nice statistical properties like unbiasedness and consistency if the linear model is correct. It also has the remarkable, though less well-known, property that the estimate is still consistent up to a proportional constant under (1.1) with $k = 1$ and the linear design Condition due to Brillinger (1977, 1983) (see also Cook, Hawkins, and Weisberg (1992), Li and Duan (1989)).

But the popularity of linear least squares regression in exploratory data analysis is also due to its straightforward interpretation without model assumptions: linear least squares regression offers a “best” llinear approximation of the regression surface; in different terms, it finds the linear combination of the independent variables that maximizes the correlation with the dependent variable.

Likewise, for popularizing SIR, this paper offers a simple interpretation on what it does without model assumptions: it is a transformation-based projection pursuit; it finds linear combinations of independent variables that maximize the correlation with the optimally transformed dependent variable. This finding complements the dimension reduction theory of SIR given in Li (1991), based on Model (1.1) and the linear design condition. It also leads to simple estimates for the standard deviations of the SIR directions.

The implementation of SIR is almost as simple as linear regression. Yet it offers a much broader perspective, particularly in concert with the use of interactive, multi-dimensional, dynamic graphing techniques. SIR can be used to take the same role as linear regression in model building, residual analysis, regression diagnosis, etc.. It supplements and greatly enhances the power of linear regression. A program in xlisp.stat (Tierney (1989)) for carrying out the SIR analysis is available from either author.

The main limitation of SIR is in finding patterns symmetric about the vertical axis. This leaves room for using other simple methods, for example, second-moment based methods like SAVE (Cook and Weisberg (1991)), SIRII (Rejoinder in Li (1991)), or pHd (Li (1992b)).

It is inappropriate to view SIR as a rival to other tools for analyzing high dimensional data. ACE, CART, MARS, or PPR all aim at approximating the regression function. They are computer-intensive and do not use graphical information. SIR approaches data analysis via a different rationale. It is a graphics-driven method and is simple to compute. It can be used whenever there is a need for visualization, which in turn can help functional approximation. SIR can be used together with other methods in many ways. It helps study the shape of their residuals at least. One can also apply SIR to the rescaled **x** variables found by ACE to check if the additivity assumption of ACE is violated or not. Finally, the idea of SIR can be used or extended in other contexts. For error-in-regressor problems, see Carroll and Li (1992), and Knickerbocker, Wang and Carroll (1992).

Acknowledgements

Part of this work was carried out when the second author was visiting the Institute of Statistical Science, Academia Sinica. The hospitality and stimulating discussion from several colleagues are very much appreciated. Some of the materials are extracted from Li (1990). Li's research is supported in part by grants from National Science Foundation, Guggenheim Fellowship, and a Research Grant from National Science Council, Taiwan.

Appendices

A. Proof of Theorem 3.1.

First, it can be verified that for any direction b, the optimal transformation is $T_b(y) = E(b'\mathbf{x}|Y = y) = b'\eta(y)$. Then a simple conditional expectation argument leads to

$$\begin{aligned}\mathrm{Cov}\,(T_b(Y), b'\mathbf{x}) &= E[T_b(Y)(b'\mathbf{x})] - [ET_b(Y)][Eb'\mathbf{x}]\\ &= E[T_b(Y)E(b'\mathbf{x}|Y)] - [ET_b(Y)E(E(b'\mathbf{x}|Y))]\\ &= b'E(\eta(Y)\eta(Y)')b - b'(E\eta(Y))(E\eta(Y))'b = b'\Sigma_\eta b\\ &= \mathrm{Var}\,(T_b(Y)).\end{aligned}$$

It follows that

$$R^2(b) = \text{Cov}\,(T_b(Y), b'\mathbf{x})/\text{Var}\,(b'\mathbf{x}) = \frac{b'\Sigma_\eta b}{b'\Sigma_{\mathbf{x}} b}.$$

Therefore the eigenvalue decomposition of Σ_η with respect to $\Sigma_{\mathbf{x}}$ solves the maximization problem (3.2), completing the proof.

B. Derivation for Standard Deviations of SIR Estimates

Let e be any vector in the orthogonal complement of the e.d.r. space; i.e.,

$$e'\beta_i = 0, i = 1, \ldots, k. \tag{1}$$

We shall derive a formula for the asymptotic variance of $e'\hat{v}_i$.

