Title
On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma

Author
Ker-Chau Li

Publication Date
2011-10-24

On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma

KER-CHAU LI*

Modern graphical tools have enhanced our ability to learn many things from data directly. With much user-friendly graphical software available, we are encouraged to plot a lot more often than before. The benefits from direct interaction with graphics have been enormous. But trailing behind these high-tech advances is the issue of appropriate guidance on what to plot. There are too many directions to project a high-dimensional data set and unguided plotting can be time-consuming and fruitless. In a recent article, Li set up a statistical framework for study on this issue, based on a notion of effective dimension reduction (edr) directions. They are the directions to project a high dimensional input variable for the purpose of effectively viewing and studying its relationship with an output variable. A methodology, sliced inverse regression, was introduced and shown to be useful in finding edr directions. This article introduces another method for finding edr directions. It begins with the observation that the eigenvectors for the Hessian matrices of the regression function are helpful in the study of the shape of the regression surface. A notation of principal Hessian directions (pHd's) is defined that locates the main axes along which the regression surface shows the largest curvatures in an aggregate sense. We show that pHd's can be used to find edr directions. We further use the celebrated Stein lemma for suggesting estimates. The sampling properties of the estimated pHd's are obtained. A significance test is derived for suggesting the genuineness of a view found by our method. Some versions for implementing this method are discussed, and simulation results and an application to real data are reported. The relationship of this method with exploratory projection pursuit is also discussed.

KEY WORDS: Projection pursuit; Sliced inverse regression; Statistical graphics; Stein's lemma.

1. INTRODUCTION

Statistical graphics is indispensable for data analysis. Histograms, stem-and-leaf plots, normal probability plots, box plots, and scatterplots, for instance, have been routinely used for describing features, summarizing information, suggesting models, and guiding statistical inference. With modern computing power, graphics can now be easily created on personal computers through software such as MacSpin, S, Systat, NCSS, Xlisp.stat, Data Desk, and so on. Tukey's works and influence have nourished the growth in this area; see the collection of Tukey's works on graphics edited by Cleveland (1988). The dynamic features such as brushing, slicing, linking, and animating, together with 3-dimensional rotation and scatterplot matrix techniques have become popular; see Cleveland and MacGill (1988) and Wegman and Depriest (1986).

With such powerful graphical tools now available, there is also a growing need for guidance on what to plot (Cook and Weisberg 1989). This is particularly the case when analyzing high-dimensional data. For instance, a data set with 10 variables can yield $\binom{10}{3} = 120$ rotation plots for inspection. Without proper guidance, this can become a time-consuming, fruitless task. To meet this need, many projection pursuit methods (commonly associated with the names of Fisferkeller, Friedman, Huber, Kruskal, Switzer, Tukey, and Wright) have been invented. These methods find useful viewing angles by machine-picking the most interesting projections via the maximization of some projection index Friedman 1987; Huber 1985). The complexity of these algorithms, however, has limited our understanding about their statistical properties. Only a few such theoretical works are available (Chen 1991; Donoho and Johnstone 1989; Hall 1989a,b).

For suggesting what to plot, Li (1991) formulated the high-dimensional data visualization problem through a dimension-reduction model

$$y = g(\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x}, \varepsilon). \tag{1.1}$$

Here we have assumed a variable of primary interest y and want to see its relationship with p other variables, $\mathbf{x}$. The β_j's are unknown vectors to be estimated from data, and ε is the random error. The set, $\mathcal{B}$, of any linear combination of β_j's will be referred to as the *effective dimension reduction (edr) space.* Model (1.1) looks like a nonlinear regression model except for one difference: g is completely unknown. This difference is very important. For instance, when $K = 2$, it yields the needed flexibility for allowing any pattern in the plot of y against $\beta_1'\mathbf{x}$, $\beta_2'\mathbf{x}$ to occur. Model (1.1) represents the weakest form for expressing the wish that useful information about y from a high-dimensional covariate $\mathbf{x}$ can be retrieved from its low-dimensional projected variable $(\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x})$ when K is small. In fact, we could have allowed K to be equal to p, in which case (1.1) becomes redundant and we do not assume any model at all. But we cannot anticipate a typical data set of size, say less than 1,000, to reveal all 10 dimensional nonlinear structures if $K = p = 10$, say. A method called sliced inverse regression (SIR) is used for finding edr directions. SIR does not depend on K.

* Ker-Chau Li is Professor of Mathematics at University of California, Los Angeles, Los Angeles, CA 90024. The authors' interest in data visualization was inspired by Dennis Cook, who introduced Tierney's Xlisp.stat to him; thanks to Dennis's and Ray Carroll's interests in SIR, the author was encouraged to pursue this work.

Journal of the American Statistical Association
December 1992, Vol. 87, No. 420, Theory and Methods

In practice, this is desirable because most often we do not know K in advance. Indeed, Li suggested a chi-squared criterion for determining whether or not an estimated component represents a real structure. In this way, he curtailed the difficult problem of estimating K by offering a useful lower bound based on the data.

The estimation of g has been deliberately avoided. SIR does not need to estimate g to find the correct edr directions. This makes the principal behind SIR distinct from other nonparametric modeling techniques such as ACE (Breiman and Friedman 1985), PPR (Friedman and Stuetzle 1981), generalized additive models (Stone 1986), and partly spline models (Rice 1986; Wahba 1986). The issue of finding important variables, or linear factors (i.e., $\beta_k'\mathbf{x}$), can be distinguished from the issue of data fitting or functional approximation. On the one hand it is not necessary to approximate the unknown regression function or to model the data before selecting the crucial variables or linear factors. On the other hand, once crucial variables or linear factors are found, we could proceed with any analysis, depending on what we have seen from the projected data. We could estimate the regression function (either parametrically or nonparametrically), build an empirical model, estimate the quantiles of ε, impose suitable constraints on the functional form for g, conduct heteroscedasticity analysis, seek for clusters, or simply decide that the reduced data set does not provide needed clues to meet the study's major goal. Clearly, data analysis has a wider scope than data–model fitting. Other powerful tools that can be used for data reduction include recursive partition and classification (Breiman, Friedman, Olshen, and Stone 1984; Loh and Vanichsetakul 1988) and many nonlinear multivariate analysis techniques, including correspondence analysis (Gifi 1990).

The first moment-based SIR has received more extensive studies (Duan and Li 1991; Hsing and Carroll 1991; Li 1989b, 1991). However, one major restriction has been its vulnerability to the symmetry of g about the mean of $\mathbf{x}$. For example, if $y = x_1^2$ and x_1, the first coordinate variable in $\mathbf{x}$, is symmetric about its mean, then the inverse regression curve, $E(\mathbf{x} \mid y)$ is degenerated. But at least with the proposed chi-squared test, SIR would conservatively admit that no interesting direction is found. Suggestions for remedy can be offered based on second moments (Cook and Weisberg 1991; Li 1989b, 1991 [rejoinder]).

This article presents another method for finding the edr space, which can handle many symmetric cases. The motivation comes from the observation that the Hessian matrix $\mathbf{H}_{\mathbf{x}}(\mathbf{x})$ of the regression function $f(\mathbf{x}) = E(y|\mathbf{x})$ at any point $\mathbf{x}$, as defined by the p by p matrix, $[(\partial^2/\partial x_i \partial x_j) f(\mathbf{x})]$, will be degenerate along any directions that are orthogonal to $\mathcal{B}$. Based on this, we use a notion of principal Hessian directions (pHd's) for identifying the edr space. Roughly, the pHd's form a new coordinate system with the property that the curvatures of the regression function along the p coordinate axes are successively the largest possible in an average sense as measured by second-partial derivatives. (See Sec. 2 for details.)

In Section 3 we propose a simple method for estimating the average Hessian matrix and the pHd's, based on a celebrated Stein's lemma (Stein 1981). This method leads to three variants of estimates, all shown to be Fisher consistent, under the assumption that the distribution of $\mathbf{x}$ is normal. Two of them take the form of a weighted covariance matrix of $\mathbf{x}$ with the weighted factor determined by y, the dependent variable, or r, the residual after fitting a linear model. The other variant is related with quadratic polynomial fitting. They are all very simple to compute.

In Section 4 we study large sample properties, establish the root n consistency, evaluate a closeness measure between the estimated edr space and the true edr space as defined in Li (1991), and propose significance tests for assessing the number of components K. In Section 5 we show how transformations on y can be applied to the pHd methodology. In particular, this brings up a connection with second-moment-based SIR.

In Section 6 we study the behavior of our estimates under the linear conditional expectation condition for $\mathbf{x}$, a much weaker assumption used in formalizing the theory of SIR. We show that the estimated pHd's are still useful in finding edr directions without the normality assumption on the input variable.

We discuss the projection pursuit aspect of our approach in Section 7, arguing for its merit as means of constructing descriptive statistics. For readers interested mainly in the application of the pHd methodology, Sections 5–7 may be skipped during the first reading.

In Section 8 we present some simulation results to illustrate the theory and apply our method to a real data set. We illustrate how pHd and SIR can be used to complement each other in finding interesting features in the data. We conclude in Section 9 by discussing some general issues in the area of data visualization and dimension reduction.

The longer proofs of Theorems 4.1 and 4.2 are given in Appendix B; other proofs are presented in Appendix A. Remark 1.1 gives a brief account for the relationship between this work and Brillinger (1977, 1983).

Remark 1.1. The distinction between the design of experiment approach and the correlation approach to regression analysis often emerges in the literature. One issue is whether or not one should condition on regressors. For experimental data, this has been controversial due to the role of randomization. In the observational study, however, the interplay between the two sides has greatly enriched the theory of regression analysis.

Yet despite the awareness of the merit in treating the regressors as random, it has been a surprise to find out that regressors with a Gaussian distribution can enjoy many robustness properties against link violation. The first of these is the discovery in Brillinger (1977, 1983). Brillinger showed that the least squares estimate for the slope vector in multiple linear regression is still root n consistent up to a proportionality scalar, even if the linear model assumption is violated and the true model takes the form of model (1.1) with $K = 1$. Brillinger proved his result under the condition that $\mathbf{x}$ is Gaussian, but pointed out the key linearity condition, similar to (6.1) in Section 6 of this article. He also linked his result to Stein's lemma. Brillinger's result was extended by

Li and Duan (1989) to most regression estimates. (See the references given in that article for other related works.) Brillinger's influence on the development of SIR and on this article is clear. Corollary 3.2, for example, can be viewed as a generalization of his result to polynomial regression.

Remark 1.2. Stein's lemma has many applications in statistics. The most well known is in the discovery of the inadmissibility of the sample mean for estimating a multivariate normal mean, the beginning of a new era in decision theory. More recently, Stein's estimate and its unbiased risk estimate have been applied to nonparametric smoothing, establishing the connection with the generalized cross-validation, and constructing honest confidence regions for nonparametric regression; see Li and Hwang (1984) and Li (1985, 1986, 1987, 1989a). Stein's lemma and its generalization also have a wide application in probability theory; see Stein (1986).

2. PRINCIPAL HESSIAN DIRECTIONS AND EFFECTIVE DIMENSION REDUCTION DIRECTIONS

In this section we define principal Hessian directions and discuss their roles in finding edr directions for dimension reduction.

2.1 Principal Hessian Directions

The Hessian matrix typically varies as $\mathbf{x}$ changes unless the surface is quadratic. Let the mean of the Hessian matrix be $\bar{\mathbf{H}}_{\mathbf{x}} = \mathbf{E}\mathbf{H}_{\mathbf{x}}(\mathbf{x})$. We define the pHd's with respect to the distribution of $\mathbf{x}$ as the eigenvectors $b_1, \ldots, b_p$ of the matrix $\bar{\mathbf{H}}_{\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}}$, where $\boldsymbol{\Sigma}_{\mathbf{x}}$ denotes the covariance matrix of $\mathbf{x}$:

$$\bar{\mathbf{H}}_{\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}} b_j = \lambda_j b_j, \qquad j = 1, \ldots, p$$
$$|\lambda_1| \geq \cdots \geq |\lambda_p|. \tag{2.1}$$

Because any scale multiple of an eigenvector is also an eigenvector, we now restrict that $b_j'\boldsymbol{\Sigma}_{\mathbf{x}} b_j = 1$ to avoid ambiguity. The pHd's are the directions to form a new set of coordinate axes along which the average curvatures of the regression function $E(y|\mathbf{x})$ are successively the largest in terms of second partial derivatives, as illustrated in the following paragraph.

