Sliced Inverse Regression for Dimension Reduction

Ker-Chau Li

Journal of the American Statistical Association, Vol. 86, No. 414. (Jun., 1991), pp. 316-327.

Sliced Inverse Regression for Dimension Reduction

KER-CHAU LI*

Modern advances in computing power have greatly widened scientists' scope in gathering and investigating information from many variables, information which might have been ignored in the past. Yet to effectively scan a large pool of variables is not an easy task, although our ability to interact with data has been much enhanced by recent innovations in dynamic graphics. In this article, we propose a novel data-analytic tool, sliced inverse regression (SIR), for reducing the dimension of the input variable $\mathbf{x}$ without going through any parametric or nonparametric model-fitting process. This method explores the simplicity of the inverse view of regression; that is, instead of regressing the univariate output variable y against the multivariate $\mathbf{x}$, we regress $\mathbf{x}$ against y. Forward regression and inverse regression are connected by a theorem that motivates this method. The theoretical properties of SIR are investigated under a model of the form, $y = f(\beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x}, \epsilon)$, where the β_k's are the unknown row vectors. This model looks like a nonlinear regression, except for the crucial difference that the functional form of f is completely unknown. For effectively reducing the dimension, we need only to estimate the space [effective dimension reduction (e.d.r.) space] generated by the β_k's. This makes our goal different from the usual one in regression analysis, the estimation of all the regression coefficients. In fact, the β_k's themselves are not identifiable without a specific structural form on f. Our main theorem shows that under a suitable condition, if the distribution of $\mathbf{x}$ has been standardized to have the zero mean and the identity covariance, the inverse regression curve, $E(\mathbf{x} \mid y)$, will fall into the e.d.r. space. Hence a principal component analysis on the covariance matrix for the estimated inverse regression curve can be conducted to locate its main orientation, yielding our estimates for e.d.r. directions. Furthermore, we use a simple step function to estimate the inverse regression curve. No complicated smoothing is needed. SIR can be easily implemented on personal computers. By simulation, we demonstrate how SIR can effectively reduce the dimension of the input variable from, say, 10 to $K = 2$ for a data set with 400 observations. The spin-plot of y against the two projected variables obtained by SIR is found to mimic the spin-plot of y against the true directions very well. A chi-squared statistic is proposed to address the issue of whether or not a direction found by SIR is spurious.

KEY WORDS: Dynamic graphics; Principal component analysis; Projection pursuit.

1. INTRODUCTION

Regression analysis is a popular way of studying the relationship between a response variable y and its explanatory variable $\mathbf{x}$, a p-dimensional column vector. Quite often, a parametric model is used to guide the analysis. When the model is parsimonious, standard estimation techniques such as the maximum likelihood or the least squares method have proved to be successful in gathering information from the data.

In most applications, however, any parametric model is at best an approximation to the true one, and the search for an adequate model is not easy. When there are no persuasive models available, nonparametric regression techniques emerge as promising alternatives that offer the needed flexibility in modeling. A common theme of nonparametric regression is the idea of local smoothing, which explores only the continuity or differentiability property of the true regression function. The success of local smoothing hinges on the presence of sufficiently many data points around each point of interest in the design space to provide adequate information. For one-dimensional problems, many smoothing techniques are available (see Eubank 1988 for a comprehensive account).

As the dimension of $\mathbf{x}$ gets higher, however, the total number of observations needed for local smoothing escalates exponentially. Unless we have a gigantic sample, standard methods, such as kernel estimates or nearest-neighbor estimates, break down quickly because of the sparseness of the data points in any region of interest. To challenge the curse of dimensionality, one hope that statisticians may capitalize on is that interesting features of high-dimensional data are retrievable from low-dimensional projections. For regression problems, the following model describes such an ideal situation:

$$y = f(\beta_1\mathbf{x}, \beta_2\mathbf{x}, \ldots, \beta_K\mathbf{x}, \epsilon). \tag{1.1}$$

Here the β's are unknown row vectors, ϵ is independent of $\mathbf{x}$, and f is an arbitrary unknown function on $\mathbf{R}^{K+1}$.

When this model holds (cf. Remark 1.1), the projection of the p-dimensional explanatory variable $\mathbf{x}$ onto the K dimensional subspace, $(\beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x})'$, captures all we need to know about y. When K is small, we may achieve the goal of data reduction by estimating the β's efficiently. For convenience, we shall refer to any linear combination of the β's as an effective dimension-reduction (e.d.r.) direction, and to the linear space B generated by the β's as the e.d.r. space. More discussion on this model, e.d.r. directions, and the relation to other approaches is given in Section 2. Our main focus in this article is on the estimation of the e.d.r. directions, leaving questions such as how to estimate main features of f for further investigation. Intuitively speaking, after estimating the e.d.r. directions, standard smoothing techniques can be more successful because

* Ker-Chau Li is Professor, Division of Statistics, Department of Mathematics, University of California, Los Angeles, CA 90024. This research was supported in part by the National Science Foundation under grants DMS86-02018 and DMS89-02494. It has been a long time since I introduced SIR in talks given at Berkeley, Bell Labs, and Rutgers in 1985. I received many useful questions and suggestions from these audiences. I am indebted to Naihua Duan for stimulating more ideas for improvement, leading eventually to Duan and Li (1991). Peter Bickel brought semiparametric literature to my attention. Jan de Leeuw broadened my knowledge in areas of dimension reduction and multivariate analysis. Without Dennis Cook, who introduced XLISP-STAT to me, the appearance of Section 6.3 would have been further delayed. As a replacement for "slice regression" or "slicing regression," used earlier, the name SIR was suggested to me by Don Ylvisaker, who also helped me clear hurdles in publishing this article. Three referees and an associate editor offered nice suggestions for improving the presentation. Finally, I would like to thank David Brillinger, whose work has inspired me so much.

the dimension has been lowered (cf. Remark 1.2). On the other hand, during the exploratory stage of data analysis, one often wants to view the data directly. Many graphical tools are available (see, for instance, the special applications section on statistical graphics, introduced by Cleveland 1987), but plotting y against every combination of $\mathbf{x}$ within a reasonable amount of time is impossible. So, to use the scatterplot-matrix techniques (Carr et al. 1987), we often focus on coordinate variables only. Likewise, 3-D rotating plots (e.g., see Huber 1987) can handle only one two-dimensional projection of $\mathbf{x}$ at a time (the third dimension is reserved for y). Therefore, to take full advantage of modern graphical tools, guidance on how to select the projection directions is clearly called for. A good estimate of the e.d.r. directions can lead to a good view of the data. Section 6.3 demonstrates how sharp the view found by our method is.

Our method of estimating the e.d.r. directions is based on the idea of inverse regression. Instead of regressing y against $\mathbf{x}$ (forward regression) directly, we regress $\mathbf{x}$ against y (inverse regression). The immediate benefit for exchanging the roles of y and $\mathbf{x}$ is that we can side-step the dimensionality problem. This comes out because inverse regression can be carried out by regressing each coordinate of $\mathbf{x}$ against y. Thus we essentially deal with a one-dimension to one-dimension regression problem, rather than the high-dimensional forward regression.

The feasibility of finding the e.d.r. directions via inverse regression will become clear. As y varies, $E(\mathbf{x} \mid y)$ draws a curve, called the inverse regression curve, in $\mathbf{R}^p$. Under (1.1), however, this curve typically will hover around a K-dimensional affine subspace. At one extreme, as shown in Theorem 3.1 of Section 3, the inverse regression curve actually falls into a K-dimensional affine subspace determined by the e.d.r. directions, provided that the distribution of $\mathbf{x}$ satisfies (3.1). If we have standardized $\mathbf{x}$ to have mean 0 and the identity covariance, then this subspace coincides with the e.d.r. space. Elliptically symmetric distributions, including the normal distribution, satisfy condition (3.1).

Exploring the simplicity of inverse regression, we develop a simple algorithm, called sliced inverse regression (SIR), for estimating the e.d.r. directions. After standardizing $\mathbf{x}$, SIR proceeds with a crude estimate of the inverse regression curve $E(\mathbf{x} \mid y)$, which is the slice mean of $\mathbf{x}$ after slicing the range of y into several intervals and partitioning the whole data into several slices according to the y value. A principal component analysis is then applied to these slice means of $\mathbf{x}$, locating the most important K-dimensional subspace for tracking the inverse regression curve $E(\mathbf{x} \mid y)$. The output of SIR is these components after an affine retransformation back to the original scale.

In Section 5, under the design condition of (3.1), we show that SIR yields root n consistent estimates for the e.d.r. directions.

Besides offering estimates of e.d.r. directions, the outputs of SIR are themselves interesting descriptive statistics containing useful information about the inverse regression curve. In Section 7, we elaborate this point further and argue that the directions produced by SIR can be used to form variables (derivable from $\mathbf{x}$ linearly) that are most predictable from y. Thus for graphical purposes, even if the design condition (3.1) is not satisfied, SIR still suggests interesting directions for viewing the data.

As a sharp contrast to most nonparametric techniques that require intensive computation, SIR is very simple to implement. Moreover, the sampling property of SIR is easy to understand, another advantage over other methods. Thus it is possible to assess the effectiveness of SIR by using the companion output eigenvalues at the principal component analysis step (see Remark 5.1 in Section 5). These eigenvalues provide valuable information for assessing the number of components in the data (see Remark 5.2). Finally, selection of the number of slices for SIR is less crucial than selection of the smoothing parameter for typical nonparametric regression problems. We further illustrate these points by simulation in Section 6.

In view of these virtues, however, SIR is not intended to replace other computer-intensive methods. Rather it can be used as a simple tool to aid other methods; for instance, it provides a good initial estimate for many methods based on the forward regression viewpoint. Because of low computing cost, one should find it easy to incorporate SIR into most statistical packages.