We consider only the case that the number of slices H is fixed. One way of slicing is to use pre-specified fixed intervals on Y. Let $\delta_h(Y)$ be an indicator function which takes value 1 (if Y falls into the hth interval) or 0 (otherwise). Define $\mu_h = E(\mathbf{x}|\delta_h(Y) = 1)$, $\Sigma_\eta^* = \sum_{h=1}^{H} p_h(\mu_h - E\mathbf{x})(\mu_h - E\mathbf{x})'$, where $p_h = E\delta_h(Y)$, and $Y^* = \sum_{h=1}^{H} \delta_h(Y)E(Y|\delta_h(Y) = 1)$. It is clear that $\hat{\Sigma}_\eta$, defined by (2.2), converges to Σ_η^*, which can be viewed as a discretized version of Σ_η when Y^* (instead of Y) is available. It is easy to check that the statements of Theorems 3.1 and 3.2 remain valid if Y is replaced by Y^* in (3.1), (3.4), and the eigenvalues λ_i and eigenvectors v_i are replaced by the λ_i^* and v_i^* respectively from the following discrete population version of SIR:

$$\Sigma_\eta^* v_i^* = \lambda_i^* \Sigma_{\mathbf{x}} v_i^*. \tag{2}$$

The following is a mild assumption:

Assumption (B.1). Any nonzero eigenvalue λ_i^* in the above eigenvalue decomposition, is simple (i.e, it is different from other eigenvalues).

In addition to (B.1), we shall need another assumption, which is similar to the homogeneous assumption in MLR. Recall that the most general form of the covariance matrix for the slope estimate $\hat{\beta}_{ls}$ is more complicated than (4.1):

$$\text{Cov}\,(\hat{\beta}_{ls}) = n^{-1}\Sigma_{\mathbf{x}}^{-1} E(\epsilon^2 \cdot (\mathbf{x} - E\mathbf{x})(\mathbf{x} - E\mathbf{x})')\Sigma_{\mathbf{x}}^{-1},$$

where the residual $\epsilon = Y - EY - b_{ls}'(\mathbf{x} - E\mathbf{x})$. Thus $\text{Var}\,(e'\hat{\beta}_{ls}) = n^{-1}E(\epsilon^2(e'\Sigma_{\mathbf{x}}^{-1}(\mathbf{x} - E\mathbf{x}))^2)$, which is reduced to the familar (and simpler) one inferred from (4.1), $n^{-1}\sigma^2 e'\Sigma_{\mathbf{x}}^{-1} e$, when errors are homogeneous so that

$$\text{Cov}\,(\epsilon^2, (e'\Sigma_{\mathbf{x}}^{-1}\mathbf{x})^2) = 0. \tag{3}$$

We note however that the common practice is to use (4.1) unless severe hetroscedasticity is found.

To introduce the homogeneity analogous to (3) in our case, let $T_i^*(Y^*) = E(v_i^{*\prime}\mathbf{x}|Y^*) = v_i^{*\prime}(\sum_{h=1}^H \delta_h(Y)\mu_h)$ be the ith optimal transformation suggested by (3.4) of Theorem 3.2 (applied to Y^*). Let b_i^* be the slope of MLR for the transformed variable $T_i^*(Y^*)$. Thus

$$b_i^* = \Sigma_{\mathbf{x}}^{-1}\,\mathrm{Cov}\,(\mathbf{x}, T_i^*(Y^*)) = \lambda_i^* v_i^*. \tag{4}$$

Define the residual

$$r^* = T_i^*(Y^*) - E(T_i^*(Y^*)) - b_i^{*\prime}(\mathbf{x} - E\mathbf{x}). \tag{5}$$

Assumption (B.2). The residual of MLR after transformation is homogeneous in the sense that $\mathrm{Cov}\,(r^{*2}, (e'\Sigma_{\mathbf{x}}^{-1}\mathbf{x})^2) = 0$.