First, observe from the chain rule that the Hessian matrix, $\mathbf{H}_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}) = [(\partial^2/\partial\tilde{x}_i\partial\tilde{x}_j) f(\tilde{\mathbf{x}})]$, taken with respect to a new coordinate system $\tilde{\mathbf{x}} = \mathbf{A}\mathbf{x}$, is related to $\mathbf{H}_{\mathbf{x}}(\mathbf{x})$:

$$\mathbf{H}_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}) = \mathbf{A}'^{-1}\mathbf{H}_{\mathbf{x}}(\mathbf{x})\mathbf{A}^{-1}.$$

This implies that $\bar{\mathbf{H}}_{\tilde{\mathbf{x}}} = \mathbf{A}'^{-1}\bar{\mathbf{H}}_{\mathbf{x}}\mathbf{A}^{-1}$. The new coordinate variables are restricted to be uncorrelated and have equal length in the sense that $\operatorname{cov}\tilde{\mathbf{x}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}' = I$, an identity matrix. Next, we want to find one system that has the largest absolute value for the expected second partial derivative along the first axis:

$$\max_{\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}'=I} |e_1'\mathbf{A}'^{-1}\bar{\mathbf{H}}_{\mathbf{x}}\mathbf{A}^{-1}e_1|, \tag{2.2}$$

where $e_1 = (1, 0, \ldots, 0)'$.

Lemma 2.1. The first row of any matrix $\mathbf{A}$ maximizing (2.2) must be equal to b_1'.

After fixing the first axis, we can find the second axis similarly by maximizing $|e_2'\bar{\mathbf{H}}_{\tilde{\mathbf{x}}}e_2|$, where $e_2 = (0, 1, 0, \ldots, 0)'$. By the same argument as in Lemma 2.1, this axis is determined by b_2. The same interpretation holds for other b_j's. Therefore, the pHd's can be used to form a new coordinate system for visualizing the relationship between $\mathbf{x}$ and y, with the property that the average curvatures of the regression function along the coordinate axes are successively the largest. In reverse order, the same coordinate system yields the smallest absolute values for the average curvatures successively.

From this interpretation, we see that our definition of pHd's is affine invariant. Specifically, consider any invertible matrix $\mathbf{T}$ and any vector $\mathbf{a}$. Transform $\mathbf{x}$ to $\tilde{\mathbf{x}} = \mathbf{a} + \mathbf{T}\mathbf{x}$. Then the pHd's with respect to $\tilde{\mathbf{x}}$ are the eigenvectors, $\tilde{b}_j$'s, for the matrix $\bar{\mathbf{H}}_{\tilde{\mathbf{x}}}\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}} = \mathbf{T}'^{-1}\bar{\mathbf{H}}_{\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{T}'$; $\mathbf{T}'^{-1}\bar{\mathbf{H}}_{\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{T}'\tilde{b}_j = \tilde{\lambda}_j\tilde{b}_j$. Multiply $\mathbf{T}'$ on both sides of this identity to obtain that $\mathbf{T}'\tilde{b}_j = b_j$ and $\tilde{\lambda}_j = \lambda_j$. Therefore, we see that $\tilde{b}_j = \mathbf{T}'^{-1}b_j$ and that the new projected variable $\tilde{b}_j'\tilde{\mathbf{x}}$ is the same as the original one, $b_j'\mathbf{x}$.

Remark 2.1. Any plot needs a suitable scale that fits in the visual perception. Our restriction on the coordinate system takes the standard deviation as the scaling factor. Obviously, we could choose others that may have better robustness properties (cf. Donoho et al. 1985; Fill and Johnstone 1984; Li and Chen 1985). Another aspect for modification is that we may allow $\boldsymbol{\Sigma}_{\mathbf{x}}$ in (2.2) to be replaced by other appropriate matrices for incorporating subjective opinions on the relative importance among the input variables.

2.2 Dimension Reduction

Recall model (1.1) from Section 1. We now study the properties of the average Hessian matrix $\bar{\mathbf{H}}_{\mathbf{x}}$ and the associated pHd's under (1.1).

Under (1.1), the regression function takes the form

$$\mathbf{E}(y|\mathbf{x}) = f(\mathbf{x}) = h(\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x}) \tag{2.3}$$

for some function h. Assume that h is twice differentiable.

Lemma 2.2. Under (2.3), the rank of the average Hessian matrix, $\bar{\mathbf{H}}_{\mathbf{x}}$, is at most K. Moreover, the pHd's with nonzero eigenvalues are in the edr space, $\mathcal{B}$.

This lemma indicates that if we can estimate the average Hessian matrix well, then the associated pHd's with significant nonzero eigenvalues can be used to find edr directions. In Section 3 we shall use Stein's lemma to suggest an estimate of the average Hessian matrix.

Remark 2.2. Another way to motivate the definition of pHd's is through the Mahalanobis metric, which defines the distance between two points $\mathbf{x}_1$, $\mathbf{x}_2$, in R^p as $((\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x}_1 - \mathbf{x}_2))^{1/2}$. Let $\mathbf{v}$ denote a unitary vector in the Mahalanobis metric, $\mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{v} = 1$. For each $\mathbf{v}$, consider the associated bundle of curves on the regression surface generated by moving each point $\mathbf{x}$ in R^p along direction $\mathbf{v}$ with the unitary speed. Measure the local nonlinearity of each curve at time $t = 0$ by the second derivative $(\partial^2/\partial t^2)f(\mathbf{x} + t\mathbf{v})|_{t=0}$, which is seen to be $\mathbf{v}'\mathbf{H}_{\mathbf{x}}(\mathbf{x})\mathbf{v}$. Let $\mathbf{v}_1$ be the direc-

tion of movement $\mathbf{v}$ that maximizes the absolute value of the aggregate nonlinearity $|\mathbf{v}'\bar{\mathbf{H}}_{\mathbf{x}}\mathbf{v}| = |\mathbf{E}\mathbf{v}'\mathbf{H}_{\mathbf{x}}(\mathbf{x})\mathbf{v}|$ for the associated bundle of regression curves. Subject to being orthogonal to $\mathbf{v}_1$ in the Mahalanobis metric, $\mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{v}_1 = 0$, let $\mathbf{v}_2$ be the $\mathbf{v}$ that maximizes $|\mathbf{v}'\bar{\mathbf{H}}_{\mathbf{x}}\mathbf{v}|$. Define $\mathbf{v}_3, \ldots, \mathbf{v}_p$ in a similar way. It is easy to check that $\mathbf{v}_j$'s are the eigenvectors for $\boldsymbol{\Sigma}_{\mathbf{x}}\bar{\mathbf{H}}_{\mathbf{x}}$. Furthermore, multiplying $\boldsymbol{\Sigma}_{\mathbf{x}}$ on both sides of (2.1), we see that $\mathbf{v}_j = \boldsymbol{\Sigma}_{\mathbf{x}}b_j$. It follows that $b_j'\mathbf{v}_{j'} = 0$ if $j \neq j'$. Now suppose we want to reduce the dimensionality of $\mathbf{x}$ to just 1 by choosing a suitable projection direction b. Because the movement of any point in R^p along a direction $\mathbf{v}$ perpendicular to b will not be detected from the projected variable $b'\mathbf{x}$, possible interesting nonlinear structures on the associated bundle of curves on the regression surface are obscured after projection. To minimize this loss, the orthogonal complement of b should contain the directions of movement for which the associated bundle of regression curves are as linear as possible. This suggests the choice of $b = b_1$, because its orthogonal complement is spanned by $\mathbf{v}_2, \ldots, \mathbf{v}_p$, the subspace with dimension $p - 1$ that generates bundles of least nonlinear regression curves. In particular, if all second derivatives are identically 0 as we move points along any directions $\mathbf{v}$ in the space spanned by $\mathbf{v}_2, \ldots, \mathbf{v}_p$, then the nonlinear structure in the response surface can be perfectly revealed from the first pHd direction, b_1. This argument applies to the cases where projections on more than one dimension are desirable.

Remark 2.3. In Remark 2.2, a better nonlinearity measure is to take the absolute value of the second derivative before taking the expectation. But its implementation probably is not simple.

Remark 2.4. The relationship between $\mathbf{v}_j$'s and b_j's are mathematically more transparent from the theory of reproducing kernel Hilbert space (see, for example, Parzen 1961). The Mahalanobis metric $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}$ is the reproducing kernel for the random vector $\mathbf{x}$ with covariance $\boldsymbol{\Sigma}_{\mathbf{x}}$. The duality relationship between b_j' and $\mathbf{v}_j$, $b_j = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{v}_j$, is now clear. Based on this, we can also generalize our method to the case in which $\mathbf{x}$ is a stochastic process with a reproducing kernel Hilbert space structure.

3. STEIN'S LEMMA AND ESTIMATES OF THE PRINCIPAL HESSIAN DIRECTIONS

We shall show how to use Stein's Lemma to estimate the pHd's when the distribution of $\mathbf{x}$ is normal.

3.1 Stein's Lemma

Recall Stein's Lemma from Stein (1981, Lemma 4).

Lemma 3.1. If the random variable z is normal, with mean ξ and variance 1, then

$$\mathbf{E}(z - \xi)l(z) = E\dot{l}(z)$$

$$\mathbf{E}(z - \xi)^2 l(z) = \mathbf{E}l(z) + \mathbf{E}\ddot{l}(z),$$

where, in each case, all derivatives involved are assumed to exist, in the sense that an indefinite integral of each is the next preceding one, and to have finite expectations.

Using Stein's lemma, it is easy to derive the following corollary.

Corollary 3.1. Suppose that $\mathbf{x}$ is normal with mean $\mu_{\mathbf{x}}$ and covariance $\boldsymbol{\Sigma}_{\mathbf{x}}$. Let μ_y be the mean of y. Then the average Hessian matrix $\bar{\mathbf{H}}_{\mathbf{x}}$ is related to the weighted covariance

$$\boldsymbol{\Sigma}_{yxx} = E(y - \mu_y)(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})'$$

through the identity

$$\bar{\mathbf{H}}_{\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{yxx}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}.$$

From this corollary, we can find pHd's based on the weighted covariance matrix $\boldsymbol{\Sigma}_{yxx}$, as the following theorem suggests.

Theorem 3.1. When $\mathbf{x}$ is normal, the pHd's b_j, $j = 1, \ldots, p$, can be obtained by the eigenvectors for the eigenvalue decomposition of $\boldsymbol{\Sigma}_{yxx}$ with respect to $\boldsymbol{\Sigma}_{\mathbf{x}}$:

$$\boldsymbol{\Sigma}_{yxx}b_j = \lambda_j\boldsymbol{\Sigma}_{\mathbf{x}}b_j, \quad \text{for} \quad j = 1, \ldots, p.$$

Observe that adding or subtracting a linear function of $\mathbf{x}$ from y does not change the Hessian matrix. Hence, instead of using y in Theorem 3.1, we may replace it by the residual after the linear least squares fit.

Theorem 3.2. Suppose that $\mathbf{x}$ is normal. Let $r = y - a - b_{ls}'\mathbf{x}$ be the residual for the linear regression of y on $\mathbf{x}$, where a, b_{ls} are the least squares estimates so that $\mathbf{E}r = 0$ and $\operatorname{cov}(r, \mathbf{x}) = 0$. Then we have

$$\bar{\mathbf{H}}_{\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{rxx}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1},$$

where

$$\boldsymbol{\Sigma}_{rxx} = \mathbf{E}r(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})'.$$

Moreover, the pHd's b_j, $j = 1, \ldots, p$, can be obtained by an eigenvalue decomposition of $\boldsymbol{\Sigma}_{rxx}$ with respect to $\boldsymbol{\Sigma}_{\mathbf{x}}$:

$$\boldsymbol{\Sigma}_{rxx}b_j = \lambda_j\boldsymbol{\Sigma}_{\mathbf{x}}b_j, \quad \text{for} \quad j = 1, \ldots, p.$$

Remark 3.1. Corollary 3.1 can also be applied to show that polynomial regression can be used to estimate pHd's, as the following corollary suggests.

Corollary 3.2. Suppose that $\mathbf{x}$ is normal and consider a polynomial fitting:

$$\min_{Q(\mathbf{x})} \mathbf{E}(y - Q(\mathbf{x}))^2,$$

where $Q(\mathbf{x})$ is any polynomial function of $\mathbf{x}$ with total degrees no greater than q. Then the average Hessian matrix for the fitted polynomial is the same as the average Hessian matrix for y, if q is larger than 1.