Remark 1.1. All models are imperfect in some sense and (1.1) should be interpreted as an approximation to reality. However, there is a fundamental difference between this and other statistical models: (1.1) takes the weakest form for reflecting our hope that a low-dimensional projection of a high-dimensional regressor variable contains most of the information that can be gathered from a sample of a modest size. Equation (1.1) does not impose any structure on how the projected regressor variable affects the output variable. In addition, we may even vary K to reflect the degree of the anticipated dimension reduction. At $K = p$, (1.1) becomes a redundant assumption. By comparison, most regression models assume $K = 1$, with additional structures on f.

Remark 1.2. A philosophical point needs to be emphasized here: The estimation of the projection angles can be a more important statistical issue than the estimation of the structure of f itself. In fact, the structure of f is impossible to identify unless we have other scientific evidence beyond the data under study. One can obtain two different versions of f to represent the same joint distribution of y and $\mathbf{x}$ [cf. (2.1)]. Thus what we can estimate are at most statistical quantities, such as the conditional mean or quantiles of y given $\mathbf{x}$. On the other hand, during the early stages of data analysis, when one does not have a fixed objective in mind, the need for estimating such quantities is not as pressing as that for finding ways to simplify the data. Our formulation of estimating the e.d.r. directions is one way to address such a need in data analysis. After finding a good e.d.r. space, we can project data into this smaller space. We are then in a better position to identify what should be pursued further: model building, response surface estimation, cluster analysis, heteroscedasticity analysis, variable selection, or inspecting scatterplots (or spin-plots) for interesting fea-

tures. This approach toward data analysis is different from that in many other works. Projection pursuit regression (Friedman and Stuetzle 1981), ACE and additive models (Breiman and Friedman 1985; Stone 1986; Hastie and Tibshirani 1986), or partial splines (Chen 1988b; Cuzik 1987; Engle, Granger, Rice, and Rice 1986; Heckman 1986; Speckman 1987; Wahba 1986), for instance, seem to have singled out the approximation of the conditional mean of the output variable (or its transformation) as their primary goal. But dimension reduction in statistics has a wider scope than functional approximation. The concept of e.d.r. space and the method of SIR aim at this general purpose of dimension reduction.

After dimension reduction, if we want to estimate the response surface, for example, we can apply standard techniques in nonparametric regression to the projected variables (e.g. Li 1987 and the references therein). In addition, suitable conditional influence as in the Box–Cox transformation study (Hinkley and Runger 1984) may be valuable. Needless to say, the door is open for further serious work.

2. A MODEL FOR DIMENSION REDUCTION

Equation (1.1) describes an ideal situation where one can reduce the dimension of the explanatory variable from p to a smaller number K without losing any information. An equivalent version of (1.1) is: The conditional distribution of y given $\mathbf{x}$ depends on $\mathbf{x}$ only through the K dimensional variable $(\beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x})$. Thus conditional on $\beta_k\mathbf{x}$'s, y and $\mathbf{x}$ are independent; the perfectly reduced variable, $(\beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x})$, is seen to be as informative as the original $\mathbf{x}$ in predicting y.

Recall the terminology of e.d.r. direction and e.d.r. space from Section 1. Observe that by changing f suitably, (1.1) can be reparameterized by any set of K linearly independent e.d.r. directions. Another interpretation is that conditioning on $(\beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x})$ is equivalent to conditioning on any nondegenerate affine transformation of this vector. Thus it is the e.d.r. space B that can be identified; the individual vectors $\beta_1, \ldots, \beta_K$ are not themselves identifiable (unless further structural conditions on f are imposed).

Let $\Sigma_{\mathbf{xx}}$ be the covariance matrix of $\mathbf{x}$. Later on, we shall find it convenient to consider the standardized version of $\mathbf{x}$, $\mathbf{z} = \Sigma_{\mathbf{xx}}^{-1/2}[\mathbf{x} - E(\mathbf{x})]$. We may rewrite (1.1) as

$$y = f(\eta_1\mathbf{z}, \ldots, \eta_K\mathbf{z}, \epsilon), \tag{2.1}$$

where $\eta_k = \beta_k\Sigma_{\mathbf{xx}}^{1/2}$ $(k = 1, \ldots, K)$. We shall call any vector in the linear space generated by the η_k's a standardized e.d.r. direction.

We will discuss the relationship of our model to others subsequently.

First of all, it is fair to say that one-component models $(K = 1)$ prevail in the literature; for instance, the generalized linear model, the Box–Cox transformation model and its generalization (Box and Cox 1964; Bickel and Doksum 1981; Carroll and Ruppert 1984) and others. Brillinger (1977, 1983) derived a surprising result about the robustness of least squares estimation under a global misspecification of the link function. Li and Duan (1989) generalized Brillinger's result to a general class of maximum likelihood type estimators.

Turning to the multicomponent model $(K > 1)$, observe that the conditional expectation $E(y \mid \mathbf{x})$, the forward regression surface, takes the form $g(\beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x})$. From the forward regression viewpoint, after projecting $\mathbf{x}$ onto any K-dimensional subspace (with a basis, say, $b_1, \ldots, b_K$), it is possible to estimate the conditional expectation $E(y \mid b_k\mathbf{x}\text{'s})$ nonparametrically. The average conditional variance $E[\mathrm{var}(y \mid b_k\mathbf{x}\text{'s})]$ is minimized when the projection space coincides with the space of the e.d.r. directions. We are led to a variant of the projection pursuit method studied in Chen (in press), which estimates the e.d.r. space by globally searching for a best K-dimensional projection that minimizes a lack-of-fit measure based on the residual sum of squares. Large-sample results were derived, but the method appears to be highly computer-intensive because one has to worry not only about how to do a global search efficiently but also about how to do the multidimensional smoothing.

If the conditional expectation $E(y \mid \mathbf{x})$ takes the additivity form, $g_1(\beta_1\mathbf{x}) + \cdots + g_K(\beta_K\mathbf{x})$, then one may use the projection pursuit regression algorithm (PPR) as described in Friedman and Stuetzle (1981) to estimate the e.d.r. directions. Donoho and Johnstone (1989), Hall (1989), and Huber (1985) add more insight to PPR.

Another possible forward regression route to attack this problem is based on the observation that under (1.1), any slope vector, the derivative of y with respect to $\mathbf{x}$ at any point, falls within the e.d.r. space B. Thus if we can estimate the slope vectors well, we may apply a principal component analysis to the estimated slope vectors to find the e.d.r. directions. The main difficulty for this approach, however, is the estimation of the derivatives for high dimensional $\mathbf{x}$.

Many recent works are related to data reduction. A short and incomplete list includes the correspondence analysis approach (e.g., van Rijckevorsel and de Leeuw 1988), classification trees (e.g., Breiman, Friedman, Olshen, and Stone 1984; Loh and Vanichsetakul 1988), ACE and additive models (Breiman and Friedman 1985; Koyak 1987; Stone 1986), and partial spline models (Chen 1988; Cuzick 1987; Engle et al. 1983; Heckman 1986; Speckman 1987; Wahba 1986), and projection pursuit density estimation (e.g., Diaconis and Freedman 1984; Friedman 1987; Huber 1985).

We conclude this section by discussing the question of how to evaluate the effectiveness of an estimated e.d.r. direction. An obvious criterion is based on the squared Euclidean distance between the estimated e.d.r. direction b (normalized to have the unitary length) and the true e.d.r. space B. This criterion, however, is not invariant under scale change or affine transformation of $\mathbf{x}$. We prefer an affine invariant criterion,

$$R^2(b) = \max_{\beta \in B} \frac{(b\Sigma_{\mathbf{xx}}\beta')^2}{b\Sigma_{\mathbf{xx}}b \cdot \beta\Sigma_{\mathbf{xx}}\beta'}, \tag{2.2}$$

the squared multiple correlation coefficient between the projected variable $b\mathbf{x}$ and the ideally reduced variables $\beta_1\mathbf{x}$, $\ldots$, $\beta_K\mathbf{x}$. For a collection of K estimated directions $b_1, \ldots, b_K$ generating a linear subspace $\hat{B}$, we use the squared trace

correlation, denoted by $\mathbf{R}^2(\hat{B})$, as our criterion: the average of the squared canonical correlation coefficients between $b_1\mathbf{x}, \ldots, b_K\mathbf{x}$ and $\beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x}$ (Hooper 1959). It is also reasonable to replace Σ_{xx} by the sample covariance matrix in our definition of the criteria.

3. THE INVERSE REGRESSION CURVE

Consider the trajectory of the inverse regression curve $E(\mathbf{x} \mid y)$ as y varies. The center of this curve is located at $E(E(\mathbf{x} \mid y)) = E(\mathbf{x})$. In general, the centered inverse regression curve, $E(\mathbf{x} \mid y) - E(\mathbf{x})$ is a p-dimensional curve in $\mathbf{R}^p$. We shall see that it lies on a K-dimensional subspace, however, with the following condition on the design distribution:

Condition 3.1. For any b in $\mathbf{R}^p$, the conditional expectation $E(b\mathbf{x} \mid \beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x})$ is linear in $\beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x}$; that is, for some constants $c_0, c_1, \ldots, c_K$, $E(b\mathbf{x} \mid \beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x}) = c_0 + c_1\beta_1\mathbf{x} + \cdots + c_K\beta_K\mathbf{x}$.

This condition is satisfied when the distribution of $\mathbf{x}$ is elliptically symmetric (e.g., the normal distribution). More discussion of this condition is given in Remark 3.3. The following theorem will be proved in the Appendix.

Theorem 3.1. Under the conditions (1.1) and (3.1), the centered inverse regression curve $E(\mathbf{x} \mid y) - E(\mathbf{x})$ is contained in the linear subspace spanned by $\beta_k\Sigma_{xx}$ ($k = 1, \ldots, K$), where Σ_{xx} denotes the covariance matrix of $\mathbf{x}$.