Theorem B.1. *Consider model* (1.1) *and assume the Linear Design condition given in Section* 2 *together with assumption* (B.1)*. The number of slices used in SIR is assumed fixed. Then for any vector* e *which is orthogonal to the e.d.r. space, we have*

$$e'\hat{v}_i = \lambda_i^* n^{-1}\Sigma_{j=1}^n r_j^* e'\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_j - E\mathbf{x}) + O_p(n^{-1}). \tag{6}$$

With the additional assumption (B.2)*, the asymptotic variance of* $e'\hat{v}_i$ *is equal to*

$$\frac{(1-\lambda_i^*)}{\lambda_i^*}\cdot n^{-1}e'\Sigma_{\mathbf{x}}^{-1}e. \tag{7}$$

Proof. The orthogonality assumption (1) implies the following identities which will be used later on:

$$e'\Sigma_{\mathbf{x}}^{-1}\mu_h = 0,\ e'\Sigma_{\mathbf{x}}^{-1}\Sigma_\eta^* = 0,\ e'v_i^* = 0. \tag{8}$$

The first identity is due to the basic theorem of SIR (Li (1991), Theorem 3.1) which implies that $\Sigma_{\mathbf{x}}^{-1}\mu_h$ falls into the e.d.r. space. The second identity follows from the first one, which also implies the third one.

Now define $\Delta v = \hat{v}_i - v_i^*$ and $\tilde{v}_i = \hat{\Sigma}_{\mathbf{x}}^{-1}\hat{\Sigma}_\eta v_i^*$. From (2.3), we can write

$$\begin{aligned} 0 &= (\hat{\Sigma}_{\mathbf{x}}^{-1}\hat{\Sigma}_\eta - \hat{\lambda}_i I)\hat{v}_i \\ &= \tilde{v}_i - \hat{\lambda}_i v_i^* + (\hat{\Sigma}_{\mathbf{x}}^{-1}\hat{\Sigma}_\eta - \hat{\lambda}_i I)\Delta v \\ &= \tilde{v}_i - \hat{\lambda}_i v_i^* + (\Sigma_{\mathbf{x}}^{-1}\Sigma_\eta^* - \lambda_i^* I)\Delta v + O_p(n^{-1}). \end{aligned}$$

Left-multiply the last expression by e' and use (8) to obtain $e'\tilde{v}_i - \lambda_i^* e'\Delta v = O_p(n^{-1})$. This gives

$$e'\Delta v = \lambda_i^{*-1}e'\tilde{v}_i + O_p(n^{-1}). \tag{9}$$

To approximate the term $e'\tilde{v}_i$, we assume $E\mathbf{x} = 0$ without loss of generality. Define $\tilde{\Sigma}_\eta = \sum_h \hat{p}_h(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})\mu_h'$. Straightforwardly,

$$\begin{aligned}
e'\tilde{v}_i &= e'\hat{\Sigma}_{\mathbf{x}}^{-1}\tilde{\Sigma}_\eta v_i^* + e'\hat{\Sigma}_{\mathbf{x}}^{-1}(\hat{\Sigma}_\eta - \tilde{\Sigma}_\eta)v_i^* \\
&= e'\hat{\Sigma}_{\mathbf{x}}^{-1}\tilde{\Sigma}_\eta v_i^* + e'\hat{\Sigma}_{\mathbf{x}}^{-1}\sum_h \hat{p}_h(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})(\bar{\mathbf{x}}_h - \bar{\mathbf{x}} - \mu_h)'v_i^* \\
&= e'\hat{\Sigma}_{\mathbf{x}}^{-1}\tilde{\Sigma}_\eta v_i^* + \sum_h \hat{p}_h e'\Sigma_{\mathbf{x}}^{-1}\mu_h(\bar{\mathbf{x}}_h - \mu_h)'v_i^* + O_p(n^{-1}) \\
&= e'\hat{\Sigma}_{\mathbf{x}}^{-1}\tilde{\Sigma}_\eta v_i^* + O_p(n^{-1}) \\
&= e'\hat{\Sigma}_{\mathbf{x}}^{-1}(n^{-1}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})T_i^*(Y_j^*)) + O_p(n^{-1}).
\end{aligned} \tag{10}$$

Without the vector e, the leading term is just the slope estimate of MLR when applied to the transformed data $T_i^*(Y_j^*) = v_i^{*\prime}(\sum_{h=1}^{H}\delta_h(Y_j)\mu_h)$, $j = 1,\ldots,n$. Its asymptotic expansion is easy to obtain. Let $r_j^*, j = 1,\ldots,n$ be the residuals, defined according to (5), and recall (4). We have

$$\begin{aligned}
\hat{\Sigma}_{\mathbf{x}}^{-1}\Big(n^{-1}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})T_i^*(Y_j^*)\Big) &= b_i^* + \hat{\Sigma}_{\mathbf{x}}^{-1}\Big(n^{-1}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})r_j^*\Big) \\
&= \lambda_i^* v_i^* + n^{-1}\sum_{j=1}^{n} r_j^*\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_j - E\mathbf{x}) + O_p(n^{-1}).
\end{aligned} \tag{11}$$

Putting (9), (10) and (11) together and applying (8) again, we obtain (6).