3.2 Estimates for Principal Hessian Directions

Theorem 3.1 can be used to suggest estimates for pHd's from an iid sample $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$. Let $\bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}$ be the sample mean and sample covariance of $\mathbf{x}$. Then:

1. Form the matrix $\hat{\boldsymbol{\Sigma}}_{yxx} = 1/n \sum_{i=1}^{n} (y_i - \bar{y})(\mathbf{x}_i - \bar{\mathbf{x}}) \times (\mathbf{x}_i - \bar{\mathbf{x}})'$.

2. Conduct an eigenvalue decomposition of $\hat{\Sigma}_{yxx}$ with respect to $\hat{\Sigma}_x$:

$$\hat{\Sigma}_{yxx}\hat{b}_{yj} = \hat{\lambda}_{yj}\hat{\Sigma}_x\hat{b}_{yj}, \qquad j = 1, \ldots, p$$
$$|\hat{\lambda}_{y1}| \geq \cdots \geq |\hat{\lambda}_{yp}|.$$

Instead of this y-based method, we may use Theorem 3.2 and suggest the same procedure, but with $y_i - \bar{y}$ replaced by the residual $\hat{r}_i = y_i - \hat{a} - \hat{b}_{ls}'\mathbf{x}_i$, where $\hat{a}$, $\hat{b}_{ls}$ are the least squares estimates for the linear regression of y against $\mathbf{x}$:

1. Find the residuals $\hat{r}_i$, $i = 1, \ldots, n$.
2. Form the matrix $\hat{\Sigma}_{rxx} = 1/n \sum_{i=1}^{n} \hat{r}_i(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$.
3. Conduct the eigenvalue decomposition of $\hat{\Sigma}_{rxx}$ with respect to $\hat{\Sigma}_x$:

$$\hat{\Sigma}_{rxx}\hat{b}_{rj} = \hat{\lambda}_{rj}\hat{\Sigma}_x\hat{b}_{rj}, \qquad j = 1, \ldots, p$$
$$|\hat{\lambda}_{r1}| \geq \cdots \geq |\hat{\lambda}_{rp}|.$$

Remark 3.2. Corollary 3.2 suggests yet another way of finding the edr directions. First, fit y by a quadratic polynomial of $\mathbf{x}$. The Hessian matrix for the fitted quadratic function, say $\hat{\mathbf{B}}$, can be easily formed from the estimated quadratic and cross product terms. Next, take the eigenvalue decomposition of the matrix $\hat{\mathbf{B}}\hat{\Sigma}_x$ to get the pHd's. This method (hereinafter called the q-based pHd) is related to the canonical analysis for exploring and exploiting quadratic response surfaces where the eigenvalue decomposition is taken for the Hessian matrix of the fitted quadratic surface *with respect to the identity matrix.* Box (1954) and Box and Draper (1987), for example, have illustrated well how their techniques have been used successfully to locate stationary points and to obtain a parsimonious description of these points in many designed chemical experiments. A fundamental assumption they made is that the response surface is well approximated by a quadratic polynomial in the region covered by the design points. But our study in this article suggests that this restriction can be relaxed because the surface fitting often can be improved by many low-dimension smoothing or model-fitting techniques after we have found the edr space. Further study on this subject is underway regarding possible ill effects due to estimating too many quadratic terms when p is large compared to the sample size.

Remark 3.3. Härdle and Stoker (1989) proposed an estimate of the average slope for the regression function without the normality assumption of $\mathbf{x}$. Their method can be extended to estimate the average Hessian $\bar{\mathbf{H}}_x$. But the needed estimation for the score function of $\mathbf{x}$ is severely subject to the curse of dimensionality.

4. SAMPLING PROPERTIES

We shall demonstrate the root n consistency of our estimates, compute the expected value for a closeness measure between the estimated edr space the the true edr space, and find two significance tests for determining the number of components K. Our development follows Li (1991) closely. Proofs are given in Appendix B.

First, because estimates for Σ_{yxx} and Σ_{rxx} are based on the method of moments, root n consistency needs no proof. To evaluate how close the estimated edr space, $\hat{\mathcal{B}}_y$ ($\hat{\mathcal{B}}_r$), is to the true edr space, we consider the measure used in Li (1991), namely, the squared trace correlation, $R^2(\hat{\mathcal{B}}_y)$ ($R^2(\hat{\mathcal{B}}_r)$), which is the average of the squared canonical correlation coefficients between $\hat{b}_{yj}'\mathbf{x}$, $j = 1, \ldots, K$ ($\hat{b}_{rj}'\mathbf{x}$, $j = 1, \ldots, K$), and $\beta_j'\mathbf{x}$, $j = 1, \ldots, K$. The closer to 1 that this measure is, the sharper the viewing angle will be. The following theorem gives an approximation for the expected value of this quantity.

Theorem 4.1. Assume that $\mathbf{x}$ is normal and that Σ_{yxx} has rank K. Then, under (2.3), we have

$$R^2(\hat{\mathcal{B}}_y) = 1 - (p - K)n^{-1} \times \sum_{j=1}^{K} (-1 + \lambda_j^{-2}\mathrm{var}((y - \mu_y)b_j'(\mathbf{x} - \mu_x))) + o(n^{-1}) \quad (4.1)$$

and

$$R^2(\hat{\mathcal{B}}_r) = 1 - (p - K)n^{-1} \times \sum_{j=1}^{K} (-1 + \lambda_j^{-2}\mathrm{var}(rb_j'(\mathbf{x} - \mu_x))) + o(n^{-1}). \quad (4.2)$$

Theorem 4.2. Under the same conditions as in Theorem 4.1, we have

$$n^{1/2} \sum_{j=k+1}^{p} \hat{\lambda}_j \sim N(0, 2(p - k)\mathrm{var}(\cdot)) \quad (4.3)$$

$$n \sum_{j=k+1}^{p} \hat{\lambda}_j^2 \sim 2\,\mathrm{var}(\cdot)\chi^2_{(p-k+1)(p-k)/2}, \quad (4.4)$$

where $\hat{\lambda}_j$ denotes $\hat{\lambda}_{yj}$ or $\hat{\lambda}_{rj}$, $\mathrm{var}(\cdot)$ equals var y or var r, and $k \leq K$ is the rank of the weighted covariance matrix $\Sigma_{yxx} = \Sigma_{rxx}$.

We can use Theorem 4.2 to suggest whether a component found is likely to be real or not, by estimating var y (var r) with the sample variance of y (the mean squares for residuals $(n - p)^{-1} \sum_{i=1}^{n} \hat{r}_i^2$).

Remark 4.1. Theorem 4.2 suggests that the residual-based estimate is more powerful in detecting a real component because var r is typically smaller than var y. But Theorem 4.1 indicates that in terms of offering a sharper viewing angle, there is no clear winner between the two.

Remark 4.2. For the q-based method suggested in Remark 3.2, the asymptotic result will be similar. We need only replace r by the residual of the quadratic fit.

5. EXTENSION

We can extend the list of possible estimates by considering nonlinear transformations of y before applying the methods in Section 3. For example, we may want to trim out large y values to decrease the sensitivity to outliers. We may also use the absolute value of the residual to form the estimate.

We now draw a connection between what is suggested here and second-moment-based SIR methodology. First, consider the population case. Partition the range of y into $\mathbf{H}$ intervals, I_h, $h = 1, \ldots, \mathbf{H}$. Then apply the indicator transformation $\tilde{y} = \delta_h(y) = 1$, or 0, depending on whether or not y falls into the hth interval. Denote $p_h = P\{y \in I_h\}$.

Then we have

$$\Sigma_{\tilde{y}\mathbf{x}\mathbf{x}} = \mathbf{E}(\delta_h(y) - p_h)(\mathbf{x} - \mu_\mathbf{x})(\mathbf{x} - \mu_\mathbf{x})'$$
$$= p_h[\mathbf{E}((\mathbf{x} - \mu_\mathbf{x})(\mathbf{x} - \mu_\mathbf{x})' \mid y \in I_h) - \Sigma_\mathbf{x}].$$

Now we can apply Theorem 3.1 to the transformed variable $\tilde{y}$.

Corollary 5.1. Assume that $\mathbf{x}$ is normal. For each slice h, conduct the eigenvalue decomposition of the sliced second-moment matrix $\mathbf{E}((\mathbf{x} - \mu_\mathbf{x})(\mathbf{x} - \mu_\mathbf{x})' \mid y \in I_h)$ with respect to $\Sigma_\mathbf{x}$. Then the eigenvectors with eigenvalues distinct from 1 are edr directions.

The sample version is easy to obtain. First, form the sliced second-moment matrix $(n_h - 1)^{-1} \sum_{\mathbf{x}_i \in I_h} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$, where n_h is the number of cases in the hth slice. Then take the eigenvalue decomposition of this matrix with respect to $\hat{\Sigma}_\mathbf{x}$. Let the eigenvalues $\hat{\lambda}_{hj}$'s be arranged to have the order $|\hat{\lambda}_{h1} - 1| \geq \cdots \geq |\hat{\lambda}_{hp} - 1|$.

Large sample results can be derived from Theorems 4.1 and 4.2. Because $\tilde{y}$ is dichotomous, the resulting formulas are in fact simpler. In particular, we have $\text{var}(\cdot) = \text{var}\ \tilde{y} = p_h(1 - p_h)$ for (4.3) and (4.4).

The sliced second-moment matrix $\mathbf{E}((\mathbf{x} - \mu_\mathbf{x})(\mathbf{x} - \mu_\mathbf{x})' \mid y \in I_h)$ discussed earlier is closely related to the conditional covariance $\text{cov}(\mathbf{x} \mid y \in I_h)$, the core of some specific suggestions for applying second moments in the SIR approach (Cook and Weisberg 1991; rejoinder to Li 1991). The difference between these two matrices is just a rank 1 matrix, $(m_h - \mu_\mathbf{x})(m_h - \mu_\mathbf{x})'$, where $m_h = \mathbf{E}((\mathbf{x} - \mu_\mathbf{x}) \mid y \in I_h)$ is the core of the first-moment-based SIR estimate.

Remark 5.1—Limitations. All methods have limitations. SIR and pHd are no exceptions. We shall identify cases in which edr directions cannot be estimated from any transformation version of pHd. For simplicity of discussion, take $K = 1$ and concentrate on the case where $\mathbf{E}(\mathbf{x} \mid y) = \mathbf{E}\mathbf{x}$, which is the condition to nullify the power of the first-moment-based SIR. Under this condition, the least squares estimate b_{ls} is equal to 0. Thus the residual-based estimate is the same as the y-based estimate. We are interested in knowing when the weighted covariance matrix $\Sigma_{\mathbf{T}(y)\mathbf{x}\mathbf{x}} = \mathbf{E}(\mathbf{T}(y) - \mathbf{E}\mathbf{T}(y))(\mathbf{x} - \mu_\mathbf{x})(\mathbf{x} - \mu_\mathbf{x})'$ will be degenerated to 0 for any transformation $\mathbf{T}(y)$, in which case no edr directions can be detected. The following Lemma offers an answer.

Lemma 5.1. Assume that $\mathbf{x}$ is normal and consider (1.1) with $K = 1$. Then,

$$\Sigma_{\mathbf{T}(x)\mathbf{x}\mathbf{x}} = 0, \quad \text{for any transformation } \mathbf{T}(y),$$

if and only if

$$\mathbf{E}[(\beta_1'(\mathbf{x} - \mu_\mathbf{x}))^2 \mid y] \quad \text{does not depend on } y. \tag{5.1}$$

It is easy to interpret this result from the inverse regression point of view. In general, the conditional distribution of $\beta_1'\mathbf{x}$ given y should depend on y under (1.1). But if this dependence is only through moments of order higher than 2, then (5.1) will hold; pHd or any first- or second-moment-based SIR will not find any significant directions. This leaves room for introducing more complicated procedures based on features other than the first two moments of the inverse regression.

6. LINEAR CONDITIONAL EXPECTATION FOR X

In this section we shall study the behavior of the eigenvectors b_j's as defined in Theorem 3.1, under the following weaker condition on the distribution of $\mathbf{x}$ used in Li (1991). For any $b \in R^p$

$$\mathbf{E}(b'\mathbf{x} \mid \beta_j'\mathbf{x}, j = 1, \ldots, K) \quad \text{is linear in } \beta_j'\mathbf{x}\text{'s.} \tag{6.1}$$

This condition includes any elliptical symmetric distributions. See Remark 6.2 for more discussion.

