UCLA
Department of Statistics Papers

Title
Helices in High Dimensional Regression

Permalink
https://escholarship.org/uc/item/6373b8g6

Author
Ker-Chau Li

Publication Date
2011-10-24

eScholarship.org Powered by the California Digital Library
University of California

**Nonlinear Confounding in High-Dimensional Regression**

Ker-Chau Li

*Annals of Statistics*, Volume 25, Issue 2 (Apr., 1997), 577-612.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199704%2925%3A2%3C577%3ANCIHR%3E2.0.CO%3B2-O

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

*Annals of Statistics* is published by Institute of Mathematical Statistics. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ims.html.

*Annals of Statistics*
©1997 Institute of Mathematical Statistics

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2003 JSTOR

http://www.jstor.org/
Wed Apr 2 10:06:11 2003

*The Annals of Statistics*
1997, Vol. 25, No. 2, 577–612

# NONLINEAR CONFOUNDING IN HIGH-DIMENSIONAL REGRESSION[1]

BY KER-CHAU LI

*University of California, Los Angeles*

It is not uncommon to find nonlinear patterns in the scatterplots of regressor variables. But how such findings affect standard regression analysis remains largely unexplored. This article offers a theory on *nonlinear confounding*, a term for describing the situation where a certain nonlinear relationship in regressors leads to difficulties in modeling and related analysis of the data. The theory begins with a measure of nonlinearity between two regressor variables. It is then used to assess nonlinearity between any two projections from the high-dimensional regressor and a method of finding most nonlinear projections is given. Nonlinear confounding is addressed by taking a fresh new look at fundamental issues such as the validity of prediction and inference, diagnostics, regression surface approximation, model uncertainty and Fisher information loss.

**1. Introduction.** This article offers a way of studying general regression models of the form:

$$y = f(\beta'\mathbf{x}, \varepsilon). \tag{1.1}$$

In contrast to most earlier works (see Section 1.4 for a brief review), our focus is not on how to find a good estimate of the unknown $p$-vector $\beta$. Rather, we shall address several statistical issues about *nonlinear confounding*, a term used for describing a situation where a certain nonlinear relationship in the $p$-dimensional regressor $\mathbf{x}$ leads to difficulties in specifying the functional form of $f$ and in related analysis of the data. Nonlinear confounding sets intrinsic limitations on how much can be learned about (1.1). Such limitations point to a new type of danger inherent in prediction and inference, which is not explored yet in the vast literature of regression diagnostics. We offer a new theory to quantify such limitations. In developing the theory, we have in mind that the main application should be on the high-dimensional cases where data analysts often look to graphical methods for enhancing their analysis.

What ignites our study is a special data pattern between a dependent variable and the regressors, first reported in the Rejoinder of Li (1991) and subsequently in several other data sets. It appears like a part of a helix or a twisted slide; for easy reference, we shall call it a quasi-helix. Figure 1 is an

Received March 1994; revised August 1996.
[1]Research supported in part by NSF Grants.
*AMS* 1991 *subject classifications.* 62J20, 62J99.
*Key words and phrases.* Adaptiveness, dimension reduction, graphics, nonlinear regression, overlinearization, quasi-helical confounding, information matrices, regression diagnostics, semiparametrics, sliced inverse regression.

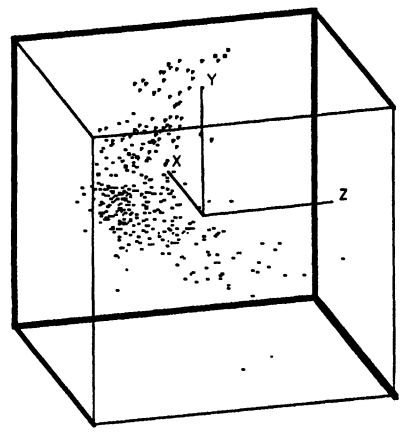


FIG. 1. *A quasi-helical pattern found in Boston Housing Data.*

exhibition of a similar structure, obtained by the methodology developed in this article. Details are to be described in Example 3.1 of Section 3.