Assumption (B.2) implies that

$$E[r_j^{*2}(e'\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_j - E\mathbf{x}))^2] = Er_j^{*2}\cdot E(e'\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_j - E\mathbf{x}))^2 = Er_j^{*2}\cdot e'\Sigma_{\mathbf{x}}^{-1}e.$$

A direct calculation similar to Theorem 3.2 (with Σ_η replaced by Σ_η^*) shows that $\mathrm{Var}\,(r_j^*) = (1-\lambda_i^*)\lambda_i^*$. This shows that the asymptotic variance of $e'\hat{v}_i$ is equal to (7). The proof of Theorem B.1 is complete.

Instead of fixing intervals, another way of slicing is to use the order statistics of Y, $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$. The slice means are now defined by $\bar{\mathbf{x}}_h^o = \sum_{j=[np_{h-1}]}^{[np_h]}\mathbf{x}_{(j)}$, where $p_1,\ldots,p_H$ are prespecified proportions and $[np_h]$ is interpreted as the integer closest to np_h. The superscript o is used hereafter to denote quantities associated with slicing by order statistics; for example, $\hat{\Sigma}_\eta^o$ denotes the covariance matrix of slice means $\bar{\mathbf{x}}_h^o$, and $\hat{v}_i^o$, $\hat{\lambda}_i^o$ are respectively the ith SIR estimate and eigenvalue.

This alternative way of slicing is related to the fixed-interval slicing with the hth interval set as $[F_Y^{-1}(p_{h-1}), F_Y^{-1}(p_h)]$, where $F_Y(\cdot)$ is the cumulative distribution function of Y. It is not hard to argue that the difference between $\bar{\mathbf{x}}_h^o$ and $\bar{\mathbf{x}}_h$

is of the order $n^{-\frac{1}{2}}$. Resulting from this, the differences, $\hat{\lambda}_i^o - \hat{\lambda}_i$,and $\hat{v}_i^o - \hat{v}_i$, are also found to be of the same order. However, with a more delicate elaboration, we shall establish

$$e'\hat{v}_i^o = e'\hat{v}_i + O_p(n^{-\frac{3}{4}}), \tag{12}$$

where $\hat{v}_i$ denotes the ith SIR estimate resulting from the fixed interval slicing. Therefore the asymptotic variance for $e'\hat{v}_i^o$ remains the same as (7).

Proof of (12). Without loss of generality, we assume $E\mathbf{x} = 0$. We begin with an argument similar to the one leading to (9):

$$\begin{aligned} 0 &= (\hat{\Sigma}_{\mathbf{x}}^{-1}\hat{\Sigma}_\eta^o - \hat{\lambda}_i^o I)\hat{v}_i^o \\ &= (\hat{\Sigma}_{\mathbf{x}}^{-1}\hat{\Sigma}_\eta^o - \hat{\lambda}_i^o I)(\hat{v}_i^o - \hat{v}_i) + \hat{\Sigma}_{\mathbf{x}}^{-1}(\hat{\Sigma}_\eta^o - \hat{\Sigma}_\eta)\hat{v}_i - (\hat{\lambda}_i^o - \hat{\lambda}_i)\hat{v}_i \\ &= (\Sigma_{\mathbf{x}}^{-1}\Sigma_\eta^* - \lambda_i^* I)(\hat{v}_i^o - \hat{v}_i) + \Sigma_{\mathbf{x}}^{-1}(\hat{\Sigma}_\eta^o - \hat{\Sigma}_\eta)v_i^* - (\hat{\lambda}_i^o - \hat{\lambda}_i)\hat{v}_i + O_p(n^{-1}). \end{aligned}$$

Left-multiplying by e', we finally get

$$e'(\hat{v}_i^o - \hat{v}_i) = \lambda_i^{*-1} e'\Sigma_{\mathbf{x}}^{-1}(\hat{\Sigma}_\eta^o - \hat{\Sigma}_\eta)v_i^* + O_p(n^{-1}).$$