Corollary 3.1. Assume that $\mathbf{x}$ has been standardized to $\mathbf{z}$. Then under (2.1) and (3.1), the standardized inverse regression curve $E(\mathbf{z} \mid y)$ is contained in the linear space generated by the standardized e.d.r. directions $\eta_1, \ldots, \eta_K$.

An important consequence of this corollary is that the covariance matrix $\operatorname{cov}[E(\mathbf{z} \mid y)]$ is degenerate in any direction orthogonal to the η_k's. We see, therefore, that the eigenvectors, $\eta_k (k = 1, \ldots, K)$, associated with the largest K eigenvalues of $\operatorname{cov}[E(\mathbf{z} \mid y)]$ are the standardized e.d.r. directions. Transforming back to the original scale, $\eta_k\Sigma^{-1/2}$ ($k = 1, \ldots, K$) are in the e.d.r. space. This leads to the SIR algorithm of the next section.

Remark 3.1. Conditional covariance $\operatorname{cov}(\mathbf{z} \mid y)$ can also reveal valuable clues for finding the standardized e.d.r. directions. To see this, simply observe the identity

$$E[\operatorname{cov}(\mathbf{z} \mid y)] = \operatorname{cov} \mathbf{z} - \operatorname{cov}[E(\mathbf{z} \mid y)] = I - \operatorname{cov}[E(\mathbf{z} \mid y)].$$

Therefore, after an eigenvalue decomposition of $E[\operatorname{cov}(\mathbf{z} \mid y)]$, we may find the standardized e.d.r. directions from the eigenvectors associated with the smallest K eigenvalues. The estimation of $E[\operatorname{cov}(\mathbf{z} \mid y)]$ is not difficult; see Remark 5.3.

Remark 3.2. Regression models are usually formed by decomposing the joint distribution of y and $\mathbf{x}$ as $h(y \mid \mathbf{x})k(\mathbf{x})$ and modeling $h(y \mid \mathbf{x})$. This is the forward view of regression. The inverse view of regression factorizes the joint density as $h(\mathbf{x} \mid y)k(y)$ and models $h(\mathbf{x} \mid y)$. Important cases of inverse formation include discriminant analysis (with logistic regression as the counterpart from the forward view). SIR is itself meaningful from the inverse view of modeling. It is interesting to observe that we may consider y as a parameter with an empirical Bayes prior.

Remark 3.3. Condition (3.1) seems to impose a stringent requirement on the distribution of $\mathbf{x}$. One implication is that, at the stage of data collection, unless the functional form of the response surface is known, we had best design the experiment so that the distribution of $\mathbf{x}$ will not blatantly violate elliptic symmetry. For example, rotatable designs, advocated by George Box (see, e.g., Box and Draper 1987) in the response surface literature, deserve to be studied more closely in the future. On the other hand, after data collection, it would help the analysis if closer examination of the distribution of $\mathbf{x}$ can be made so that outliers can be removed or clusters can be separated before analysis.

An interesting extension of Corollary 3.1 will be to quantify how far away from the standardized e.d.r. space the inverse regression curve $E(\mathbf{z} \mid y)$ is when (3.1) is mildly violated. If the projection of $E(\mathbf{z} \mid y)$ on the orthogonal complement of the standardized e.d.r. space is small, then the directions picked up by the principal component analysis on $\operatorname{cov}[E(\mathbf{z} \mid y)]$ will still be close to the standardized e.d.r. directions. The situation is similar to that in Brillinger (1977, 1983) where consistency of least squares in estimating β_1 for one-component models under (3.1) is proved. Furthermore, empirical evidence was reported indicating that his result is not sensitive to violation of (3.1). A comprehensive account of this robustness issue for the least squares and other commonly used regression estimates is given in Li and Duan (1989).

After the first version of our article was written, this design robustness issue was further addressed in three articles. First, a bound to bias in estimation was obtained in Duan and Li (in press) for $K = 1$. Moreover, Li (1989) argued that for most directions b, we can expect the linearity in (3.1) to hold approximately, borrowing a powerful result from Diaconis and Freedman (1984) where they showed that most low-dimension projections of a high-dimension data cloud are close to being normal. Li (1989) also demonstrated how SIR may find the directions that violate (3.1) most seriously. Li (1990b) extended the discussion to a framework for the uncertainty analysis of mathematical models.

4. SLICED INVERSE REGRESSION

A scheme for sliced inverse regression operates on the data $(y_i, \mathbf{x}_i)$ $(i = 1, \ldots, n)$, in the following way:

1. Standardize $\mathbf{x}$ by an affine transformation to get $\tilde{\mathbf{x}}_i = \hat{\Sigma}_{xx}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$ $(i = 1, \ldots, n)$, where $\hat{\Sigma}_{xx}$ and $\bar{\mathbf{x}}$ are the sample covariance matrix and sample mean of $\mathbf{x}$ respectively.
2. Divide range of y into H slices, $I_1, \ldots, I_H$; let the proportion of the y_i that falls in slice h be $\hat{p}_h$; that is $\hat{p}_h = (1/n)\sum_{i=1}^n \delta_h(y_i)$, where $\delta_h(y_i)$ takes the values 0 or 1 depending on whether y_i falls into the hth slice I_h or not.
3. Within each slice, compute the sample mean of the $\tilde{\mathbf{x}}_i$'s, denoted by $\hat{m}_h$ $(h = 1, \ldots, H)$, so that $\hat{m}_h = (1/n\hat{p}_h)\sum_{y_i \in I_h} \tilde{x}_i$.
4. Conduct a (weighted) principal component analysis

for the data $\hat{m}_h$ ($h = 1, \ldots, H$) in the following way: Form the weighted covariance matrix $\hat{V} = \sum_{h=1}^{H} \hat{p}_h \hat{m}_h \hat{m}_h'$, then find the eigenvalues and the eigenvectors for $\hat{V}$.

5. Let the K largest eigenvectors (row vectors) be $\hat{\eta}_k$ ($k = 1, \ldots, K$). Output $\hat{\beta}_k = \hat{\eta}_k \hat{\Sigma}_{xx}^{-1/2}$ ($k = 1, \ldots, K$).

Steps 2 and 3 produce a crude estimate of the standardized inverse regression curve $E(\mathbf{z} \mid y)$. Although it is feasible to use more sophisticated nonparametric regression methods such as kernel, nearest neighbor, or smoothing splines to yield a better estimate of the inverse regression curve, we advocate only the method of slicing due to its simplicity. Since we only need the main orientation (but not any other detailed aspects) of the estimated curve, possible gains due to smoothing are not likely to be substantial.

The weighting adjustment for principal component analysis in Step 4 takes care of the case where there may be unequal sample sizes in different slices. The first K components locate the most important subspace to track the standardized inverse regression curve $E(\mathbf{z} \mid y)$. Finally, Step 5 retransforms the scale back to the original one. Thus $\hat{\beta}_k$'s can be used as estimates of the e.d.r. directions and the e.d.r. space B is estimated by $\hat{B}$, the space generated by the $\hat{\beta}_k$'s.

A few remarks about the actual implementation are in order.

Remark 4.1. It is not necessary to transform each individual $\mathbf{x}_i$ to $\tilde{\mathbf{x}}_i$. All we need is to transform the slice means before conducting the principal component analysis to save computing time. Let $\hat{\Sigma}_1$ be $\sum_{h=1}^{H} \hat{p}_h(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})'$, where $\bar{\mathbf{x}}_h$ denotes the sample mean of the $\mathbf{x}_i$'s in the hth slice. Then the $\hat{\beta}_k$'s are just the eigenvectors for the eigenvalue decomposition of $\hat{\Sigma}_1$ with respect to $\hat{\Sigma}_{xx}$. Using the terminology of MANOVA (e.g., Mardia, Kent, and Bibby 1979, chap. 12), $\hat{\Sigma}_1$ describes the between-slice variation.

Remark 4.2. The range for each slice may be set to have equal length; but in section 6, we prefer to allow it to vary so that the number of observations in each slice can be as close to each other as possible.

Remark 4.3. The choice of the number of slices may affect the asymptotic variance of the output estimate. However, the difference is not significant for practical sample sizes in our simulation study. This issue here is less critical than the choice of a smoothing parameter in nonparametric regression. Theoretically an inappropriate choice of smoothing parameter in nonparametric regression or density estimation may lead to a slower rate of convergence, while for our case we can still have root n consistency no matter how H is chosen; see Remark 5.3 in Section 5. For a comprehensive treatment of adaptive choice of smoothing parameter in nonparametric regression, see Li (1987), and Härdle, Hall, and Marron (1988).

Remark 4.4. When standardizing $\mathbf{x}$, it is not necessary to base the affine transformation on the sample mean and sample covariance matrix. Some robust versions of them may be preferable (see Donoho, Johnstone, Rousseeuw, and Stahel 1985; Fill and Johnstone 1984; and Li and Chen 1985). At least we should downweight or cut out those influential design points. But this issue is probably less crucial and is relatively easy to handle because we are dealing with the design points that are under our control (even in the observational study, we may screen out some bad design points; if the percentage of the remaining points is high enough we can still have a good analysis). Note that the efficiency of the affine transformation is not the main concern because we need only a consistent estimate to make SIR work.