1.1. *Nonlinear confounding in bivariate regression.* To help bring the issue into focus, we begin with a bivariate example.

EXAMPLE 1.1a. Consider a simple linear model with two regressors $x_1$, $x_2$:

$$y = x_2 + \varepsilon \tag{1.2}$$

where $\varepsilon$ is a normal random variable with standard deviation $\sigma = 0.1$. The regressors are generated from

$$\begin{aligned} x_1 &\sim \text{uniform}[0, 1] \\ x_2 &= \log x_1 + e, \qquad e \sim \text{uniform}[-\delta^*, \delta^*] \end{aligned} \tag{1.3}$$

with $\delta^* = 0.3$. Figure 2(a) shows a strong nonlinear trend between the regressors.

EXAMPLE 1.1b. Consider the same regressors as generated in Example 1.1a. Instead of (1.2), generate $y$ from the model

$$y = \log x_1 + \varepsilon \tag{1.4}$$

Figures 2(b)–(e) show several angles of the three-dimensional scatterplot.

To save space, we do not display the three-dimensional plot for Example 1.1a. It has a shape which is hard to distinguish from what is seen in Example 1.1b. In view of the exchanged role of $x_1$ and $x_2$, this is somewhat anticipated. In either case, data points are found to cling to a curve which draws a descending arc, bent and twisted around the $y$-axis. When we spin the three-dimensional plot about the $y$-axis fast enough, the structure appears like a portion of a helical-spiral-slide-shaped object winding around on the screen.