Theorem 6.1. Under (1.1) and (6.1), the edr space $\mathcal{B}$ is invariant under the transformation induced by the matrix $\Sigma_\mathbf{x}^{-1}\Sigma_{y\mathbf{x}\mathbf{x}}$, in the sense that

$$\{\Sigma_{y\mathbf{x}\mathbf{x}} b : b \in \mathcal{B}\} \subseteq \{\Sigma_\mathbf{x} b : b \in \mathcal{B}\}.$$

Because the invariance spaces of a matrix are spanned by its eigenvectors, this theorem suggests that the eigenvectors b_j's can be used to find edr directions. For instance, if $K = 1$, then one of the b_j's must be an edr direction unless $\Sigma_{y\mathbf{x}\mathbf{x}}\beta_1 = 0$, or equivalently,

$$\text{cov}(y, (\beta_1'\mathbf{x} - \mu_\mathbf{x})^2) = 0. \tag{6.2}$$

Thus, although it is not clear which b_j is the right one to use, for the purpose of data visualization we can display all p bivariate plots, y against b_j's, and then choose the one that shows the most interesting structure. If (6.2) does happen, then we cannot find the edr direction by this method. In such a situation, we can still hope that some transformation on y might avoid (6.2). Suitably combining second-moment SIR estimates is likely to be more productive. Likewise, the case where $K = 2$ leads to viewing $\binom{p}{2}$ sets of three-dimensional plots. Some troubles may begin to occur when K is larger, because the combination number increases quickly. But our experience shows that eigenvalues still offer good indication of the importance of the associated directions, even though pathological cases can exist.

Now we consider the elliptically symmetric distribution in further detail.

Theorem 6.2. Assume that $\mathbf{x}$ follows an elliptically symmetric distribution. Under (1.1), for the eigenvalues λ_j defined in Theorem 3.1, at least $p - K$ of them take a common value. In addition, all other eigenvectors are edr directions if $p - K$ is greater than K.

This theorem does not say anything about the size of the common eigenvalue. But we expect it to be small for most cases. One can see this from the last term of the expression in the proof of this theorem given in Appendix A.6. If p is large, the random variable, $\|\mathbf{x}\|^2 - \|P_1\mathbf{x}\|^2 = \|P_2\mathbf{x}\|^2$, becomes nearly independent of $P_1\mathbf{x}$ and hence is nearly independent of y.

Remark 6.1. Clearly, our discussion applies to the residual-based eigenvectors as defined in Theorem 3.2.

Remark 6.2. Li (1989b, 1991) argued that (6.1) is expected to hold approximately for many data sets, based on

the result of Diaconis and Freedman (1984), who showed that almost all low-dimension projections of high-dimension data are approximately normal. This is now more rigorously stated and proved in Hall and Li (1992). Li (1989b) demonstrated the ability of SIR to find the directions for which (6.1) are severely violated. Brillinger (1991) demonstrated how to achieve (6.1) by normal subsampling; see also Li and Duan (1989) for more discussion.

7. PROJECTION INDEX

Like SIR, we can also discuss the pHd method in terms of projection pursuit. (See Huber 1985 for a comprehensive account.) Projection pursuit finds interesting directions by maximizing a projection index defined for each direction. Different projection indices lead to different projection pursuit methods.

First, observe that for any $b \in R^p$, we have $b'\Sigma_{yxx}b = \operatorname{cov}(y, (b'(\mathbf{x} - \mu_x))^2)$. It follows that the first eigenvector b_1, as defined in Theorem 3.1, is the solution of the maximization problem:

$$\max_{\operatorname{var} b'\mathbf{x}=1} |\operatorname{cov}(y, (b'(\mathbf{x} - \mu_x))^2)|.$$

Subject to being uncorrelated to the proceeding solutions, maximization can be achieved by the b_j's. Therefore, the y-based pHd method can be viewed as a projection pursuit method if, for each projected variable $u = b'\mathbf{x}$, the projection indexes is set to be $|\operatorname{cov}(y, (u - \mu_u)^2)|$, where μ_u is the mean of u.

Similarly, the residual-based pHd's are the directions found by using $|\operatorname{cov}(r, (u - \mu_u)^2)| = |\operatorname{cov}(r, u^2)|$ as the projection index. This index is related to the index given by the R-squared value for fitting r with a quadratic polynomial of u, which, after a straightforward derivation, turns out to be

$$\left|\operatorname{corr}\left(r, u^2 - \frac{\operatorname{cov}(u^2, u)}{\operatorname{var} u}u\right)\right|^2.$$

Indeed, when $\mathbf{x}$ is elliptically symmetric with mean 0, $\operatorname{cov}(u^2, u) = 0$ and the standard deviation of u^2 is the same for any direction b, implying that maximum cov is the same as maximizing corr. Therefore, these two projection indices lead to the same solution.

As we have seen from the preceding discussion, the projection index for pHd is more like fitting a quadratic function on each direction. This can be compared to the more complicated projection pursuit regression (PPR) index based on nonparametric curve fitting (Friedman and Stuetzle 1981). It is interesting to observe that simple indices like those discussed here can still work well for a complicated model like (1.1), which can be very different from a quadratic one; see Section 8 for some specific examples. For users of PPR, it would then be interesting to find out how much additional help PPR can offer to solve those cases where pHd fails.

The limitation of pHd is also better understood from the associated projection index: pHd can fail if the R-squared value of the quadratic fit is too small. For such cases, transformation on y can be applied to increase the R-square value (see Remark 7.1). On the other hand, incorporation of transformation on y is not so simple for the already complicated PPR, where many obscured aspects of smoothing might affect the algorithm substantially.

Remark 7.1. Li (1989b) showed that the first-moment-based SIR can be derived from the projection pursuit viewpoint by introducing the following projection index for a projection direction b:

$$\max_{T(\cdot)} \operatorname{corr}^2(T(y), b'\mathbf{x}),$$

where $T(\cdot)$ can be any transformation. This can be compared to the PPR index

$$\max_{T(\cdot)} \operatorname{corr}^2(y, T(b'\mathbf{x})).$$

One may wish to propose a new index based on transformations on both sides, but the question is how to do this effectively.

We can apply transformation on y to pHd. For each direction b with $\operatorname{var}(b'\mathbf{x}) = 1$, define the projection index as

$$\max_{\operatorname{var} T(y)=1} \operatorname{cov}(T(y), (b'(\mathbf{x} - \mu_x))^2).$$

Maximization can be achieved by taking the transformation $T(y) = c_b^{-1}\mathrm{E}((b'(\mathbf{x} - \mu_x))^2 | y)$, where c_b^2 equals $\operatorname{var}(\mathrm{E}((b'(\mathbf{x} - \mu_x))^2 | y))$. The maximum equals c_b. Maximization of this projection index is related to the SIR-II estimate; see the rejoinder to Li (1991).

Remark 7.2. We could apply transformation to the q-based pHd method as well. To find the "optimal" transformation, we can apply SIR for y against the linear and the second-order variables, $x_1, \ldots, x_p, x_1^2, \ldots, x_p^2, x_1x_2, \ldots, x_{(p-1)}x_p$. Details will be reported elsewhere.

8. EXAMPLES

Three simulation examples are reported to demonstrate the performance of the pHd method. The fourth example applies SIR and pHd to help analyze the ozone data taken from Breiman and Friedman (1985). All works were done on a MacII, using Xlisp.stat (Tierney 1990).

Example 8.1. The model used to generate the data is given by

$$y = \cos(2\beta_1'\mathbf{x}) - \cos(\beta_2'\mathbf{x}) + .5\varepsilon, \qquad (8.1)$$

where $\mathbf{x}$ has $p = 10$ dimensions, $\beta_1 = (1, 0, \cdots)'$, $\beta_2 = (0, 1, 0, \cdots)'$, and all coordinates of $\mathbf{x}$ and ε are iid standard

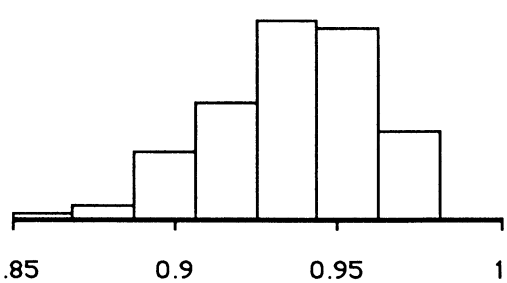

Figure 1. Histogram of $R^2(\hat{B})$ for 100 Runs.

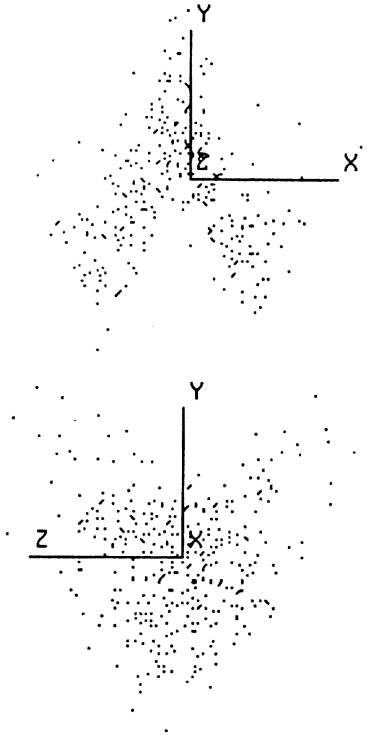

Figure 2. Best Views for Example 8.1.

normal random variables. For $n = 400$, we study the performance of the residual-based estimate $\hat{b}_{r1}$'s after 100 simulation runs. A histogram of the closeness measure $R^2(\hat{\mathcal{B}}_r)$ is given in Figure 1. The views from the first two directions found in a typical run are given in Figure 3, compared to the best views, views from β_1, β_2, given in Figure 2. One could better appreciate how the views are similar to each other by spinning the two rotation plots and view the data cloud from all angles. Table 1 gives the output directions, the absolute eigenvalues, and the p values based on Theorem 4.2. The jth p value is calculated under the null hypothesis that there are $j - 1$ components that are real, so a small value supports the genuineness for the jth direction. Only two directions are found to be significant.

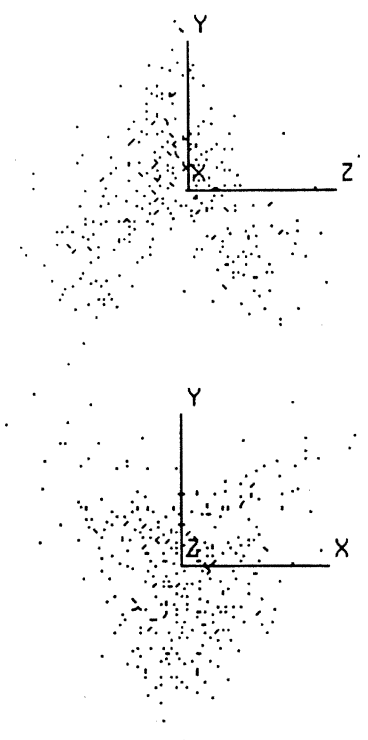

Figure 3. Views Found by pHd for Example 8.1.