Remark 4.5. If the standardized inverse regression curve falls within a proper subspace of the standardized e.d.r. space, then SIR cannot recover all e.d.r. directions. For instance, if $y = g(\beta_1 \mathbf{x}) + \epsilon$ for some symmetric function g, and $\beta_1 \mathbf{x}$ is also symmetric about 0, then $E(\mathbf{x} \mid y)$ equals 0, and $\hat{\beta}_1$ is a poor estimate of β_1. For handling such cases, one approach is to explore higher conditional moments of $\mathbf{x}$ given y. For instance, if $\mathbf{x}$ is normal, then for any direction $b\mathbf{x}$ orthogonal to the $\beta_k \mathbf{x}$'s, we see that $\text{var}(b\mathbf{x} \mid y)$ remains invariant as y changes. In particular, for standardized $\mathbf{x}$, the eigenvectors of $\text{cov}(\mathbf{x} \mid y \in I_h)$ with eigenvalues different from 1 are in the e.d.r. space. Thus if an eigenvalue decomposition on the sample covariance of the $\tilde{\mathbf{x}}_i$'s, denoted as COV_h, for each slice h is conducted, then we may combine those eigenvectors with eigenvalues significantly different from 1 from each slice in a suitable way to estimate β_k's. Details on ways of combination are under investigation. It is not always necessary to conduct the eigenvalue decomposition separately for each slice, however. For instance, one may treat it as an approximation problem of fitting each $\text{COV}_h - I$ separately by a nonnegative definite matrix of rank K with the constraint that the fitted matrices have a common range. Another second moment method and a method based on the notion of principal Hessian directions were suggested in Li (1989, 1990a). More recently, the author also learned that Cook and Weisberg have independently obtained some good estimates based on the second moments.

5. SAMPLING PROPERTIES

In this section we present a brief argument to show how the output of SIR provides root n consistent estimates for the e.d.r. directions.

Let $p_h = \Pr\{y \in I_h\}$ and $\mathbf{m}_h = E(\mathbf{z} \mid y \in I_h)$, where $\mathbf{z}$ stands for the standardized $\mathbf{x}$, as defined in Section 2. Elementary probability theory shows that $\hat{m}_h$ converges to $\mathbf{m}_h$ at rate $n^{-1/2}$. Let V be the matrix $\sum_{h=1}^{H} p_h \mathbf{m}_h \mathbf{m}_h'$. Clearly the weighted covariance $\hat{V}$ in Step (4) of SIR converges to V at the root n rate. Consequently, the eigenvectors of $\hat{V}$, $\hat{\eta}_k$ ($k = 1, \ldots, K$), converge to the corresponding eigenvectors for V at the root n rate. Now we use Corollary 3.1 and the simple identity $\mathbf{m}_h = E[E(\mathbf{z} \mid y) \mid y \in I_h]$ to see that the first K eigenvectors of V fall in the standardized e.d.r. space. Since $\hat{\Sigma}_{xx}^{-1/2}$ converges to $\Sigma_{xx}^{-1/2}$, we see that each $\hat{\beta}_k$ converges to an e.d.r. direction at rate root n.

The case where the range of each slice varies in order to ensure an even distribution of observations is related to the following choice of intervals:

$$I_h = (F_y^{-1}((h-1)/H), F_y^{-1}(h/H)),$$

where $F_y(\cdot)$ is the cdf of y. The root n consistency result still holds.

Remark 5.1. It is possible to establish the asymptotic normality of the $\hat{\beta}_k$'s and to calculate the asymptotic covariance matrices using the delta method as in Mallows (1961) or Tyler (1981). In the Appendix we show how to approximate the expectation of $R^2(\hat{B})$, the squared trace correlation between $\beta_k\mathbf{x}$'s and $\hat{\beta}_k\mathbf{x}$'s (see the last paragraph of section 2). For the normal $\mathbf{x}$, we have the following simple approximation:

$$E[R^2(\hat{B})] = 1 - \frac{p-K}{n}\left(-1 + \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\lambda_k}\right) + o\left(\frac{1}{n}\right), \quad (5.1)$$

where λ_k is the kth eigenvalue of V. A crude estimate of this quantity is given by substituting the kth largest eigenvalue of $\hat{V}$ for λ_k.

Remark 5.2. To be really successful in picking up all K dimensions for reduction, the inverse regression curve cannot be too straight. In other words, the first K eigenvalues for V must be significantly different from zero compared to the sampling error. This can be checked by the companion output eigenvalues of $\hat{V}$ in Step (4) of the SIR. The asymptotic distribution of the average of the smallest $p - K$ eigenvalues, denoted by $\bar{\lambda}_{(p-K)}$, for $\hat{V}$ can be derived based on perturbation theory for finite-dimensional spaces (Kato 1976, chapter 2). For normal $\mathbf{x}$, we have the following result.

Theorem 5.1. If $\mathbf{x}$ is normally distributed, then $n(p - K)\bar{\lambda}_{(p-K)}$ follows a χ^2 distribution with $(p - K)(H - K - 1)$ df asymptotically.

We may use this result to give a conservative assessment of the number of components in the model. Thus if the rescaled $\bar{\lambda}_{(p-k)}$ is larger than the corresponding χ^2 value (say the 95th percentile), then we may infer that there are at least $k + 1$ (significant) components in the model. For other elliptically symmetric distributions, the result is more complicated. Although it is possible to estimate the asymptotic distribution using some version of the bootstrap method, we feel it is good enough to use the normal case result as a guideline to keep our procedure as simple as possible. An outline for the proof of the asymptotic result discussed here is given in the Appendix. A referee pointed out the similarity between this result and the result of a likelihood ratio test in MANOVA (Mardia et al. 1979, p. 342; further connection between SIR and sec. 12.5.4 of that reference can also be drawn). We remind the reader, however, that the fundamental assumption about the error distribution in MANOVA is not satisfied for our case. The conditional distribution of $\mathbf{x}$ given y is not normal, even if $\mathbf{x}$ is normal unconditionally. Furthermore, the conditional variance of $\mathbf{x}$ given $y \in I_h$ depends on h. Hence one has to be very careful if linking of SIR with MANOVA is desired.

Remark 5.3. Can SIR still yield reasonable estimates if the number of slices increases too fast and the number of observations in each slice is too small for $\hat{m}_h$ to consistently estimate $\mathbf{m}_h$? Remark 3.1 offers an answer. First, we see that the following is one way to estimate $E[\text{cov}(\mathbf{z} \mid y)]$:

(a) Introduce a large number of slices for partitioning the range of y.

(b) Within each slice, form the sample covariance of $\tilde{\mathbf{x}}_i$'s that fall into that slice.

(c) Form an average of the estimated conditional covariances of (b).

Intuitively, in order to get rid of the bias for estimating the conditional variance $\text{cov}(\mathbf{z} \mid y)$ for each y in (b), we hope that the range of each slice will converge to 0, so that only local points will contribute to the estimation. But when the number of slices is too large, the sampling variance in each estimate of $\text{cov}(\mathbf{z} \mid y)$ may not diminish, even for large n. Fortunately, the averaging process of (c) will stabilize the final estimate by the law of large numbers. As a matter of fact, even if the slice number is $n/2$, so that each slice contains only two observations, the resulting estimate will still be root n consistent.

The interesting connection of this estimate of $E[\text{cov}(\mathbf{z} \mid y)]$ to SIR is that this estimate is proportional to $I - \hat{V}$ because of the sample version of the identity given in Remark 3.1. Because of this conjugate relationship, a principal component analysis on the above estimate of $E[\text{cov}(\mathbf{z} \mid y)]$ for the smallest K components is equivalent to a principal component analysis on $\hat{V}$ for the largest K components. This explains why a large number of slices may still work, bolstering our earlier claim that the selection of H is not as crucial as the choice of a smoothing parameter in most nonparametric regression or density estimation problems.

Remark 5.4. It is interesting to study the asymptotic behavior of the SIR estimate when both the sample size and the dimension p of $\mathbf{x}$ increase simultaneously, but the number of components K and the number of slices H are kept fixed. One can see that a sufficient condition for the $\hat{\eta}_k$ to converge to η_k (in the sense that the angle between the two converges to 0) in probability is that the maximum singular value for $\hat{V} - V$ converges to 0 in probability. When the eigenvalues of V are bounded away from 0 and infinity as n increases, the above sufficient condition is implied by the condition that p/n converges to 0 in such a way that the difference between the maximum eigenvalue of the sample covariance $\hat{\Sigma}_{\mathbf{xx}}$ and that of $\Sigma_{\mathbf{xx}}$ converges to 0. This shows the potential of SIR for handling high-dimensional data. Asymptotic settings that allow p to increase are more appropriate in reflecting the situations where dimension-reduction techniques are called for. Diaconis and Freedman (1984) illustrated this well. See also Portnoy (1985) and references therein for the context of robust regression.

6. SIMULATION RESULTS

To demonstrate how SIR works, we have conducted simulation studies, and some of the results are reported here. The first subsection describes the behavior of the estimates; the second subsection discusses the eigenvalues and their role in estimation; and the third subsection demonstrates the graphical aspect of the e.d.r. direction estimation.

6.1. Behavior of the SIR estimates

First we use the linear model

$$y = x_1 + x_2 + x_3 + x_4 + 0x_5 + \epsilon \qquad (6.1)$$

to generate $n = 100$ data points. The dimension p equals 5 and the x_i's and ϵ are independent, with the standard normal distribution. There is only one component in this model, $K = 1$, and any vector proportional to $\beta = (1, 1, 1, 1, 0)$ is an e.d.r. direction. Table 1 gives the mean and the standard deviation (in parentheses) of $\hat{\beta}_1$ of SIR for $H = 5$, 10, and 20, after 100 replicates. Here we have to standardize the length and the sign of $\hat{\beta}_1$.

As we see from this table, the estimates are very good. The means are all quite close to the normalized target (.5, .5, .5, .5, 0). On the other hand, the least squares estimate for each coordinate of β, based on the correct linear model, has the standard deviation $1/\sqrt{n} = .1$. Since the target vector is half of β, the value .05 can be used as a benchmark for comparing the standard deviations of the SIR estimates. We also see that the performance of SIR is not sensitive to the number of slices.