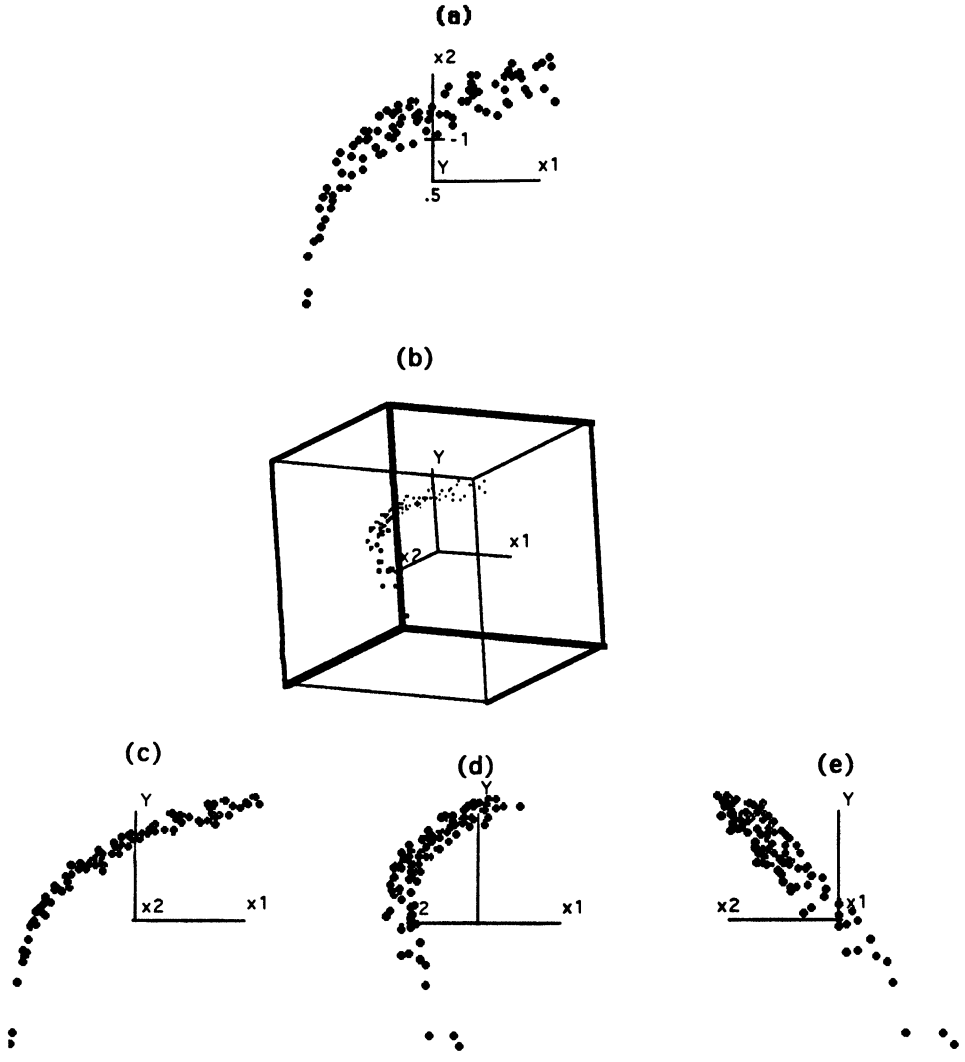


FIG. 2. *Three-dimensional plot for Example* 1.1*b*. (*a*) *shows the nonlinear trend in regressor.* (*b*)–(*e*) *are some views of the three-dimensional scatterplot of y against the two regressors. A quasi-helical pattern is revealed by rotation.*

What should data analysts do in face of such data patterns? This is indeed a hard question to answer. It is not easy to use graphical methods to tell which model the data may come from. Take the data from Example 1.1b. Both (1.2) and (1.4) appear to fit the data quite well. The true model (1.4), can be written as $y = x_2 + (\varepsilon - e)$. Strictly speaking, in view of (1.3), $e$ is not independent of $x_2$. However, this subtlety is hardly noticeable from Figure 2(e). Furthermore, (1.2) and (1.4) are just two special cases of (1.1). Other

equally compelling suggestions can be found by rotating the three-dimensional plot to different angles.

Such ambiguity in modeling–namely, different projections suggesting different models—occurs because the input–output relationship between $y$ and $\mathbf{x}$ is confounded with the nonlinear relationship between the regressor variables. This is the situation we earlier called *nonlinear confounding*.

We regard nonlinear confounding as an intrinsic weakness in the data. We urge that this weakness be fully reported in the final analysis. Failing to point out the presence of nonlinear confounding may leave certain questionable claims unchallenged. One such danger is in prediction. Consider the prediction of $y$ at the point $(x_{10}, x_{20})' = (0.5, -1)'$, for example. Suppose, due to simplicity or other reasons, only (1.2) is accepted as the tentative model. The predicted value will be $Ey = x_{20} = -1$, which differs substantially from the value predicted by using (1.4), $Ey = \log .5 = -0.69$. Thus our prediction would be grossly wrong if the data are indeed generated from (1.4). Unfortunately this danger is not easy to notice from the plot suggesting the model [Figure 2(e)]. The linear fit seems good. Moreover, the prediction point appears to be within the range of the fitting; $x_{20}$, is seen to fall well inside the predictor's sample space (in fact, $x_{20} = Ex_2$).

The ill effect of nonlinear confounding is of course easy to observe in bivariate regression. The aforementioned danger in prediction is exposed after inspecting the scatterplot of $x_1, x_2$. The prediction point $(0.5, -1)'$ is seen to fall on the boundary of the sample points—it is not a clear-cut interpolation problem as observed earlier.

1.2. *High-dimensional regression.* The above illustration on nonlinear confounding can be extended to the general regression model (1.1) with more than two regressor variables. The role of $x_1$ and $x_2$ in (1.3) may be played by two projected variables from $\mathbf{x}$, $\mathbf{b}'\mathbf{x}$ and $\beta'\mathbf{x}$; the logarithmic function may also be replaced by other nonlinear functions. Nonlinear confounding gets severe as the nonlinear relationship becomes stronger [as $\delta^*$ gets smaller in (1.3), for instance].