Thus to obtain (12), it suffices to show that for each h

$$e'\Sigma_{\mathbf{x}}^{-1}(\bar{\mathbf{x}}_h^o - \bar{\mathbf{x}}_h) = O_p(n^{-\frac{3}{4}}). \tag{13}$$

Let $\delta_h^o(Y_j)$ be the hth indicator function associated with slicing by order statistics. We can express $\bar{\mathbf{x}}_h^o$ as $[nh_p]^{-1}\sum_{j=1}^n \delta_h^o(Y_j)\mathbf{x}_j$. Thus, the left side of (13) becomes

$$\begin{aligned} e'\Sigma_{\mathbf{x}}^{-1}(\bar{\mathbf{x}}_h^o - \bar{\mathbf{x}}_h) &= e'\Sigma_{\mathbf{x}}^{-1}[np_h]^{-1}\sum_{j=1}^n(\delta_h^o(Y_j) - \delta_h(Y_j))\mathbf{x}_j + ([np_h]^{-1} - (n\hat{p}_h)^{-1})e'\Sigma_{\mathbf{x}}^{1}\bar{\mathbf{x}}_h \\ &= e'\Sigma_{\mathbf{x}}^{-1}[np_h]^{-1}\sum_{j=1}^n(\delta_h^o(Y_j) - \delta_h(Y_j))\mathbf{x}_j + 0_p(n^{-1}). \end{aligned}$$

Define $m(Y_j) = E(\mathbf{x}_j|Y_j)$ and $u_j = \mathbf{x}_j - m(Y_j)$. Similar to the argument leading to (8), we have $e'\Sigma_{\mathbf{x}}^{-1}m(Y_j) = 0$. Thus conditional on $Y_1, \ldots, Y_n$, we can write the leading term in the last expression as

$$\begin{aligned} &e'\Sigma_{\mathbf{x}}^{-1}[np_h]^{-1}\sum_{j=1}^n(\delta_h^o(Y_j) - \delta_h(Y_j))m(Y_j) + e'\Sigma_{\mathbf{x}}^{-1}[np_h]^{-1}\sum_{j=1}^n(\delta_h^o(Y_j) - \delta_h(Y_j))u_j \\ &= [np_h]^{-1}\sum_{j=1}^n(\delta_h^o(Y_j) - \delta_h(Y_j))e'\Sigma_{\mathbf{x}}^{-1}u_j. \end{aligned}$$

Due to the independence between u_j's, the conditional variance of this term is equal to $[np_h]^{-2}\sum_{j=1}^{n}(\delta_h^o(Y_j)-\delta_h(Y_j))^2\mathrm{Var}\,(e'\Sigma_{\mathbf{x}}^{-1}u_j|Y_j)$, which is no greater than

$$\begin{aligned}&[np_h]^{-2}\sum_{j=1}^{n}(\delta_h^o(Y_j)-\delta_h(Y_j))^2\mathrm{Var}\,(e'\Sigma_{\mathbf{x}}^{-1}\mathbf{x}_j)\\ &=e'\Sigma_{\mathbf{x}}^{-1}e\cdot[np_h]^{-2}\sum_{j=1}^{n}|\delta_h^o(Y_j)-\delta_h(Y_j)|=n^{-2}O_p(\sqrt{n}).\end{aligned}$$

This establishes (13), completing the proof of (12).

Remark B.1. We can verify that if the conditional variance $\mathrm{Var}\,(e'\Sigma_{\mathbf{x}}^{-1}\mathbf{x}|\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x})$ does not depend on $\beta_1^{\mathbf{x}}, \ldots, \beta_k'\mathbf{x}$, then Assumption (B.2) holds. This is the case when $\mathbf{x}$ is normal. In general, as one often does in the standard MLR application, a diagnostic check on (B.2) can be performed by plotting the residuals of MLR (after transformation) against the variate $e'\Sigma_{\mathbf{x}}^{-1}\mathbf{x}$. Note also that even without assumption (B.2), we can still use (6) to see that the asymptotic variance of $e'\hat{v}_i$ is equal to $n^{-1}\lambda_i^{*2}e'\Sigma_{\mathbf{x}}^{-1}E(r^{*2}(\mathbf{x}-E\mathbf{x})(\mathbf{x}-E\mathbf{x})')\Sigma_{\mathbf{x}}^{-1}e$.