Turning to the multicomponent case, we shall concentrate on the case $K = 2$. Two models are studied:

$$y = x_1(x_1 + x_2 + 1) + \sigma \cdot \epsilon \qquad (6.2)$$

and

$$y = \frac{x_1}{0.5 + (x_2 + 1.5)^2} + \sigma \cdot \epsilon. \qquad (6.3)$$

In addition to ϵ, x_1, x_2, we also generate $x_3, \ldots, x_p$, all variables being independent and following the standard normal distribution. We take $p = 10$ together with $\sigma = .5$ and $\sigma = 1$. The sample size is set at $n = 400$. The true e.d.r. directions are the vectors in the plane generated by (1, 0, ..., 0) and (0, 1, 0, ..., 0). The first two components of SIR will be used as estimates of e.d.r. directions. Recall the performance measure $R^2(\cdot)$ from Section 2. With the number of slices H set at 5, 10, and 20 respectively, Tables 2 and 3 report the mean and the standard deviation of $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ after 100 replicates.

For both models, despite the change in the noise level, the first component is very close to the e.d.r. space as the R^2 values hover in the neighborhood of 90%. The second component is more sensitive to the noise level. But even for the high noise level case, the sample correlation between the projected one-dimensional variable $\hat{\beta}_2\mathbf{x}$ and the perfectly reduced data, the square root of R^2, is still strong (above .7) on the average. Again, the number of slices has only minor effects on the results. It is interesting to observe that SIR is doing better for the rational function model (6.3) than for the quadratic model (6.2), despite the fact that the strength of the signal as measured by the standard deviation of $E(y \mid \mathbf{x})$ is weaker for the rational function model [about .8 for (6.3) vs. about 2.0 for (6.2)]. The key to the success of SIR hinges not on the signal-to-noise ratio, but on the eigenvalues of $\mathrm{cov}(E(\mathbf{z} \mid y))$.

We do not report the average for $R^2(\hat{B})$, because it is very close to the average of $1/2[R^2(\hat{\beta}_1) + R^2(\hat{\beta}_2)]$.

Table 1. Mean and Standard Deviation of $\hat{\beta}_1 = (\hat{\beta}_{11}, \ldots, \hat{\beta}_{15})$ for the linear model (6.1), n = 100; the Target is (.5, .5, .5, .5, 0)*

H	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$
5	.505	.498	.494	.488	.002
	(.052)	(.049)	(.056)	(.056)	(.066)
10	.502	.500	.492	.491	.001
	(.046)	(.045)	(.055)	(.049)	(.060)
20	.500	.502	.497	.487	−.003
	(.048)	(.046)	(.053)	(.054)	(.060)

*Numbers in parentheses represent standard deviations.

Table 2. Mean and Standard Deviation of $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ for the Quadratic Model (6.2), p = 10, n = 400

	σ = 0.5		σ = 1	
H	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$
5	.91	.75	.88	.52
	(.05)	(.15)	(.07)	(.21)
10	.92	.80	.89	.55
	(.04)	(.13)	(.08)	(.24)
20	.93	.77	.88	.49
	(.04)	(.15)	(.08)	(.26)

See Table 1 note.

6.2 Eigenvalues

How many components are there in the data? Perhaps this question is too ambitious to ask. But the companion output eigenvalues at step (4) of SIR do provide us with valuable information for a more practical question: Is an estimated component real or spurious?

Table 4 gives the empirical quantiles and the mean of $\bar{\lambda}_{(8)}$, $\bar{\lambda}_{(9)}$, and $\bar{\lambda}_{(10)}$ for the same 100 replicates used in obtaining the columns of $H = 10$ in Tables 2 and 3 (the conclusions are similar for other H's). For $\bar{\lambda}_{(8)}$, the numbers are close to the rescaled χ^2 values, as anticipated by Theorem 5.1. Thus guided by χ^2, we will not often falsely conclude that the third component is real (or mistakenly claim that there are more than two components in the data).

Turning to $\bar{\lambda}_{(9)}$, we expect the numbers to be larger than those given by using the rescaled χ^2 that falsely assumes only one component in the model. For the rational function model with $\sigma = .5$, this is clearly so, as we see that the 1% quantile of $\bar{\lambda}_{(9)}$ is close to the 99% quantile of the rescaled χ^2. Thus in this case we correctly infer that there are at least 2 components in the model in each of the 100 replicates. As confirmed by the corresponding $R^2(\hat{\beta}_2)$ reported in Table 3, a high value of $\bar{\lambda}_{(9)}$ leads to good performance of $\hat{\beta}_2$ as an e.d.r. direction. On the other hand, the distribution of $\bar{\lambda}_{(9)}$ for the quadratic model with $\sigma = 1$

Table 3. Mean and Standard Deviation of $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ for the Rational Function Model (6.3), p = 10, n = 400

	σ = 0.5		σ = 1	
H	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$
5	.96	.83	.89	.51
	(.02)	(.08)	(.06)	(.23)
10	.96	.88	.90	.56
	(.02)	(.06)	(.06)	(.23)
20	.96	.89	.90	.53
	(.02)	(.06)	(.06)	(.24)

See Table 1 note.

Table 4. Sample Quantiles and Means of $\bar{\lambda}_{(8)}$, $\bar{\lambda}_{(9)}$, and $\bar{\lambda}_{(10)}$ for the 100 Replicates Used in Obtaining the Columns of $H = 10$ in Tables 2 and 3

	Model	1%	5%	10%	25%	50%	75%	90%	95%	99%	Mean
$10\bar{\lambda}_{(8)}$	quadratic $\sigma = 1$	.10	.12	.13	.14	.16	.18	.20	.20	.24	.16
	quadratic $\sigma = .5$	.09	.12	.13	.15	.16	.18	.20	.22	.27	.17
	rational function $\sigma = 1$	.09	.11	.13	.14	.16	.19	.20	.21	.27	.16
	rational function $\sigma = .5$	.13	.13	.14	.16	.18	.20	.23	.24	.27	.18
$\frac{1}{320}\chi^2_{56}$		.11	.12	.13	.15	.17	.20	.22	.23	.26	.175
$10\bar{\lambda}_{(9)}$	quadratic $\sigma = 1$	.17	.18	.19	.22	.24	.27	.29	.32	.33	.24
	quadratic $\sigma = .5$	.19	.22	.24	.27	.30	.33	.35	.37	.43	.30
	rational function $\sigma = 1$	.16	.18	.20	.22	.24	.27	.30	.32	.35	.25
	rational function $\sigma = .5$	.28	.29	.30	.34	.36	.40	.43	.46	.53	.37
$\frac{1}{360}\chi^2_{72}$		.13	.15	.16	.18	.20	.22	.24	.26	.29	.20
$10\bar{\lambda}_{(10)}$	quadratic $\sigma = 1$	.28	.33	.34	.38	.42	.47	.53	.55	.58	.43
	quadratic $\sigma = .5$	.39	.43	.45	.49	.53	.57	.64	.66	.70	.54
	rational function $\sigma = 1$	.34	.36	.38	.40	.43	.49	.51	.55	.61	.44
	rational function $\sigma = .5$	.58	.63	.65	.69	.74	.79	.82	.85	.90	.74
$\frac{1}{400}\chi^2_{90}$		.15	.17	.18	.20	.22	.25	.27	.28	.31	.225

shows a substantial overlap with the rescaled χ^2. This is reflected in the relatively lower average and higher standard deviation of $R^2(\hat{\beta}_2)$ in Table 2. But a positive point is that by comparing $\bar{\lambda}_{(9)}$ with the rescaled χ^2, we realize that our data do not strongly support the claim that the estimated second component is real.

Finally, $\bar{\lambda}_{(10)}$ is well above the associated χ^2, assuring the high average and the low standard deviation of $R^2(\hat{\beta}_1)$ in all cases.

6.3. Graphics

We shall demonstrate how effective the estimated e.d.r. directions for (6.3) can be when used to view the data via spinning plots. For comparison, we also present the best view of the data, y against x_1, x_2. The best view is only possible in simulation study. First, Figure 1 shows the picture of the response surface in (6.3).

The rest of the study is carried out using XLISP-STAT (Tierney 1989). We set $p = 10$ and $\sigma = .5$ to generate $n = 400$ cases according to (6.3). The best view of the data is given in Figure 2, a–d, which shows four different angles from which to view the plot as we rotate it along the y axis every 45°. We then run SIR on the generated data with $H = 5$ and 30 and plot y against $\hat{\beta}_1\mathbf{x}$, $\hat{\beta}_2\mathbf{x}$ using the spin-plot command in XLISP-STAT (Figures 3 and 4). Evidently, SIR yields a very sharp view of the data. We also see that the choice of H has very little visual effect, confirming our theoretical argument given in Remark 4.3 and Remark 5.3.

7. DESCRIPTIVE STATISTICS AND SIR

We conclude this article by arguing that besides offering e.d.r. estimates, SIR gives useful descriptive statistics for cases in observational studies where we may think of $\mathbf{x}$ as the dependent variable and y as the independent variable.

According to the interpretation following (2.1), the forward view of data reduction aims at seeking a K-dimensional variable (derivable from $\mathbf{x}$ linearly) that predicts y most effectively. Now, reversing the role of y and $\mathbf{x}$, let us first ask which one-dimensional variable (derivable from $\mathbf{x}$ linearly) is most predictable from y. This question was asked and answered by Hotelling (1935), with the restriction that the prediction rules must be linear. Hotelling's solution, for a multivariate y, leads to the analysis of canonical correlation.