An extreme case occurs when there is a perfect nonlinear relationship between $\mathbf{b}'\mathbf{x}$ and $\beta'\mathbf{x}$: $\beta'\mathbf{x} = h(\mathbf{b}'\mathbf{x})$ for some nonlinear function $h(\cdot)$. Clearly, (1.1) can also be expressed as a function of $\mathbf{b}'\mathbf{x}$: $y = f(h(\mathbf{b}'\mathbf{x}), \varepsilon)$. We cannot tell whether $\beta$ or $\mathbf{b}$ should be the true direction to estimate. Factor $\beta'\mathbf{x}$ and factor $\mathbf{b}'\mathbf{x}$ are thus totally confounded, a situation similar to the confounding/aliasing between interactions (and/or main effects) in the fractional factorial design literature.

Nonlinear confounding is geometrically different from confounding due to colinearity between regressors. If $\mathbf{b}'\mathbf{x}$ and $\beta'\mathbf{x}$ are linearly dependent, $\beta'\mathbf{x} = c_1 + c_2\mathbf{b}'\mathbf{x}$, then except for a location shift and a scale change in the horizontal axis, the scatterplot of $y$ against $\mathbf{b}'\mathbf{x}$ should appear the same as the scatterplot of $y$ against $\beta'\mathbf{x}$. Both plots would suggest the same form of regression functions. Under colinearity, although we cannot decide which

vector to estimate, such ambiguity does not increase the difficulty in specifying the functional form of (1.1). To simplify our discussion, from now on we shall assume

*the covariance matrix* $\Sigma_{\mathbf{x}}$ *of* $\mathbf{x}$ *is nondegenerate.*

We also assume that any conditional expectation operator appearing in this article is well defined.

1.3. *Section outlines.* Dangers of nonlinear confounding get harder to detect as the regressor dimension increases. To find good lower-dimensional projections that may reveal clear patterns, like Example 1.1, for closer examination, we first need some measure of nonlinearity between regressors. Such a nonlinearity measure, called the $\kappa$ measure, is introduced in Section 2. With this measure, we can assess the degree of nonlinearity between any two projections $\mathbf{v}'\mathbf{x}$ and $\mathbf{b}'\mathbf{x}$. This brings out the notion of the most nonlinear projection direction $\mathbf{v}$ against a given direction $\mathbf{b}$: a direction $\mathbf{v}$ such that the nonlinear trend in the scatterplot of the variable $\mathbf{v}'\mathbf{x}$ against the given variable $\mathbf{b}'\mathbf{x}$ is the strongest. A simple method is offered for finding $\mathbf{v}$ once $\mathbf{b}$ is specified.

In Section 3, a new diagnostic plot is introduced for enhancing the analysis of multiple linear regression. This is a three-dimensional scatter diagram constructed by plotting the response variable against the projection of the regressor along two special directions: the direction of the estimated slope vector and the most nonlinear direction against it. This plot serves as a diagnostic checking for possible ill effects of nonlinear confounding. A simulation study and an application to the Boston Housing Data are reported.

The justification of using the $\kappa$ measure is pursued from several perspectives in the subsequent sections.

In Section 4, we open an issue on linearizing the regression surface by multiple linear regression. The traditional view in this area proposes that (1) multiple linear regression still offers a best linear approximation to the unknown regression function even if it is nonlinear; (2) the pattern of departure from linearity can be detected through plots of residuals. While such viewpoints are intuitively appealing, it turns out that without a careful examination on the $\kappa$ measure, they can also be quite misleading. We argue that due to the minimization nature of the least squares procedure, the regression function as perceived from the least squares direction typically appears more linear than it truly is. Such discrepancy can be described by an over-linearization term. An inequality is then established, which sharply bounds the over-linearization term by the $\kappa$ measure. If the distribution of $\mathbf{x}$ is elliptically symmetric (a case where the $\kappa$ measure is zero), then the over-linearization term vanishes. This is a case where plots of residuals can be as effective as one usually anticipates. On the other hand, if the $\kappa$ measure turns out large, then severe over-linearization may occur and residual plots could fail to detect nonlinearity.