References

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *J. Amer. Stat. Assoc.* **76**, 296-311.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26**, 211-252.

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. (with discussion). *J. Amer. Stat. Assoc.* **80**, 580-619.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees.* CA: Wadsworth.

Brillinger, D. R. (1977), The identification of a particular nonlinear time series system. *Biometrika* **64**, 509-515.

Brillinger, D. R. (1983). A generalized linear model with 'Gaussian' regressor variables. In *A Festschrift for Erich L. Lehmann* (Edited by Peter J. Bickel, Kjell A. Doksum and J. L. Hodges, Jr.), 97-114. Wadsworth, Belmont CA.

Brillinger, D. R. (1991). Comments on "Sliced inverse regression for dimension reduction". by K. C. Li. *J. Amer. Stat. Assoc.* **86**, 333-333.

Carroll, R. J. and Li, K. C. (1992). Measurement error regression with unknown link: Dimension reduction and data visualization. *J. Amer. Stat. Assoc.* **87**, 1040-1050.

Carroll, R. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman & Hall. London.

Carroll, R. J., Wu, C. F. and Ruppert, D. (1988). The effect of estimating weights in weighted least squares. *J. Amer. Stat. Assoc.* **83**, 1045-1054.

Chaudhuri, P., Huang, M. C., Loh, W. Y. and Yao, R. (1994). Piecewise-polynomial regression trees. *Statist. Sinica* **4**, 143-167.

Chen, H. (1991). Estimation of a projection-pursuit type regression model. *Ann. Statist.* **19**, 142-157.

Cook, R. D. (1994). On the interpretation of regression plots. *J. Amer. Stat. Assoc.* **89**, 177-189.

Cook, R. D., Hawkins, D. M. and Weisberg, S. (1992). Comparison of model misspecification diagnostics using residuals from least mean of squares and least median of squares fits. *J. Amer. Stat. Assoc.* **87**, 419-424.

Cook, R. D. and Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *J. Amer. Statist. Assoc.* **89**, 592-599.

Cook, R. D. and Weisberg, S. (1991). Comment on "Sliced inverse regression for dimension reduction". by K. C. Li. *J. Amer. Stat. Assoc.* **86**, 328-332.

Cook, R. D. and Wetzel, N. (1994). Exploring regression structure with graphics. (with discussion). *Test* **2**, 33-100.

de Leeuw, J. (1981). The GIFI-system of nonlinear multivariate analysis. In *Data Analysis and Informatics III* (Edited by E. Diday, M. Jambu, L. Lebart, J. Pages and R. Tomassone), 415-424. North Holland, Amsterdam.

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

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

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76**, 817-823.

Friedman, J. H. (1990). Multivariate adaptive regression splines. (with discussion). *Ann. Statist.* **19**, 1-141.

Gifi, A. (1990). *Nonlinear Multivariate Analysis.* John Wiley, Chichester.

Hall, P. (1989). On projection pursuit regression. *Ann. Statist.* **17**, 573-588.

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

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *J. Environmental Economics Management* **5**, 81-102.

Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data. (with discussion). *J. Amer. Statist. Assoc.* **79**, 302-320.

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

Huber, P. J. (1985). Projection pursuit. (with discussion). *Ann. Statist.* **13**, 435-525.

Knickerbocker, R. K., Wang, C. Y. and Carroll, R. J. (1992). Dimension reduction in a semiparametric regression model with errors in covariates. Manuscript.

Koyak, R. A. (1987). On measuring internal dependence in a set of random variables. *Ann. Statist.* **15**, 1215-1228.

Li, K. C. (1990). Data visualization with SIR: a transformation-based projection pursuit method. *UCLA Statist Ser.* **24**.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, K. C. (1992a). Uncertainty analysis for mathematical models with SIR. In *Probability and Statistics* (Edited by Z. P. Jiang, S. J. Yan, P. Cheng and R. Wu), 138-162. World Scientific, Singapore.

Li, K. C. (1992b). On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *J. Amer. Statist. Assoc.* **87**, 1025-1039.

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

Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009-1052.

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

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

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24**, 1053-1068.

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

Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan.

E-mail: cchen@stat.sinica.edu.tw

Department of Mathematics, University of California, Los Angeles, Los Angeles, CA 90024, U.S.A.

E-mail: kcli@math.ucla.edu

(Received August 1994; accepted March 1997)