Without the linearity constraint on the prediction rules, for a variable $b\mathbf{x}$, the best prediction (under the squared error loss) is given by $E(b\mathbf{x} \mid y)$, a nonlinear function of y in general. Thus the most predictable variable is the one which maximizes

$$\frac{\operatorname{var}(E[b\mathbf{x} \mid y])}{\operatorname{var}(b\mathbf{x})} = \frac{(b\Sigma_{\mathbf{xx}}^{1/2})\operatorname{cov}[E(\mathbf{z} \mid y)](b\Sigma_{\mathbf{xx}}^{1/2})'}{\|b\Sigma_{\mathbf{xx}}^{1/2}\|^2},$$

where $\mathbf{z}$ is the standardized $\mathbf{x}$ as defined before. Clearly, the solution is given by $\eta_1\Sigma_{\mathbf{xx}}^{-1/2}$, where η_1 denotes the larg-

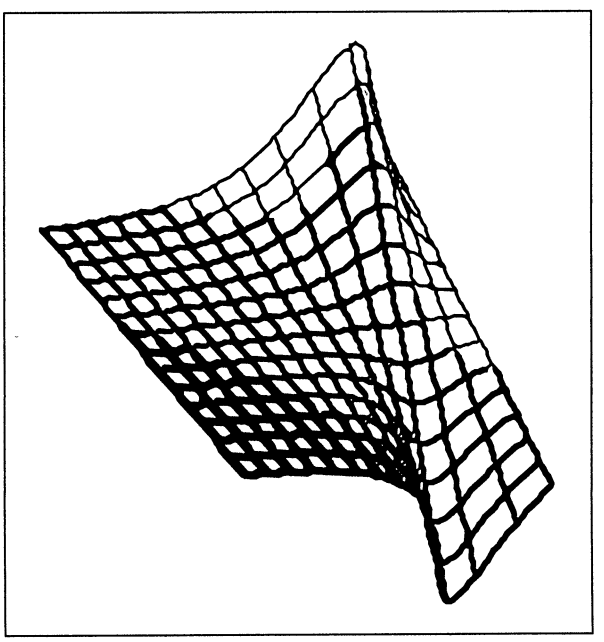

Figure 1. Response Surface of (6.3).

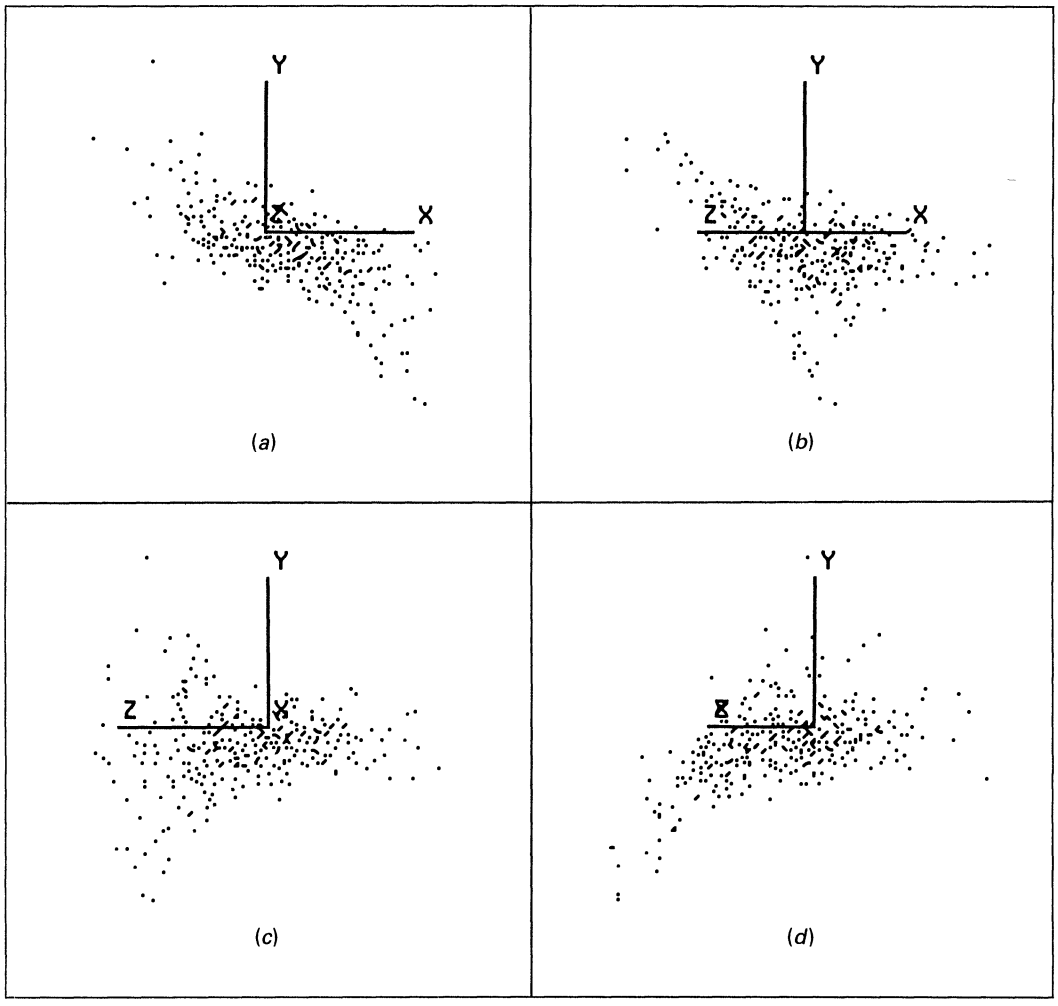

Figure 2. Best View of Data Generated From (6.3) With σ = .5, p = 10.

est eigenvector of cov[$E(\mathbf{z} \mid y)$]. This is what the first component of SIR, $\hat{\beta}_1$, attempts to estimate.

Generalizing the above argument, the projection direction yielding the variable that is most predictable from y, subject to being uncorrelated with the first most predictable variable, is a direction estimated by $\hat{\beta}_2$. Similar intepretation extends to the other $\hat{\beta}_k$.

Since the appearance of the first version of this article, four related works have been written. Duan and Li (in press) studied SIR for the case where $K = 1$ in more detail. They generalized the implementation of SIR by allowing a general weighting scheme for conducting the principal component analysis. Li (1989) examined SIR from a different angle. A projection pursuit approach was taken there, based on the notion of dependent variable transformation to provide a projection index. It brought out a nice connection with other works related to transformations (e.g., correspondence analysis, ACE, and the dummy variable approach to multivariate analysis by the Gifi school) Li (1990a) proposed a new method for handling the case where the regression function may be symmetric. Li (1990b) applied SIR to uncertainty analysis of mathematical models or computer models. SIR was used to visualize and simplify the models.

APPENDIX

A.1. Proof of Theorem 3.1. Without loss of generality, assume that $E(\mathbf{x}) = 0$. Consider any vector b in the orthogonal complement of the space spanned by $\beta_k\Sigma_{\mathbf{xx}}(k = 1, \ldots, K)$; that is, $\beta_k\Sigma_{\mathbf{xx}}b' = 0$. We need to show that $bE(\mathbf{x} \mid y) = 0$ with probability 1. Equation (1.1) implies

$$bE(\mathbf{x} \mid y) = E[E(b\mathbf{x} \mid \beta_k\mathbf{x}\text{'s}, y) \mid y] = E[E(b\mathbf{x} \mid \beta_k\mathbf{x}\text{'s}) \mid y].$$

Hence it suffices to show that $E(b\mathbf{x} \mid \beta_k\mathbf{x}\text{'s}) = 0$; or equivalently, $E[(E(b\mathbf{x} \mid \beta_k\mathbf{x}\text{'s}))^2] = 0$. By conditioning, the left term can be written as $E[E(bx \mid \beta_k\mathbf{x}\text{'s})\mathbf{x}'b']$, which equals $E[(c_0 + \Sigma_{k=1}^K c_k\beta_k\mathbf{x})\mathbf{x}'b'] = \Sigma_{k=1}^K c_k\beta_k\Sigma_{\mathbf{xx}}b' = 0$. The proof of Therem 3.1 is now complete.

A.2. Derivation of formula (5.1) for $E[R^2(\hat{B})]$. Due to the affine invariance, we may assume that $\mathbf{x}$ has mean 0 and covariance $\Sigma_{\mathbf{xx}} = I$. The squared trace correlation $R^2(\hat{B})$ reduces to

$$K^{-1} \operatorname{tr} \hat{P}_1P_1 = 1 - K^{-1} \operatorname{tr} (\hat{P}_1 - P_1)P_1(\hat{P}_1 - P_1),$$

where P_1 and $\hat{P}_1$ are symmetric projection matrices associated with the e.d.r. space B and the estimated space $\hat{B}$ respectively. From steps 4 and 5 of SIR, $\hat{P}_1$ is related to $\tilde{P}_1 = \Sigma_{k=1}^K \hat{\eta}_k'\hat{\eta}_k$, a projection matrix:

$$\hat{P}_1 = \hat{\Sigma}_{\mathbf{xx}}^{-1/2}(\tilde{P}_1\hat{\Sigma}_{\mathbf{xx}}^{-1}\tilde{P}_1)^+\hat{\Sigma}_{\mathbf{xx}}^{-1/2},$$