The above issue of overfitting also arises in nonlinear approximation. This is discussed in Section 5.

Section 6 studies another aspect of model uncertainty regarding the loss of information in estimating the slope vector. A semiparametric approach is taken here. We allow the linear model to be embedded in larger parametric models, using additional nuisance parameters to reflect small departures from linearity. The incorporation of the additional nuisance parameters generally reduces the Fisher information for the slope vector. A least favorable model, an augmented parametric model yielding the most serious information loss, is constructed. Again, nonlinearity in $\mathbf{x}$ turns out to be crucial. A major component of the information loss is attributable to the $\kappa$ measure.

Section 7 discusses the validity of commonly used F-tests on the direction of $\beta$. We find that nonlinear confounding blows up the nominal $\alpha$-level and weakens the power of the test substantially. We offer an alternative to the F-test, based on the least favorable model found in Section 6. The level of our test is better protected against nonlinear confounding.

Section 8 summarizes the findings of our study. Proofs and some technical details are given in the Appendix.

1.4. *Background for* (1.1). Regression analysis is often guided by (1.1). Under various specifications on the structure of $f$ and the distribution of $\varepsilon$, it leads to linear regression, the well-known Box–Cox transformation [Box and Cox (1964)], the generalized linear model with a canonical link [Nelder and Wedderburn (1972)], and many others.

It may appear intractable to study (1.1) without specifying the functional form of $f$ in the first place. But a surprisingly simple solution is first given by Brillinger (1983) who shows that up to a constant of proportionality, the least squares estimate $\hat{\beta}_{ls}$ from fitting the linear model,

$$y = \alpha + \beta'\mathbf{x} + \varepsilon, \tag{1.5}$$

is still consistent for $\beta$ in (1.1) even if the data are generated with a nonlinear function $f$. A key condition to Brillinger's result is

$$E(\mathbf{x}|\beta'\mathbf{x} = t) \text{ is linear in } t. \tag{1.6}$$

The linearity condition is satisfied by Gaussian regressors, or more generally, those with elliptically contoured distributions.

Li and Duan (1989) extend this result to general regression estimates. Related earlier works can be found from the references therein. All such consistency results require the key condition (1.6).

If (1.6) were to be required for all nonzero $\beta$ in $R^p$, then the impact of these consistency results would be limited because only elliptically contoured distributions can satisfy the condition [Cook and Weisburg (1991)]. But the fact is that (1.6) is not required for any noneffective dimension reduction direction; for the entire $R^p$, only the direction $\beta$ associated with (1.1) is critical in (1.6). Thus this crucial linearity condition can be satisfied by

nonelliptically contoured distributions [Li (1991), Rejoinder]. Nevertheless, the condition remains hard to verify directly, because the vector $\beta$ is still to be estimated.

A different idea of looking at (1.6) is suggested in Li [(1991), Rejoinder] and further enforced by Hall and Li (1993). Hall and Li assess the size of the set of all $\beta$ directions in $R^p$ for which (1.6) approximately holds. Under certain regularity conditions, they prove that this set indeed includes nearly every vector in $R^p$ when the dimension of $\mathbf{x}$ gets large. Thus for high-dimensional regression problems, if we adopt a Bayesian argument and, for example, consider a flat prior distribution for the unknown $\beta$ in (1.1), then there is a very high probability that (1.6) may turn out adequate.

Hall and Li's result supports the idea that (1.6) can be a reasonable assumption to make, unless evidence pointing to the opposite is found in the data. This leads to a complementary suggestion which forms the essence of our article. In terms of the $\kappa$ measure of Section 2, (1.6) holds if the $\kappa$ measure of nonlinearity along the direction $\beta$ equals 0. We recommend that, whenever a tentative candidate of $\beta$, say $\mathbf{b}$, is available, we make a diagnostic check of (1.6) by applying the technique of Section 2 in this article to find the $\kappa$ measure of nonlinearity along the direction $\mathbf{b}$. If $\kappa$ is sizable, then we should not take (1.6) for granted.