Table 1. One Simulation Output for Model (8.1)

$\hat{b}_{r1}$	.07	−.93	−.05	.09	−.25	−.03	−.10	−.19	−.11	.05
$\hat{b}_{r2}$	.97	.01	−.05	−.05	.04	−.01	−.17	.13	.04	.15
Eigenvalues	.78	.63	.26	.20	.15	.13	.11	.06	.05	.01
p values	0	0	.71	.92	.97	.97	.97	.99	.94	.90

Example 8.2. This example is used to study how violation of the linear conditional expectation condition (6.1) might affect the estimation. We consider the model

$$y = \beta'\mathbf{x} \sin(2\beta'\mathbf{x}), \tag{8.2}$$

where $\mathbf{x}$ is uniform on a 10-dimensional cube, $[-\frac{1}{2}, \frac{1}{2}]^{10}$. First, when a direction for β is chosen at random, the pHd method finds the true direction as well as if $\mathbf{x}$ were indeed normal. This confirms Remark 6.2. Instead of reporting these favorable cases, however, we want to study the worst situation. Consider $\beta'\mathbf{x}$ as a sum of p independent random variables and borrow insight from the central limit theorem. We can anticipate the worst case to happen when β is 0 on all but two coordinates, the case when $\beta'\mathbf{x}$ is the least normal in a sense. Now for those directions on the plan spanned by first two coordinates, there are four good directions for which the linear conditional expectation condition holds: the two coordinate axes and the two diagonal lines. Hence we decide to choose $\beta = (1, 2, 0, \cdots)'$ on the grounds that this direction is midway between the two good directions $(1, 1, \cdots)'$ and $(0, 1, 0, \cdots)'$. We generate $n = 400$ observations and use the y-based method to find the edr direction. From the output given in Table 2, we see some bias in the first direction found. But a close look at the p values reveals that the second direction is marginally significant. In fact, a combination of the first two directions, as shown in Figure 4 (right side), yields a high-quality reconstruction of the true curve, shown on the left side. By pitching the rotation plots used to produce Figure 4 until the y axis is perpendicular to the screen. Figure 5 shows how well the distribution for the first two projected directions matches the distribution of the first two coordinates of $\mathbf{x}$. This demonstrates the potential of our method to find directions b that violate the linear conditional expectation most seriously. One can also argue that under our parameter specification, we can view (8.2) as a two-component model with $\beta_1 = (1, 0, \cdots)'$ and $\beta_2 = (0, 1, \cdots)'$. The linear conditional expectation condition is now satisfied, explaining why we can find two directions. Of course, the p

Table 2. One Simulation Output for Model (8.2)

$\hat{b}_{y1}$	−.54	−.84	.03	−.014	−.01	−.05	−.03	−.06	.04	−.00
$\hat{b}_{y2}$	−.78	.52	−.07	−.05	−.19	.02	.04	−.05	.13	.21
Eigenvalues	.78	.63	.26	.20	.15	.13	.11	.06	.05	.01
p values	0	0	.71	.92	.97	.97	.97	.99	.94	.90

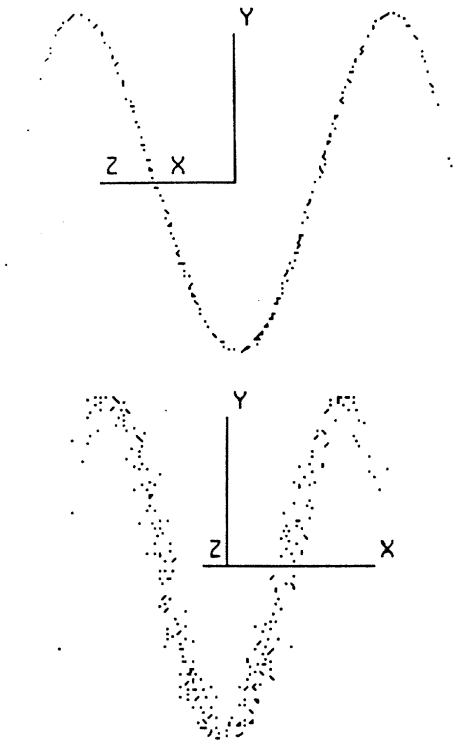

Figure 4. Best View (top) and pHd's View (bottom) for Example 8.2.

values are only suggestive, because of the violation of normality. Judgment based on the pattern of the whole sequence of p values should be more informative than the individual numbers. We see the drastic increase from .07 to .70 as a strong indication that the third component is not likely to be informative. The residual-based method is also attempted; it yields almost the same result as the one reported here. We conclude this example by reporting that as we enlarge the range of $\mathbf{x}$ so that the response curve looks more like an M-shape, pHd begins to lose power in detecting the edr direction. This is because the conditional variance of $\beta'\mathbf{x}$ given y becomes more homogeneous, and Lemma 5.1 begins to take effect. It would be interesting to see how well PPR works in such cases.

Example 8.3. This example shows how simple transformations can help pHd. We consider the model

$$y = \frac{1}{3}(\beta_1'\mathbf{x})^3 - (\beta_1'\mathbf{x})(\beta_2'\mathbf{x})^2$$

for generating the data. The surface of this function is known as the "monkey saddle;" see Figure 6. We take $\beta_1 = (1, 0, \cdots)'$ and $\beta_2 = (0, 1, 0, \cdots)'$ and generate $n = 300$ data points. First, a histogram of y (Fig. 7) suggests a long-tail distribution. To avoid the dominance of large y in the analysis, we cut out those cases with the absolute value of y greater than 2. This leaves 261 points in the sample. We find the y-based and the residual-based methods unsuccessful, as indicated by the p values. Then we take the absolute value transformation on the residuals, treat them as y, and proceed with the pHd method. Two directions are found significant (see Table 3). The best views for y and the views based on the estimated directions are given in Figures 8–9. Three branches going upward and downward in the monkey saddle can be identified well by spinning these plots on the computer. Other transformations and other methods of handling large y values are worth trying.

Example 8.4. This example demonstrates how the pHd method can be used to complement SIR. For illustration, we take a data set from Breiman and Friedman (1985), the data for studying the atmospheric ozone concentration in the Los Angeles basin. We use the daily measurement of ozone concentration in Upland as the output variable y and

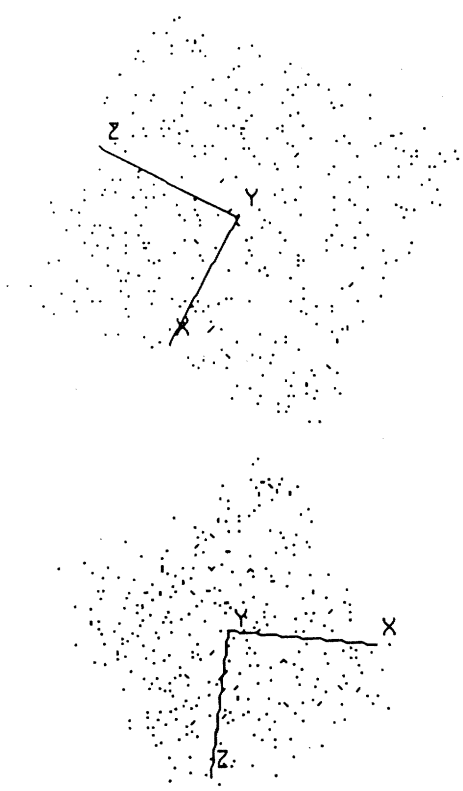

Figure 5. Distribution of the First Two Coordinates of x (top), Compared With That for the First Two pHd Directions (bottom).

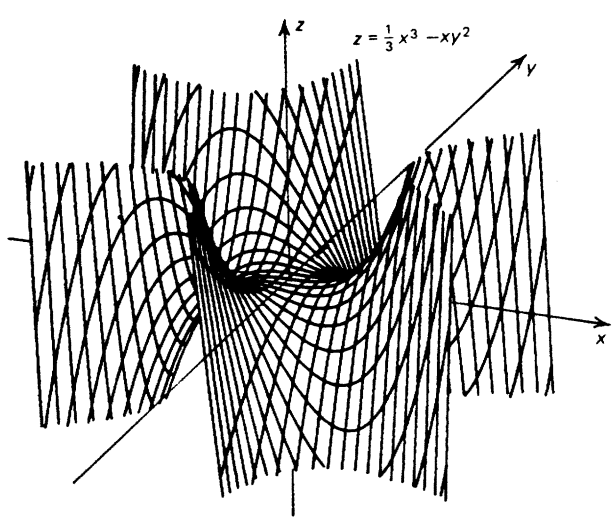

Figure 6. A Monkey Saddle.

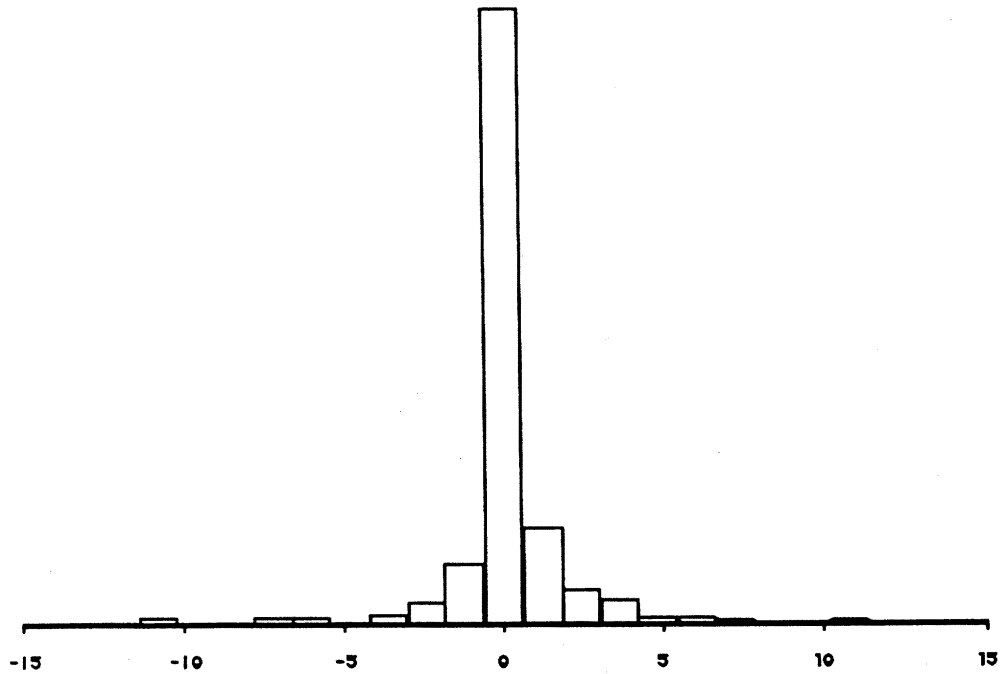

Figure 7. A Histogram for y Generated by (8.3).

want to find its relationship with eight meteorological variables (see Table 4). There are $n = 330$ observations in the study. First, we apply SIR to the data and find one significant component. This component is almost identical to the $\hat{b}'_{ls}\mathbf{x}$, the component found by the linear least squares fitting. For certain slice sizes, we can find a marginally significant second component as well, but we decide to ignore this. We then use a forward selection method to find the important variables contributing to the first component. Three variables, x_1, x_2, and x_6, are found that explain more than 99% of the total variation of the first component. We run SIR again using only x_1, x_2, and x_6 as the input variables. The scatterplot of y against $\hat{b}'_s\mathbf{x}$, the first component found, is given in Figure 10(a). We use 30 slices here for SIR, but other choices yield almost identical scenes. The correlation between $\hat{b}'_{ls}\mathbf{x}$ and $\hat{b}'_s\mathbf{x}$ is above .99 as well. The value of $\hat{b}_s$ is given in the first row of Table 5.

A quadratic trend is visible in Figure 10(a). After fitting a quadratic polynomial,

$$y = c_0 + c_1 u_1 + c_2 u_1^2 + \varepsilon, \tag{8.4}$$

where u_1 denotes the variable $\hat{b}'_s\mathbf{x}$. A summary is given in Table 6, and the residual plot is given in Figure 10(b).

Now we take a closer look at the residual by applying the pHd method, treating the residual as y. One component is found to be significant. Again we use a forward selection procedure to find that this component can be explained by x_3, x_5, and x_6 with about 90% R-squared. (If including x_8, then R-squared can be about 96%.) We then rerun pHd, using only x_3, x_5, and x_6 as the regressors. Again one component is found, denoted as $\hat{b}_{phd}$ (see Table 5). Figure 10(c) gives the plot of the residual against this component. A quadratic pattern in this figure is detected by eye and is confirmed by fitting a quadratic polynomial (see Table 7). For comparison, we also apply SIR to the residual and find one significant component, which gives the view in Figure 10(d).

Table 3. One Simulation Output for Model (8.3)

$\hat{b}_1$	−.37	.91	−.06	−.02	−.05	.12	−.08	.07	.05	−.07
$\hat{b}_2$	.91	.29	−.18	.10	.07	−.16	.02	−.02	.08	.08
Eigenvalues	.32	.20	.14	.10	.08	.07	.04	.03	.01	.01
p values	0	.03	.58	.92	.96	.99	.99	.99	.98	.81

Figure 8. Best Views for Model (8.3).

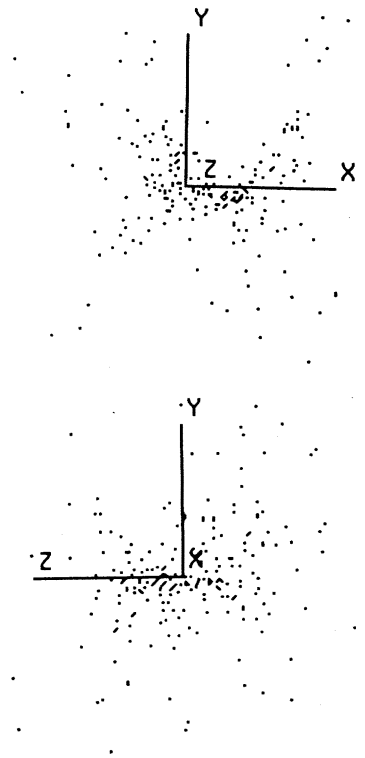

Figure 9. Views Found by pHd for (8.3).