where the superscript + denotes the Moore–Penrose generalized

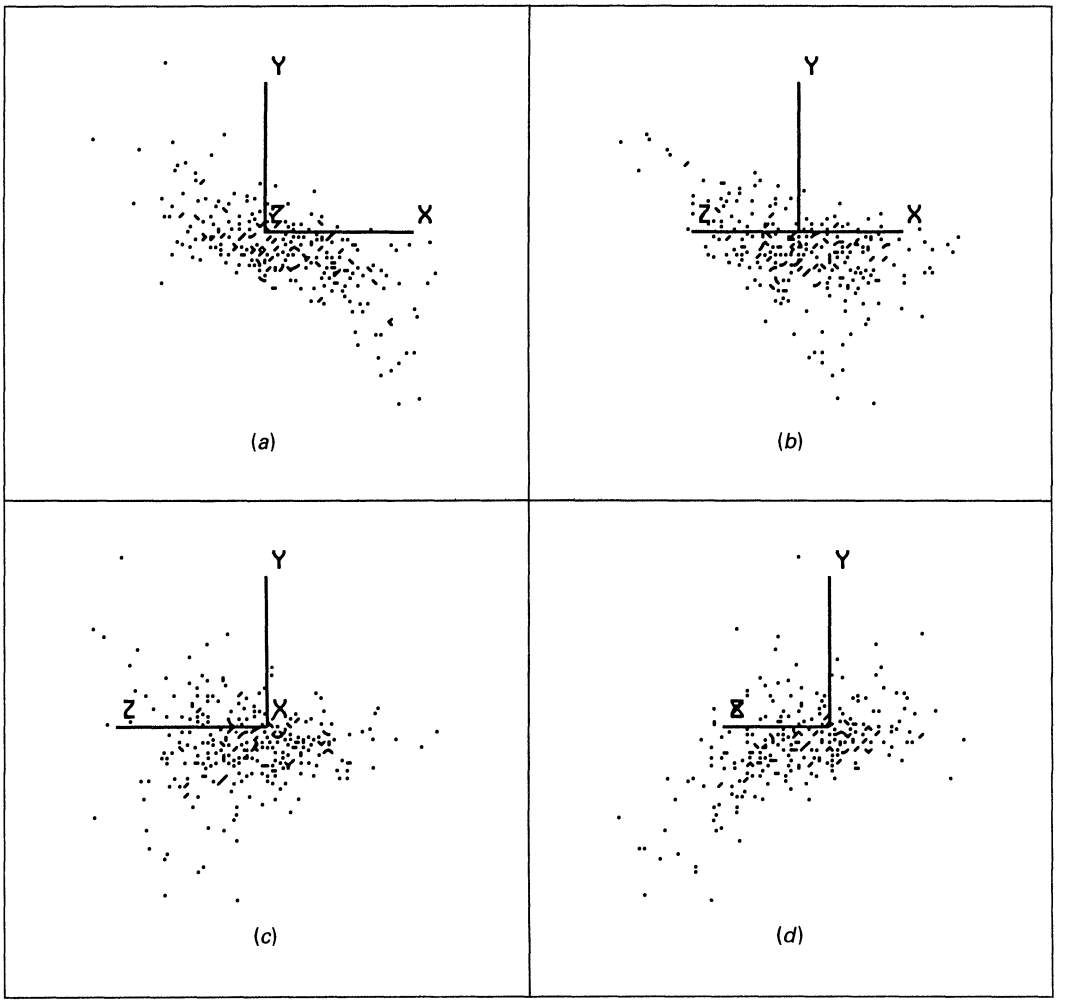

Figure 3. SIR's View of the Same Data as in Figure 2, H = 5.

inverse of a matrix. Furthermore, we need the following approximation which will be proved later:

$$\tilde{P}_1 = P_1 + P_2(\hat{V} - \tilde{V})\tilde{V}^+ + V^+(\hat{V} - \tilde{V})P_2 + o_p(n^{-1/2}), \quad \text{(A.1)}$$

where $P_2 = I - P_1$ and $\tilde{V}$ is defined by $\Sigma_{h=1}^H \hat{p}_h m_h m_h'$ with $m_h = E(\mathbf{x} \mid y \in I_h)$.

Now, approximate $\hat{\Sigma}_{\mathbf{xx}}$ by $I + \Delta$, where $\Delta = n^{-1}\,\Sigma_{i=1}^n \mathbf{x}_i\mathbf{x}_i' - I$. Take A to be the $p \times H$ matrix $(\hat{p}_1^{1/2}m_1, \ldots, \hat{p}_H^{1/2}m_H)$, and verify that $\tilde{V} = AJA'$, where J denotes the projection matrix $I - (\hat{p}_1^{1/2}, \ldots, \hat{p}_H^{1/2})(\hat{p}_1^{1/2}, \ldots, \hat{p}_H^{1/2})'$. Then we can derive

$$(\hat{P}_1 - P_1)P_1 = P_2\mathcal{C}_1JA'\tilde{V}^+ - P_2\Delta P_1 + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where the $n \times H$ matrix, $\mathcal{C}_1$, equals $(\hat{p}_h^{1/2}(\bar{\mathbf{x}}_h - m_h))$, and $\bar{\mathbf{x}}_h$ is defined in Remark 4.1.

It remains to calculate

$$E[R^2(\hat{B})] = 1 - K^{-1}(E_1 + E_2 - 2E_3) + o(n^{-1}),$$

where

$$E_1 = E[\mathrm{tr}(P_2\mathcal{C}_1JA'\tilde{V}^+)(\ \ldots\ )'] = E(\mathrm{tr}\tilde{V}^+\mathcal{C}_1'P_2\mathcal{C}_1)$$

$$E_2 = E[\mathrm{tr}(P_2\Delta P_1)(\ \ldots\ )'] = n^{-1}E(P_2\mathbf{x}\mathbf{x}'P_1\mathbf{x}\mathbf{x}'P_2)$$

$$= n^{-1}E[(\mathbf{x}'P_1\mathbf{x})(\mathbf{x}'P_2\mathbf{x})]$$

$$E_3 = \mathrm{tr}\,E[(P_2\mathcal{C}_1JA'\tilde{V}^+)(P_2\Delta P_1)'].$$

The normality assumption on $\mathbf{x}$ implies that conditional on $\hat{p}_k (k = 1, \ldots, K)$, the columns of $P_2\mathcal{C}_1$ are independent normal, with mean 0 and covariance $n^{-1}P_2$, leading to $E(\mathcal{C}_1'P_2\mathcal{C}_1) = (p - K)n^{-1}I$. Since $\tilde{V}$ converges to V, we obtain

$$E_1 = (p - K)n^{-1}\sum_{k=1}^{K}\lambda_k^{-1} + o(n^{-1}).$$

The second term E_2 equals $K(p - K)n^{-1}$, because of the independence between $P_1\mathbf{x}$ and $P_2\mathbf{x}$.

Write E_3 as $\mathrm{tr}[E(P_1\Delta P_2\mathcal{C}_1JA'V^+)] = \mathrm{tr}[E(P_1\mathbf{x}_1\mathbf{x}_1'P_2\hat{A})JA'V^+]$. The term inside the parentheses is a p by H matrix with the hth column equal to

$$n^{-1}\hat{p}_h^{-1/2}P_1\mathbf{x}_1\mathbf{x}_1'P_2\sum_{i=1}^{n}\delta_h(y_i)\mathbf{x}_i,$$

where δ_h is defined in step 2 of SIR. Replace $\hat{p}_h^{-1/2}$ by $p_h^{-1/2}$ in the above expression and then take the expectation. The result turns out to be $n^{-1/2}p_h^{1/2}(p - K)P_1m_h$ because of the independence between $P_2\mathbf{x}_1$ and y_1. Therefore we have

$$E_3 = n^{-1}(p - K)\,\mathrm{tr}(AJA'V^+) + \text{lower order term}$$

$$= n^{-1}(p - K)\mathrm{K} + \text{lower order term.}$$

Putting E_1, E_2, E_3 together, we have derived (5.1).

Proof of (A.1). The quickest way to derive this expansion is to use Lemma 4.1 in Tyler (1981). Instead of expanding $\tilde{P}_1$ about P_1, we first expand $\tilde{P}_2 = I - \tilde{P}_1$ about P_2. One advantage is that there is only one eigenvalue associated with $P_2\tilde{V}P_2$, which is 0;

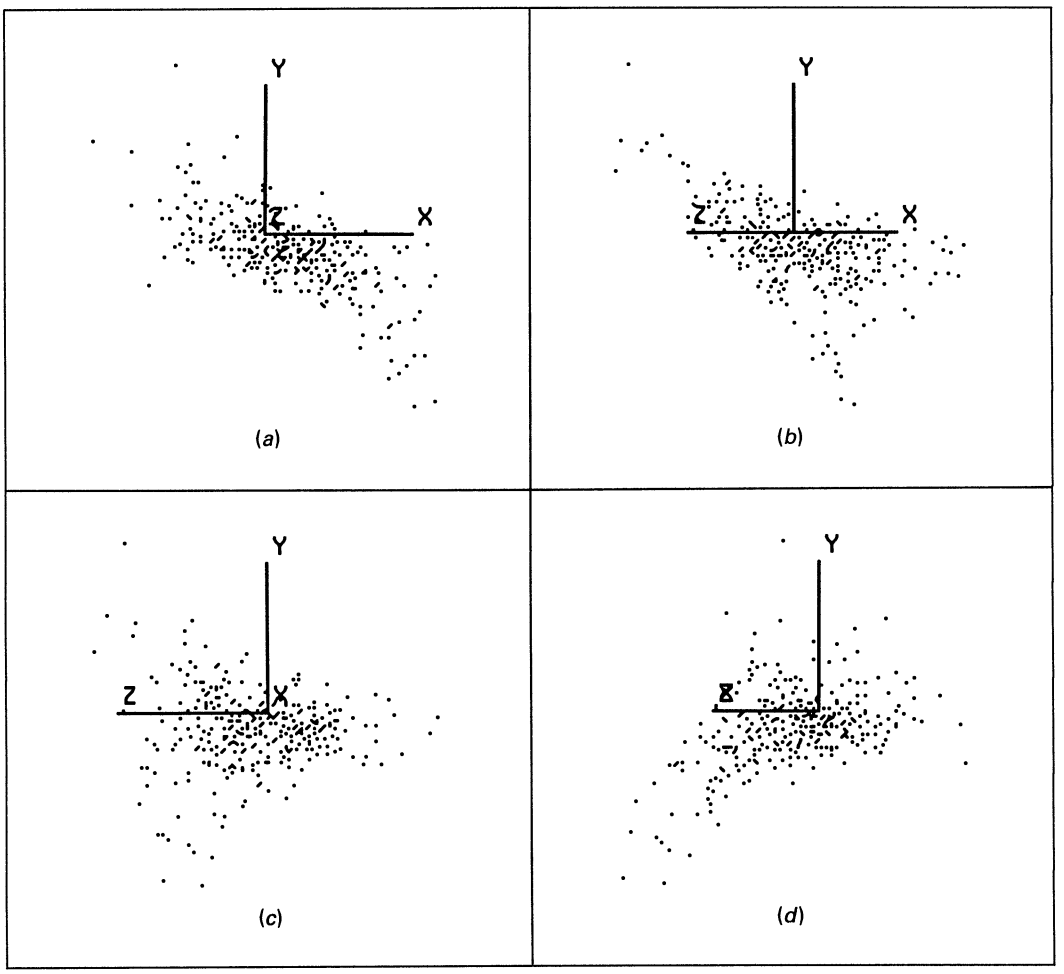

Figure 4 (a)–(d). Best View of the Same Data as in Figure 2, H = 30.

so the Taylor expansion formula is simpler. This leads to

$$\tilde{P}_2 = P_2 - [P_2(\hat{V} - \tilde{V})\tilde{V}^+ + \tilde{V}^+(\hat{V} - \tilde{V})P_2] + o_p(n^{-1/2}),$$

implying (A.1).