In introducing sliced inverse regression (SIR), the following extension of (1.1) is considered by Li (1991):

$$y = f(\beta_1'\mathbf{x}, \ldots, \beta_d'\mathbf{x}, \varepsilon). \tag{1.7}$$

This model states that the dependence of $y$ on $\mathbf{x}$ is through a lower- (if $d < p$) dimensional projection of $\mathbf{x}$, $(\beta_1'\mathbf{x}, \ldots, \beta_d'\mathbf{x})$. The space spanned by the $\beta$ vectors is called the *effective dimension reduction* (e.d.r.) space. By finding an e.d.r. space, we can reduce a higher- $(p)$ dimensional problem to a lower- $(d)$ dimensional problem without any loss of information for regression analysis. This notion is further elaborated by Cook (1994), who brings out the definition of *minimum dimension reduction space*: it is an e.d.r. space with the smallest dimension. Cook's definition is useful for discussing nonlinear confounding. Let us return to Examples 1.1a and 1.1b in Section 1.2 and consider the extreme case, $\delta^* = 0$ in (1.3). The minimum dimension reduction space is not unique here. It can be the one-dimensional space generated either by $(1, 0)'$ or by $(0, 1)'$, for instance. In general, if a totally confounded situation as described in Section 1.3 occurs, then uniqueness does not hold.

Cook and Weisberg (1994) offer an excellent account on dimension reduction methods, accessible at the undergraduate level. SIR and other related dimension reduction tools [Li (1992), Cook and Weisberg (1991)] all need a condition similar to (1.6):

$$E(\mathbf{x}|\beta_1'\mathbf{x} = t_1, \ldots, \beta_d'\mathbf{x} = t_d) \text{ is linear in } t_1, \ldots, t_d. \tag{1.8}$$

Estimation of $\beta$ in (1.1) is still possible without (1.6), and it can be attempted in several ways. For example, with resampling and reweighting

techniques, we can induce elliptical symmetry on $\mathbf{x}$ before applying multiple linear regression, SIR or related methods; see Brillinger (1991), Li and Duan (1989) and Cook and Natchheim (1994). But how well such methods perform in case of severe nonlinear confounding remains open for investigation.

Projection pursuit regression offers another approach for estimating $\beta$ [Friedman and Stuetzle (1981), Hall (1989), Chen (1991) and Härdle, Hall, and Ichimura (1993)]. Again there is not much evidence suggesting that this approach should work well for data patterns like those from Examples 1.1a and 1.1b, especially in the high-dimensional situation. The success of this approach seems to hinge on rather delicate curve smoothing techniques. Because of the visual difficulty in differentiating competing directions [Figure 1(b)–(e)], the choice of the user-specified smoothing parameter could be critical. Other methods of estimating $\beta$, which require smoothing, are studied in Härdle and Stoker (1989) and Samarov (1993).

**2. A measure of nonlinearity in $x$.** We begin with a bivariate regressor $\mathbf{x} = (x_1, x_2)'$ and consider the following decomposition:

$$x_2 = l(x_1) + r, \tag{2.1}$$

where $l(x_1) = a + bx_1$ is the population version of the least squares solution to linear regression of $x_2$ on $x_1$:

$$(a, b) = \arg\min_{a', b'} E(x_2 - a' - b'x_1)^2.$$

The residual $r$ is further decomposed into a deterministic trend and a random component:

$$r = k(x_1) + e, \tag{2.2}$$

where $k(x_1) = E(r|x_1)$ and $e = r - E(r|x_1)$. Since the linear trend is already removed from $r$, $k(x_1)$ represents the strictly nonlinear component in $E(x_2|x_1)$: $E(x_2|x_1) = l(x_1) + k(x_1)$.