Table 4. Variables Used for the Ozone Data

x_1	Sandburg Air Force Base temperature (C°)
x_2	Inversion base height (ft)
x_3	Dagget pressure gradient
x_4	Visibility (miles)
x_5	Vandenburg 500 millibar height (m)
x_6	Humidity (percent)
x_7	Inverse base temperature (F°)
x_8	Wind speed (mph)

These two viewing angles have a low correlation of about .2. They do reveal different patterns. SIR fails to find the view of Figure 10(c), because the pattern is more or less symmetric about the y axis. On the other hand, the untransformed pHd method used here cannot detect Figure 10(d), because it is symmetric about the x axis. We also take transformation on the residuals (e.g., the absolute value, or the logarithm of the absolute value) before applying the pHd method to further study the heteroscedasticity aspect of the residuals. The results are not reported here.

9. CONCLUSIONS

In this article we have introduced another method for finding edr directions. This method is based on the estimation of the pHd's. We have discussed when and why this method

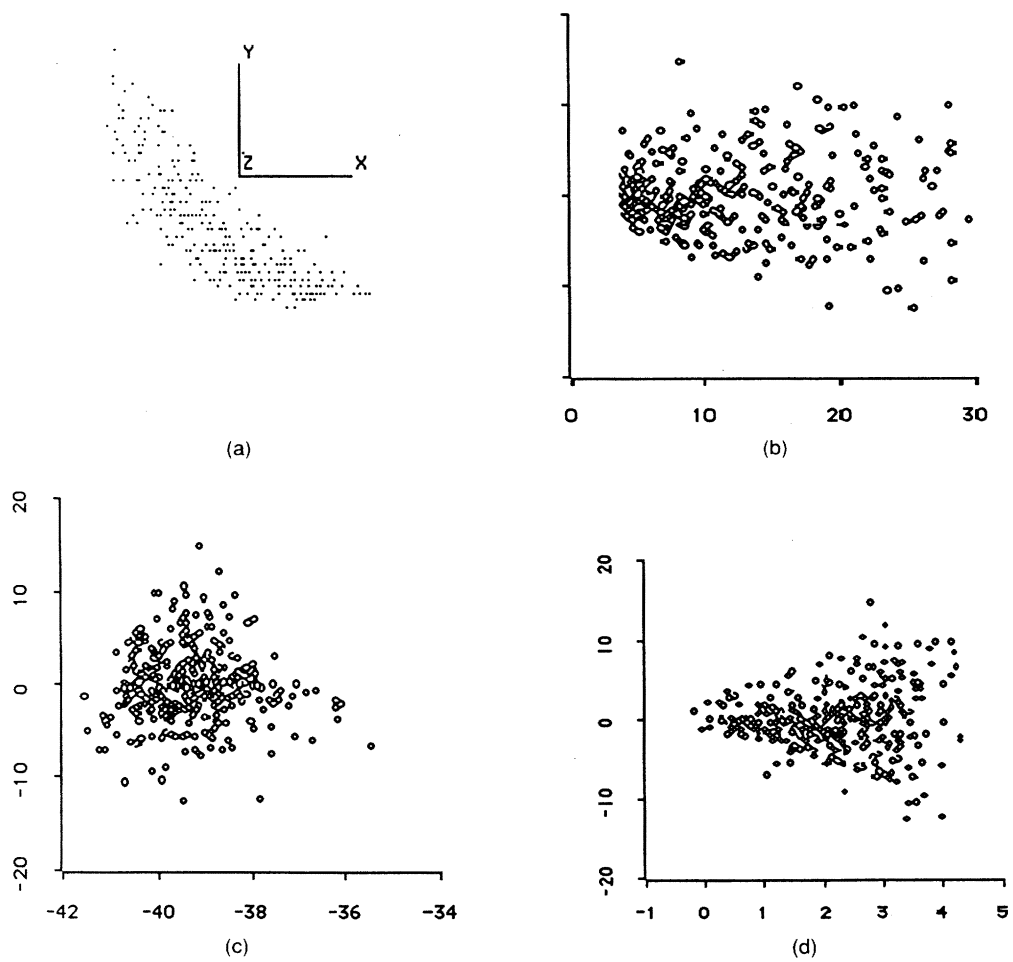

Figure 10. Plots Used in the Ozone Study. (a) Ozone against the first direction of SIR; (b) Residual plot for model (8.4); (c) Residuals against the direction found by pHd; (d) Residuals against the direction found by SIR.

Table 5. Significant Directions Found for the Ozone Data

$\hat{b}_s$	.046	1.7e-4	0	0	0	−.014	0	0
Eigenvalues	.76	.11	.08	—	—	—	—	—
p values	0	.22	.42	—	—	—	—	—
$\hat{b}_{phd}$	0	0	−.03	0	−.0072	.051	0	0
Eigenvalues	1.5	.68	.33	—	—	—	—	—
p values	0	.12	.29	—	—	—	—	—

Table 7. A Summary for Quadratic Regression for the Residual of (8.4)

Constant	−911 (216)
Linear	−46.9 (11.1)
Quadratic	−.60 (.14)
R-squared	.05

is likely to be useful. We have provided some versions for the implementation.

This method can be used to complement the first-moment-based SIR, which tends to be more stable than the pHd method because the latter involves second moments. On the other hand, the pHd method is more useful for symmetric cases that the first-moment-based SIR fails to handle. Of course, we often do not know how symmetric the case under study is before analyzing the data, so we cannot decide which one may work better in the beginning. It may also happen that each method finds only part of the edr directions. In view of this, our suggestion is to use both SIR and pHd all the time. In addition to any subjective decision that can be made after inspecting the scatterplots found by either method, we can also rely on significance tests to judge whether or not a direction is real. We have demonstrated one possible way of analyzing the ozone data, using both SIR and pHd.

We have mentioned the gain of robustifying the procedure by trimming out large values in y. The more difficult task is to handle outliers in $\mathbf{x}$. A two-stage procedure like the following is worth trying: (1) Run one round of pHd and plot y against the first few components, and (2) exclude outliers in the plots and then run the pHd method again.

Any method in statistics has its own limitation. We do not anticipate that SIR and pHd will always find structures. This is why it is important to have a measure to indicate whether or not component found is likely to be spurious. It is also important to know, at least at the conceptual level, what types of structures are likely to be found or missed by our methods. For SIR the shape of inverse regression curve is the determining factor. For pHd the determining factor is the average curvature of the regression surface as measured by the average Hessian matrix. Thus for $K = 1$, if the regression curve has too many turns, then the positive curvature on convex portions of the curve is likely to cancel out the negative curvature on the concave portions. This leaves room for inventing other methods to handle such situations. We have thought about subsampling techniques, for example. A suitable split of the data set, either linearly or spherically on the input variables, is worth trying. A more systematic strategy for splitting is to be studied.

We have also related our method to simple projection indices. In addition to the phase of statistical inference, the estimated pHd's can also be considered as descriptive statistics. Some messages on projection pursuit from our limited experience with SIR and pHd's can be summarized in the following list. Some of these may echo those found in the context of exploratory projection pursuit (see, for example, Friedman 1987; Hall 1989b; and Huber 1985).

Table 6. A Summary for Quadratic Regression (8.4)

Constant	7.5 (2.1)
Linear	5.20 (1.42)
Quadratic	1.86 (.22)
R-squared	.74

1. Under (1.1) there is more than one projection index that can be used to find edr directions.
2. Any index has its own strength and weakness; it is important to know what patterns it might miss.
3. Use of two simple, complementary indices may be preferable to a single complex index.
4. For any index, it is necessary to have a criterion for suggesting how spurious a projection might be.
5. Even if data fitting might be one of the ultimate goals, good projection indices are not necessarily confined to those directly translated from goodness-of-fit criteria; it is useful to distinguish the issue of data fitting from the issue of dimension reduction.
6. Sampling properties are easier to analyze for some projection indices than for others. Other things being equal, this becomes an important factor in guiding the choice of projection indices.

APPENDIX A: PROOFS FOR SECTIONS 2, 3, AND 6

A.1 Proof of Lemma 2.1

The restriction on $\mathbf{A}$ implies that we can represent $\mathbf{A}$ as $\mathcal{O}\Sigma_{\mathbf{x}}^{-1/2}$ for some unitary matrix $\mathcal{O}$. Plugging this into (2.2), our problem reduces to finding a unitary vector $\mathbf{a} = \mathcal{O}'e_1$, so that $|\mathbf{a}'\Sigma_{\mathbf{x}}^{1/2}\bar{\mathbf{H}}_{\mathbf{x}}\Sigma_{\mathbf{x}}^{1/2}\mathbf{a}|$ is maximized. The solution $\mathbf{a}$ is the largest eigenvector (in terms of the absolute value of the eigenvalues) for the eigenvalue decomposition of $\Sigma_{\mathbf{x}}^{1/2}\bar{\mathbf{H}}_{\mathbf{x}}\Sigma_{\mathbf{x}}^{1/2}$, which is seen to be $\Sigma_{\mathbf{x}}^{1/2}b_1$, after multiplying $\Sigma_{\mathbf{x}}^{1/2}$ on both sides of the identity (2.1). Now we see that the first row of $\mathbf{A}$ equals

$$e_1'\mathbf{A} = e_1'\mathcal{O}\Sigma_{\mathbf{x}}^{-1} = \mathbf{a}'\Sigma_{\mathbf{x}}^{-1/2} = (b_1'\Sigma_{\mathbf{x}}^{-1/2})\Sigma_{\mathbf{x}}^{-1/2} = b_1',$$

as desired. The proof is now complete.

A.2 Proof of Lemma 2.2

Let $\mathbf{B} = (\beta_1, \ldots, \beta_K)$ and $\mathbf{t} = (\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x})' = \mathbf{B}'\mathbf{x}$. Then $f(\mathbf{x}) = h(\mathbf{B}'\mathbf{x})$ and, by the chain rule, $\mathbf{H}_{\mathbf{x}}(\mathbf{x}) = \mathbf{B}\mathbf{H}_{\mathbf{t}}(\mathbf{t})\mathbf{B}'$. Now it is clear that for any direction, v, in the orthogonal complement of $\mathcal{B}$, we have $\mathbf{H}_{\mathbf{x}}(\mathbf{x})v = 0$. Hence the rank of $\bar{\mathbf{H}}_{\mathbf{x}}$ is at most K. In addition, for any pHd b_j with $\lambda_j \neq 0$, we have $0 = v'\bar{\mathbf{H}}_{\mathbf{x}})\Sigma_{\mathbf{x}}b_j = v'\lambda_j b_j$, implying that b_j is orthogonal to v. Therefore, b_j falls into the edr space $\mathcal{B}$. The proof is complete.

A.3 Proof of Corollary 3.1

First, standardize $\mathbf{x}$ to have mean 0 and the identity covariance by an affine transformation like $\mathbf{z} = \Sigma_{\mathbf{x}}^{-1/2}(\mathbf{x} - \mu)$. Define $\bar{\mathbf{H}}_{\mathbf{z}}$ to be the average Hessian matrix when the partial derivatives are taken with respect to $\mathbf{z}$. Then, applying Stein's lemma, we see that

$$\bar{\mathbf{H}}_{\mathbf{z}} = \mathbf{E}(y - \mu_y)\mathbf{z}\mathbf{z}'.$$

The relationship between $\mathbf{H}_{\mathbf{x}}(\mathbf{x})$ and $\mathbf{H}_{\mathbf{z}}(\mathbf{z})$ is obtained by the chain rule: $\mathbf{H}_{\mathbf{x}}(\mathbf{x}) = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{H}_{\mathbf{z}}(\mathbf{z})\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}$. On the other hand, $\boldsymbol{\Sigma}_{yxx} = \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}\mathbf{E}(y - \mu_y)\mathbf{z}\mathbf{z}'\boldsymbol{\Sigma}_{\mathbf{x}}^{1/2} = \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}\bar{\mathbf{H}}_{\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}$. Now it is clear that $\bar{\mathbf{H}}_{\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{yxx}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}$, as desired. The proof is complete.

A.4 Proof of Corollary 3.2

Let $\tilde{r}$ be the residual, $y - \tilde{Q}(\mathbf{x})$, where $\tilde{Q}(\mathbf{x})$ is the fitted polynomial. Then $\tilde{r}$ is uncorrelated with any polynomial of $\mathbf{x}$ with degree q or less. In particular, it is uncorrelated with any element in the random matrix $(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})'$. Now we see that

$$\begin{aligned}&\mathbf{E}(y - \mu_y)(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})'\\ &= \mathbf{E}(y - \tilde{r} - \mu_y)(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})'\\ &= \mathbf{E}(\tilde{Q}(\mathbf{x}) - \mathbf{E}\tilde{Q}(\mathbf{x}))(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})'.\end{aligned}$$

Corollary 3.1 implies that the average Hessian matrices for y and $\tilde{Q}(\mathbf{x})$ are the same, completing the proof.