A.3. Asymptotic expansion for $\tilde{\lambda}_{(p-K)}$. The following lemma is the key to our asymptotic expansion.

Lemma A.1. Consider the expansion

$$T(\omega) = T + \omega T^{(1)} + \omega^2 T^{(2)} + o(\omega^2),$$

where $T(\omega)$, T, $T^{(1)}$, $T^{(2)}$ are symmetric matrices. Suppose that T is nonnegative definite with rank K. Then the average of the smallest $p - K$ eigenvalues of $T(\omega)$, $\lambda(\omega)$, has the expansion

$$\lambda(\omega) = \frac{1}{p - K}[\omega\lambda^{(1)} + \omega^2\lambda^{(2)}] + o(\omega^2),$$

where $\lambda^{(1)} = \operatorname{tr} T^{(1)}\Pi$, $\lambda^{(2)} = \operatorname{tr}[T^{(2)}\Pi - T^{(1)}T^+T^{(1)}\Pi]$, and Π is the symmetric projection matrix of rank $p - K$ such that $\Pi T = T\Pi = 0$.

This lemma is a simplified version of a result in the perturbation theory for finite dimensional spaces (see chap. 2 of Kato 1976, p. 79, eq. (2.33)). To use this lemma, obtain, after a straightforward asymptotic expansion, that

$$\begin{aligned}\hat{V} - V = {} & [AJ(\mathscr{E}_1' + A'\mathscr{E}_2) + (\mathscr{E}_1 + \mathscr{E}_2A)JA'] \\ & + [(\mathscr{E}_1 + \mathscr{E}_2A)J(\mathscr{E}_1' + A'\mathscr{E}_2) + AJA'\mathscr{E}_1\mathscr{E}_2 \\ & + \mathscr{E}_2'\mathscr{E}_1'AJA'] + o_p(1/n),\end{aligned}$$

where $\mathscr{E}_2 = \hat{\Sigma}_{xx}^{-1/2} - I$. Thus we may substitute V for T, $n^{-1/2}$ for ω, $\hat{V}$ for $T(\omega)$, $n^{1/2}$ times the first bracketed term for $T^{(1)}$, n times the second bracketed term for $T^{(2)}$, and Π by P_2.

Straightforward computation leads to

$$\omega\lambda^{(1)} = \operatorname{tr} T^{(1)}P_2P_2 = \operatorname{tr} P_2T^{(1)}P_2 = 0,$$

$$\omega^2\lambda^{(2)} = \operatorname{tr} P_2\mathscr{E}_1Q\mathscr{E}_1'P_2,$$

where $Q = J - JA'\tilde{V}^+AJ$, a projection matrix with trace $H - K - 1$. Based on a conditional probability argument similar to that used in deriving E_1 in (A.2), we can show that $n \cdot \operatorname{tr}(P_2\mathscr{E}_1Q\mathscr{E}_1'P_2)$ follows a χ^2 distribution with $(p - K)(H - K - 1)$ degrees of freedom. This completes the proof. Note that the normality assumption on $\mathbf{x}$ is not necessary if the conditional covariance of $P_2\mathbf{x}$ given $y \in I_h$ does not depend on h because $\bar{\mathbf{x}}_h$ is asymptotically normal by the central limit theorem.

[*Received July 1988. Revised March 1990.*]

REFERENCES

Bickel, P. J., and Doksum, K. A. (1981), "An Analysis of Transformations Revisited," *Journal of the American Statistical Association*, 76, 296–311.

Breiman, L., and Friedman, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association* 80, 580–597.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*. Belmont, 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 Erick L. Lehmann*, Belmont, CA: Wadsworth, pp. 97–114.

Box, G., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

Box, G., and Draper, IV. (1987), *Empirical Model-Building and Response Surfaces*, New York: John Wiley.

Carr, D. B., Littlefield, R. J., Nicholson, W. L., and Littlefield, J. S. (1987), "Scatterplot Matrix Techniques for Large *N*," *Journal of the American Statistical Association*, 82, 424–437.

Carroll, R., and Ruppert, D. (1984), "Power Transformations When Fitting Theoretical Models to Data," *Journal of the American Statistical Association*, 79, 321–328.

Chen, H. (1988), "Convergence Rates for Parametric Components in a Partly Linear Model." *The Annals of Statistics*, 16, 136–146.

——— (in press), "Rates of Convergence for Projection Pursuit Regression," *The Annals of Statistics*.

Cleveland, W. S. (1987), "Research in Statistical Graphics," *Journal of the American Statistical Association* 82, 419–423.

Cuzik, J. (1987), "Semiparametric Additive Regression," unpublished manuscript.

Diaconis, P., and Freedman, D. (1984), "Asymptotics of Graphical Projection Pursuit," *The Annals of Statistics*, 12, 793–815.

Donoho, D., Johnstone, I., Rousseeuw, P., and Stahel, W. (1985), "Discussion of On Projection Pursuit," *The Annals of Statistics*, 13 496–499.

Donoho, D., and Johnstone, I. (1989), "Projection-Based Smoothing, and a Duality With Kernel Methods," *The Annals of Statistics* 17, 58–106.

Duan, N., and Li, K. C. (in press), "Slicing Regression: a Link-Free Regression Method," *The Annals of Statistics*.

Engle, R. F., Granger, C. W. I., Rice, J., and Weiss, A. (1986), "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310–320.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

Fill, J. A., and Johnstone, I. (1984), "On Projection Pursuit Measures of Multivariate Location and Dispersion," *The Annals of Statistics*, 12, 127–141.

Friedman, J. (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association*, 82, 249–266.

Friedman, J., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

Härdle, W., Hall, P., and Marron, S. (1988), "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum," *Journal of the American Statistical Association*, 83, 86–101.

Hall, P. (1989), "On Projection Pursuit Regression," *The Annals of Statistics* 17, 573–588.

Hastie, T., and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 1, 297–318.

Heckman, N. (1986), "Spline Smoothing in Partly Linear Models," *Journal of the Royal Statistical Society* Ser. B, 48, 244–248.

Hinkley, D. V., and Runger, G. (1984), "The Analysis of Transformed Data," with discussion, *Journal of the American Statistical Association*, 79, 302–320.

Hooper, J. (1959), "Simultaneous Equations and Canonical Correlation Theory," *Econometrica*, 27, 245–256.

Hotelling, H. (1935), "The Most Predictable Criterion," *Journal of Educational Psychology*, 139–142.

Huber, P. (1985), "Projection Pursuit," with discussion, *The Annals of Statistics*, 13, 435–526.

Huber, P. (1987), "Experiences With Three-Dimensional Scatterplots," *Journal of the American Statistical Association*, 82, 448–454.

Kato, T. (1976), *Perturbation Theory for Linear Operators* (2nd ed.), Berlin: Springer-Verlag.

Koyak, R. (1987), "On Measuring Internal Dependence in a Set of Random Variables," *The Annals of Statistics*, 15, 1215–1228.

Li, G., and Chen, Z. (1985), "Projection Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo." *Journal of the American Statistical Association*, 80, 759–766.

Li, K. C. (1987), "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975.

——— (1989), "Data Visualization With SIR: a Transformation Based Projection Pursuit Method," UCLA statistical series 24.

——— (1990a), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," UCLA technical report, Dept. of Mathematics.

——— (1990b), "Uncertainty Analysis for Mathematical Models With SIR," UCLA technical report, Dept. of Mathematics.

Li, K. C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics*, 17, 1009–1052.

Loh, W. Y., and Vanichsetakul, N. (1988), "Tree-Structured Classification via Generalized Discriminant Analysis," *Journal of the American Statistical Association* 83, 715–728.

Mallows, C. L. (1961), "Latent Vectors of Random Symmetric Matrices," *Biometrika*, 48, 133–149.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*. New York: Academic Press.

Portnoy, S. (1985), "Asymptotic Behavior of *M*-Estimators of p Regression Parameters When p^2/n is Large II: Normal Approximation," *The Annals of Statistics*, 13, 1403–1417.

Speckman, P. (1987), "Kernel Smoothing in Partial Linear Models," unpublished manuscript.

Stone, C. (1986), "The Dimensionality Reduction Principle for Generalized Additive Models," *The Annals of Statistics*, 13, 689–705.

Tierney, L. (1989), "XLISP.STAT: A Statistical Environment Based on the XLISP Language," (Beta Test Version 2.0). School of Statistics, University of Minnesota.

Tyler, D. (1981), "Asymptotic Inference for Eigenvectors," *The Annals of Statistics*, 9, 725–736.

van Rijckevorsel, L. A., and de Leeuw, J. (1988), *Component and Correspondence Analysis*, New York: John Wiley.

Wahba, G. (1986), "Partial and Interaction Splines for Semiparametric Estimation of Functions of Several Variables," in Computer Science and Statistics: Proceedings of the 18-th Symposium on the Interface, Washington, D.C., pp. 75–80.