A.5 Proof of Theorem 6.1

Consider any vector $\mathbf{u}$ such that $\mathbf{u}'\boldsymbol{\Sigma}_{\mathbf{x}}b = 0$ for any b in $\mathcal{B}$. Then (6.1) implies that $\mathbf{E}(u'\mathbf{x}|\beta_j'\mathbf{x}, j = 1, \ldots, K) = 0$. It follows that $\mathbf{u}'\boldsymbol{\Sigma}_{yxx}b = \mathbf{E}((y - \mu_y)\mathbf{E}(u'\mathbf{x}|\beta_j'\mathbf{x}, j = 1, \ldots, K)\mathbf{x}'b) = 0$. This completes the proof.

A.6 Proof of Theorem 6.2

Due to affine invariance, it suffices to consider the case where $\mathbf{x}$ is spherically symmetric with identity covariance and mean 0. Let $\mathbf{P}_1$ be the projection matrix of rank K with $\mathcal{B}$ as the range space, and let $\mathbf{P}_2 = I - \mathbf{P}_1$. We need only show that the range of $\mathbf{P}_2$ is a subspace of some eigenspace of $\boldsymbol{\Sigma}_{yxx}$. First, the result of Theorem 6.1 implies that $\mathbf{P}_1\boldsymbol{\Sigma}_{yxx}\mathbf{P}_2 = 0$, or equivalently, that $\boldsymbol{\Sigma}_{yxx}\mathbf{P}_2 = \mathbf{P}_2\boldsymbol{\Sigma}_{yxx}\mathbf{P}_2$. Fundamental properties from elliptical distributions show that, given $\mathbf{P}_1\mathbf{x}$ and $\|\mathbf{x}\|^2$, $\mathbf{P}_2\mathbf{x}$ is still spherically symmetric with mean 0, and the covariance matrix is $(p - K)^{-1}(\|\mathbf{x}\|^2 - \|\mathbf{P}_1\mathbf{x}\|^2)\mathbf{P}_2$. From this we see that

$$\begin{aligned}\mathbf{P}_3\boldsymbol{\Sigma}_{yxx}\mathbf{P}_2 &= \mathbf{E}((y - \mu_y)\mathbf{E}(\mathbf{P}_2\mathbf{x}\mathbf{x}'\mathbf{P}_2|\mathbf{P}_1\mathbf{x}, \|\mathbf{x}\|^2))\\ &= (p - K)^{-1}[\mathbf{E}(y - \mu_y)(\|\mathbf{x}\|^2 - \|\mathbf{P}_1\mathbf{x}\|^2)]\mathbf{P}_2.\end{aligned}$$

Thus $\boldsymbol{\Sigma}_{yxx}\mathbf{P}_2$ is proportional to $\mathbf{P}_2$, implying that the range space of $\mathbf{P}_2$ is contained in an eigenspace of $\boldsymbol{\Sigma}_{yxx}$. This proves the theorem.

APPENDIX B: PROOFS OF THEOREMS 4.1 AND 4.2

We need the following lemma from perturbation theory (see, for example, Kato 1976). It has been used in Li (1991). Let the superscript $^+$ denote the Moore–Penrose generalized inverse of a matrix.

B.1 Lemma B.1

Consider the expansion

$$\mathbf{T}(w) = \mathbf{T} + w\mathbf{T}^{(1)} + o(w),$$

where $\mathbf{T}(w)$, $\mathbf{T}$, and $\mathbf{T}^{(1)}$ are symmetric matrices. Suppose that $\mathbf{T}$ has rank k. Let $\lambda(w)$ be the sum of the $p - k$ eigenvalues of $\mathbf{T}(w)$ with absolute values closest to 0, and let $\boldsymbol{\Pi}(w)$ be the projection matrix associated with corresponding $p - k$ eigenvectors. Let $\boldsymbol{\Pi}$ be the projection matrix associate with the null space of $\mathbf{T}$ so that $\boldsymbol{\Pi}\mathbf{T} = \mathbf{T}\boldsymbol{\Pi} = 0$. Then we have

$$\begin{aligned}\boldsymbol{\Pi}(w) &= \boldsymbol{\Pi} - w\boldsymbol{\Pi}\mathbf{T}^{(1)}\mathbf{T}^{+}\mathbf{T}^{+}\mathbf{T}^{(1)}\boldsymbol{\Pi} + o(w)\\ \lambda(w) &= w\lambda^{(1)} + o(w),\end{aligned}$$

where $\lambda^{(1)} = \operatorname{tr}\mathbf{T}^{(1)}\boldsymbol{\Pi}$. In addition, if we have a second-order expansion, then

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

where $\mathbf{T}^{(2)}$ is also symmetric. Then

$$\lambda(w) = w\lambda^{(1)} + w^2\lambda^{(2)} + o(w^2),$$

where $\lambda^{(2)} = \operatorname{tr}[\mathbf{T}^{(2)}\boldsymbol{\Pi} - \mathbf{T}^{(1)}\mathbf{T}^{+}\mathbf{T}^{(1)}\boldsymbol{\Pi}]$.

B.2 Proof of Theorem 4.2 for the y-Based Estimate

Due to the affine invariance of the problem, we shall assume that

$$\mu_{\mathbf{x}} = 0, \qquad \mu_y = 0, \qquad \operatorname{cov}\mathbf{x} = I,$$

without loss of generality. First, we shall apply Lemma B.1 to find the asymptotic expansion of $\sum_{j=(k+1)}^{p}\hat{\lambda}_{yj}$. To do this, we have to find the expansion for the matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}\hat{\boldsymbol{\Sigma}}_{yxx}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}$ first, because it gives the needed eigenvalues.

It is straightforward to obtain

$$\hat{\boldsymbol{\Sigma}}_{yxx} = n^{-1}\sum_{i=1}^{n} y_i(\mathbf{x}_i\mathbf{x}_i' - I) - \bar{\mathbf{x}}(\mathbf{E}\mathbf{x}'y) - (\mathbf{E}y\mathbf{x})\bar{\mathbf{x}}' + o_p(n^{-1/2}). \tag{B.1}$$

To simplify the expression, define

$$\mathbf{A} = \mathbf{E}y\mathbf{x}\mathbf{x}', \quad b = \mathbf{E}y\mathbf{x}, \quad \psi_i = y_i(\mathbf{x}_i\mathbf{x}_i' - I) - \mathbf{A}, \quad \xi_i = \mathbf{x}_i\mathbf{x}_i' - I.$$

Then (B.1) becomes

$$\hat{\boldsymbol{\Sigma}}_{yxx} = \mathbf{A} + \bar{\psi} - \bar{\mathbf{x}}b' - b\bar{\mathbf{x}}' + o_p(n^{-1/2})$$

and we also have

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{x}} = I + \bar{\xi} + o_p(n^{-1/2}).$$

It follows that

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}\hat{\boldsymbol{\Sigma}}_{yxx}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2} &= \mathbf{A} + \mathbf{B}_n + \mathbf{D}_n\\ \mathbf{B}_n &= \bar{\psi} - \bar{\mathbf{x}}b' - b\bar{\mathbf{x}}' - 1/2\bar{\xi}\mathbf{A} - 1/2\mathbf{A}\bar{\xi},\end{aligned} \tag{B.2}$$

where $\mathbf{D}_n$ is of the order $o_p(n^{-1/2})$.

Now Lemma B.1 can be applied. Let $\mathbf{P}_2$ be the projection matrix associated with the null space of $\mathbf{A}$. Here we take $\mathbf{A}$, $n^{1/2}\mathbf{B}_n$, and $\mathbf{P}_2$ as $\mathbf{T}$, $\mathbf{T}^{(1)}$, and $\boldsymbol{\Pi}$ and note that $\mathbf{A}$ has rank k. Hence

$$\sum_{j=(k+1)}^{p}\hat{\lambda}_{yj} = \operatorname{tr}\mathbf{B}_n\mathbf{P}_2 + o_p(n^{-1/2}). \tag{B.3}$$

Using the fact that $\mathbf{A}\mathbf{P}_2 = 0$ and $\mathbf{P}_2b = 0$, the right side of (B.3) can be simplified as

$$\begin{aligned}&\operatorname{tr} n^{-1}\sum\psi_i\mathbf{P}_2 + o_p(n^{-1/2})\\ &\qquad = n^{-1}\sum y_i\operatorname{tr}((\mathbf{P}_2\mathbf{x}_i)(\mathbf{P}_2\mathbf{x}_i)' - \mathbf{P}_2) + o_p(n^{-1/2}).\end{aligned}$$

Due to the independence between y_i and $\mathbf{P}_2\mathbf{x}_i$, we can verify that

$$\mathbf{E}(y_i\operatorname{tr}((\mathbf{P}_2\mathbf{x}_i)(\mathbf{P}_2\mathbf{x}_i)' - \mathbf{P}_2)) = \mathbf{E}y_i\mathbf{E}\operatorname{tr}((\mathbf{P}_2\mathbf{x}_i)(\mathbf{P}_2\mathbf{x}_i)' - \mathbf{P}_2) = 0$$

and

$$\begin{aligned}&\operatorname{var}(y_i\operatorname{tr}((\mathbf{P}_2\mathbf{x}_i)(\mathbf{P}_2\mathbf{x}_i)' - \mathbf{P}_2))\\ &\qquad = \mathbf{E}y^2\mathbf{E}\operatorname{tr}((\mathbf{P}_2\mathbf{x}_i)(\mathbf{P}_2\mathbf{x}_i)' - \mathbf{P}_2) = 2(p - k)\mathbf{E}y^2.\end{aligned}$$

Now, applying the central limit theorem, we complete the proof of the first part of Theorem 4.2.

We proceed with the proof of the second part of Theorem 4.2 by observing that $\sum_{j=k+1}^{p}\hat{\lambda}_{yj}^2$ equals the sum of the smallest $p - k$ eigenvalues of the matrix

$$(\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}\hat{\boldsymbol{\Sigma}}_{yxx}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2})(\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}\hat{\boldsymbol{\Sigma}}_{yxx}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2})',$$

which can be expanded as

$$\mathbf{A}\mathbf{A} + (\mathbf{B}_n\mathbf{A} + \mathbf{A}\mathbf{B}_n) + (\mathbf{B}_n\mathbf{B}_n + \mathbf{D}_n\mathbf{A} + \mathbf{A}\mathbf{D}_n) + o_p(n^{-1}).$$

Now, by Lemma B.1, we have

$$\begin{aligned}\sum_{j=k+1}^{p} \hat{\lambda}_{yj}^2 &= \operatorname{tr}(\mathbf{B}_n\mathbf{A} + \mathbf{A}\mathbf{B}_n)\mathbf{P}_2 + [\operatorname{tr}(\mathbf{B}_n\mathbf{B}_n + \mathbf{D}_n\mathbf{A} + \mathbf{A}\mathbf{D}_n)\mathbf{P}_2 \\ &\quad - \operatorname{tr}(\mathbf{B}_n\mathbf{A} + \mathbf{A}\mathbf{B}_n)(\mathbf{A}\mathbf{A})^+(\mathbf{B}_n\mathbf{A} + \mathbf{A}\mathbf{B}_n)\mathbf{P}_2] + o_p(n^{-1}) \\ &= [\operatorname{tr}\mathbf{P}_2\mathbf{B}_n\mathbf{B}_n\mathbf{P}_2 - \operatorname{tr}\mathbf{P}_2\mathbf{B}_n\mathbf{A}(\mathbf{A}\mathbf{A})^+\mathbf{A}\mathbf{B}_n\mathbf{P}_2] + o_p(n^{-1}) \\ &= \operatorname{tr}\mathbf{P}_2\mathbf{B}_n\mathbf{P}_2\mathbf{B}_n\mathbf{P}_2 + o_p(n^{-1}).\end{aligned}$$

Now

$$\begin{aligned}\mathbf{P}_2\mathbf{B}_n\mathbf{P}_2 &= n^{-1}\sum_{i=1}^{n} y_i \operatorname{tr}\mathbf{P}_2(\mathbf{x}_i\mathbf{x}_i' - I)\mathbf{P}_2 + o_p(n^{-1}) \\ &\equiv n^{-1/2}C_n + o_p(n^{-1}).\end{aligned}$$

Due to the independence between y_i and $\mathbf{P}_2\mathbf{x}_i$, C_n is asymptotically normal with mean 0. A straightforward computation shows that the asymptotic variance for each diagonal element is $2\operatorname{var}(y)$ and the asymptotic variance for each off-diagonal element is $\operatorname{var}(y)$. C_n is symmetric, but all distinct elements have asymptotic covariances equal to 0. Finally, the trace of the matrix, $n \operatorname{tr}\mathbf{P}_2\mathbf{B}_2\mathbf{P}_2\mathbf{B}_2\mathbf{P}_2$, equals the sum of the squares of the elements in C_n, which in turn follows a rescaled χ^2 distribution with $(p-k+1)(P-k)/2$ degrees of freedom asymptotically. The proof of Theorem 4.2 is complete.

B.3 Proof of Theorem 4.1 for the *y*-Based Estimate

We make the same assumptions and adopt the same notations as in the proof of Theorem 4.2. The strategy of our proof is similar to the argument given in Li (1991, app. A.2). First, the squared trace correlation $R^2(\hat{\mathcal{B}})$ reduces to

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

where $\mathbf{P}_1$ and $\hat{\mathbf{P}}_1$ are symmetric projection matrices associated with the edr space $\mathcal{B}$ and the estimated space $\hat{\mathcal{B}}$. In our case, $\mathcal{B}$ is spanned by the first K coordinate axes. Let $\tilde{\mathbf{P}}_1$ be the projection matrix associated with the first K eigenvectors for the eigenvalue decomposition of $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}\hat{\mathbf{\Sigma}}_{yxx}\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}$.

$\hat{\mathbf{P}}_1$ is related to $\tilde{\mathbf{P}}_1$ via

$$\hat{\mathbf{P}}_1 = \hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}(\hat{\mathbf{P}}_1\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\tilde{\mathbf{P}}_1) + \hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}. \tag{B.4}$$

Furthermore, we need the following approximation, which follows from Lemma B.1:

$$I - \tilde{\mathbf{P}}_1 = \mathbf{P}_2 - (\mathbf{P}_2\mathbf{B}_n\mathbf{A}^+ + \mathbf{A}^+\mathbf{B}_n\mathbf{P}_2) + o_p(n^{-1/2}). \tag{B.5}$$

Then, from (B.4) and (B.5), we can derive

$$\begin{aligned}\hat{\mathbf{P}}_1 &= (I - 1/2\bar{\xi})[\tilde{\mathbf{P}}_1(I - \bar{\xi})\tilde{\mathbf{P}}_1]^+(I - 1/2\bar{\xi}) + o_p(n^{-1/2}) \\ &= (I - 1/2\bar{\xi})\tilde{\mathbf{P}}_1(I + \bar{\xi})\tilde{\mathbf{P}}_1(I - 1/2\bar{\xi}) + o_p(n^{-1/2}) \\ &= \mathbf{P}_1 + (\mathbf{P}_2\mathbf{B}_n\mathbf{A}^+ + \mathbf{A}^+\mathbf{B}_n\mathbf{P}_2) \\ &\quad + (\mathbf{P}_1\bar{\xi}P_1 - 1/2\bar{\xi}\mathbf{P}_1 - 1/2\mathbf{P}_1\bar{\xi}) + o_p(n^{-1/2}).\end{aligned}$$

Hence, using $\mathbf{A}^+\mathbf{P}_1 = \mathbf{A}^+$ and (B.2), we have

$$\begin{aligned}(\hat{\mathbf{P}}_1 - \mathbf{P}_1)\mathbf{P}_1 &= \mathbf{P}_2\mathbf{B}_n\mathbf{A}^+ - 1/2\mathbf{P}_2\bar{\xi}\mathbf{P}_1 + o_p(n^{-1/2}) \\ &= \mathbf{P}_2(\bar{\psi}\mathbf{A}^+ + \bar{\mathbf{x}}b'\mathbf{A}^+) - \mathbf{P}_2\bar{\xi}\mathbf{P}_1 + o_p(n^{-1/2}) \\ &= n^{-1}\sum_{i=1}^{n}(y_i\mathbf{P}_2\mathbf{x}_i\mathbf{x}_i'\mathbf{A}^+ - \mathbf{P}_2\mathbf{x}_ib'\mathbf{A}^+ - \mathbf{P}_2\xi_i\mathbf{P}_1) \\ &\quad + o_p(n^{-1/2}) \\ &\equiv n^{-1}\sum_{i=1}^{n}\zeta_i + o_p(n^{-1/2}).\end{aligned}$$

It remains to calculate

$$\mathbf{E}R^2(\hat{\mathcal{B}}) = 1 - K^{-1}n^{-1}\mathbf{E}\operatorname{tr}\zeta_i\zeta_i' + o_p(n^{-1}).$$

Now it is straightforward to see that $\zeta_i = \mathbf{P}_2\mathbf{x}_i((y_i\mathbf{x}_i - b)'\mathbf{A}^+ - \mathbf{x}_i'\mathbf{P}_1)$. By independence between $\mathbf{P}_2\mathbf{x}_i$ and $\mathbf{P}_1\mathbf{x}_i$, y, and the identity $\mathbf{A}^+ = \sum_{j=1}^{K} b_jb_j'$, we have

$$\begin{aligned}\mathbf{E}\operatorname{tr}\zeta_i\zeta_i' &= \mathbf{E}\|\mathbf{P}_2\mathbf{x}_i\|^2\mathbf{E}\|(y_i\mathbf{x}_i - b)'\mathbf{A}^+ - \mathbf{x}_i\mathbf{P}_1\|^2 \\ &= (p-K)\mathbf{E}\left\|\sum_{j=1}^{K}\lambda_j^{-1}(y_i\mathbf{x}_i - b - \lambda_j\mathbf{x}_i)'b_jb_j'\right\|^2 \\ &= (p-K)\sum_{j=1}^{K}\lambda_j^{-2}\mathbf{E}((y_i\mathbf{x}_i - b)'b_j - \lambda_j\mathbf{x}_i'b_j)^2 \\ &= (p-K)\sum_{j=1}^{K}\lambda_j^{-2}(\operatorname{var}(y_i\mathbf{x}_i'b_j) + \lambda_j^2\operatorname{var}\mathbf{x}_i'b_j \\ &\qquad - 2\lambda_jb_j'\mathbf{E}y_i\mathbf{x}_i\mathbf{x}_i'b_j) \\ &= (p-K)\sum_{j=1}^{K}[\lambda_j^{-2}\operatorname{var}(y_i\mathbf{x}_i'b_j) - 1],\end{aligned}$$

where the last identity uses the fact that b_j is the jth eigenvector for $\mathbf{E}y_i\mathbf{x}_i\mathbf{x}_i'$. The proof is complete.

B.4 Proof of Theorems 4.2 and 4.1 for the *r*-Based Estimate

This is similar to the proof of Theorems 4.2 and 4.1 for the y-based estimate. First, we make the same assumptions as were made to standardize $\mathbf{x}$ and y. Let $r_i = y_i - b_{ls}\mathbf{x}_i$, where $b_{ls} = \mathbf{E}y\mathbf{x} = b$. Now a straightforward expansion for the least squares leads to

$$\hat{r}_i = r_i - \bar{r} - \left(n^{-1}\sum_{j=1}^{n} r_j\mathbf{x}_j'\right)\mathbf{x}_i + o_p(n^{-1/2}),$$

where $\bar{r} = n^{-1}\sum_{i=1}^{n} r_i$. It follows that

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{rxx} &= n^{-1}\sum_{i=1}^{n}(r_i - \bar{r})(\mathbf{x}_i - \hat{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \\ &\quad - n^{-1}\sum_{i=1}^{n}\left[n^{-1}\sum_{j=1}^{n} r_j\mathbf{x}_j'\right]\mathbf{x}_i\xi_i + o_p(n^{-1/2}) \\ &= \mathbf{A} + \bar{\phi} - n^{-1}\sum_{i=1}^{n}\left[n^{-1}\sum_{j=1}^{n} r_i\mathbf{x}_j'\right]\mathbf{x}_i\xi_i + o_p(n^{-1/2}),\end{aligned} \tag{B.6}$$

where $\bar{\phi} = n^{-1}\sum_{i=1}^{n}\phi_i$ and $\phi_i = r_i(\mathbf{x}_i\mathbf{x}_i' - I) - \mathbf{A}$. A straightforward evaluation of the variance of the third term in (B.6) shows that it is negligible. Therefore, we have

$$\hat{\mathbf{\Sigma}}_{rxx} = \mathbf{A} + \bar{\phi} + o_p(n^{-1/2}).$$

The rest of the proof is omitted, because it follows the same manipulation starting from (B.2) by ignoring b and replacing $\bar{\psi}$ with $\bar{\phi}$.

[*Received March 1990. Revised August 1991.*]

REFERENCES

Box, G. (1954), "The Exploration and Exploitation of Response Surfaces: Some General Considerations and Examples," *Biometrics*, 10, 16–60.

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

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.

——— (1983), "A Generalized Linear Model With 'Gaussian' Regressor Variables," in *A Festschrift for Erick L. Lehmann*, eds. P. J. Bickel, K. A. Doksum, and J. L. Hodges, New York: Wadsworth, pp. 97–114.

——— (1991), *Journal of the American Statistical Association,* Comment on "Sliced Inverse Regression for Dimension Reduction," by K. C. Li, 86, 333.

Chen, H. (1991), "Estimation of a Projection-Pursuit Type Regression Model," *The Annals of Statistics,* 19, 142–157.

Cleveland, W. S. (1988), *The Collected Works of John W. Tukey, Volume V, Graphics: 1965–1985.* Pacific Grove, CA: Wadsworth & Brooks/Cole.

Cleveland, W. S., and MacGill, M. E. (1988), *Dynamic Graphics for Statistics,* Pacific Grove, CA: Wadsworth & Brooks/Cole.

Cook, R. D., and Weisberg, S. (1989), "Regression Diagnostics With Dynamic Graphics," (with discussions), *Technometrics,* 31, 277–308.

——— (1991), Comment on "Sliced Inverse Regression for Dimension Reduction" by K. C. Li, *Journal of the American Statistical Association,* 86, 328–332.

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

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

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

Duan, N., and Li, K. C. (1991), "Slicing Regression: A Link-Free Regression Method," *The Annals of Statistics,* 19, 505–530.

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.

Gifi, A. (1990), *Nonlinear multivariate analysis,* Chichester, U.K.: John Wiley.

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

——— (1989b), "On Polynomial-Based Projection Indices for Exploratory Projection Pursuit," *The Annals of Statistics,* 17, 589–605.

Hall, P., and Li, K. C. (in press), "On Almost Linearity of Low Dimensional Projection from High Dimensional Data," *The Annals of Statistics,*

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

Härdle, W., and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association,* 84, 986–995.

Hsing, T., and Carroll, R. J., (1992),"Asymptotic Properties of Sliced Inverse Regression," *The Annals of Statistics,* 20, 1040–1062.

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

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

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. (1985), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross-Validation," *The Annals of Statistics,* 13, 1352–1377.

——— (1986), "Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression With Application to Spline Smoothing," *The Annals of Statistics,* 14, 1101–1112.

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

——— (1989a), "Honest Confidence Regions for Nonparametric Regression," *The Annals of Statistics,* 17, 1001–1008.

——— (1989b), "Data Visualization with SIR: A Transformation-Based Projection Pursuit Method," UCLA Statistical Series #24, University of California, Los Angeles.

——— (1991), "Sliced Inverse Regression for Dimensional Reduction" (with discussion), *Journal of the American Statistical Association,* 86, 316–342.

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

Li, K. C., and Huang, J. T. (1984), "The Data-Smoothing Aspect of Stein Estimates," *The Annals of Statistics,* 12, 887–897.

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

Parzen, E. (1961), "Regression Analysis of Continuous Parameter Time Series," in *Proceedings of the Fourth Berkeley Symposium,* University of California, Berkeley, pp. 649–689.

Rice, J. (1986), "Convergence Rate for Partially Splined Models," *Statistics & Probability Letters,* 4, 203–208.

Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics,* 9, 1135–1151.

——— (1986), *Approximate Computation of Expectations,* Lecture Notes Monograph Series, Vol. 7, Hayward, CA: Institute of Mathematical Statistics.

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

Tierney, L. (1990), *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics,* 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 18th Symposium on the Interface,* Washington, D.C. pp. 75–80.

Wegman, E. J., and Depriest, D. J. (1986), *Statistical Image Processing and Graphics,* New York: Marcel Dekker.