DEFINITION 2.1. For any two noncollinear random variables $x_1, x_2$ admitting the decomposition (2.1) and (2.2), define the $\kappa$ measure of nonlinearity for $x_2$ against $x_1$ to be

$$\kappa_{x_2|x_1} = \frac{\operatorname{var}(k(x_1))}{\operatorname{var}(r)}.$$

Note that the above definition does not apply to collinear random variables. The $\kappa$ measure takes a value between 0 and 1. It represents the proportion of variation in the residual that can be recovered by nonlinear regression. If $\kappa_{x_2|x_1} = 1$, then $x_2 = l(x_1) + k(x)$ and the scatterplot of $x_2$ against $x_1$ will display a noise-free nonlinear curve. On the other hand, if $\kappa_{x_2|x_1} = 0$, then $E(x_2|x_1)$ is strictly linear; no nonlinear trend should be present in the scatterplot of $x_2$ against $x_1$. This nonlinearity measure is not symmetric between $x_1$ and $x_2$; see Remark 2.1 at the end of this section for further discussion.

A simple corollary following from Definition 2.1 is that for any constants $a_0, a_1, a_2 (\neq 0)$,

$$\kappa_{a_2 x_2 + a_1 x_1 + a_0 | x_1} = \kappa_{x_2 | x_1}. \tag{2.3}$$

Now we generalize this nonlinearity measure to multivariate regressors.

DEFINITION 2.2. For any $p$-dimensional random variable $\mathbf{x}$, the $\kappa$ measure of nonlinearity along the direction $\mathbf{b}$ is defined as

$$\kappa_{\mathbf{b}} = \max_{\mathbf{v} \in R^p} \kappa_{\mathbf{v}'\mathbf{x} | \mathbf{b}'\mathbf{x}}$$

Any direction $\mathbf{v}$ that achieves $\kappa_{\mathbf{b}}$ can be considered as a most nonlinear direction against $\mathbf{b}$. Because of (2.3), any linear combination of a most nonlinear direction $\mathbf{v}$ and $\mathbf{b}$ is also a most nonlinear direction for projection. But the degree of linear association varies. For the clearest exhibition of nonlinearity in the plot of $\mathbf{v}'\mathbf{x}$ against $\mathbf{b}'x$, we may choose a most nonlinear direction $\mathbf{v}$ from those that are orthogonal to the direction $\mathbf{b}$ with respect to the covariance $\Sigma_{\mathbf{x}}$ of $\mathbf{x}$:

$$\kappa_{\mathbf{b}} = \max_{\rho(\mathbf{v}'\mathbf{x}, \mathbf{b}'\mathbf{x}) = 0} \kappa_{\mathbf{v}'\mathbf{x} | \mathbf{b}'\mathbf{x}}, \tag{2.4}$$

where $\rho$ denotes the correlation coefficient.

There is an easy way to find $\kappa_{\mathbf{b}}$. Let $\mathbf{r}$ be the residual for the linear regression of $\mathbf{x}$ on $\mathbf{b}'\mathbf{x}$:

$$\mathbf{r} = \mathbf{x} - L_{\mathbf{b}}(\mathbf{b}'\mathbf{x}) \tag{2.5}$$

$$L_{\mathbf{b}}(\mathbf{b}'\mathbf{x}) = E\mathbf{x} + (\operatorname{var}(\mathbf{b}'\mathbf{x}))^{-1}(\Sigma_{\mathbf{x}}\mathbf{b})\mathbf{b}'(\mathbf{x} - E\mathbf{x}). \tag{2.6}$$

Define

$$\mathscr{K}_{\mathbf{b}}(\mathbf{b}'\mathbf{x}) = E(\mathbf{r} | \mathbf{b}'\mathbf{x}). \tag{2.7}$$

Now take the eigenvalue decomposition of the matrix

$$\Sigma_{\mathbf{b}} = \operatorname{cov}(\mathscr{K}_{\mathbf{b}}(\mathbf{b}'\mathbf{x})) \tag{2.8}$$

with respect to $\Sigma_{\mathbf{x}}$:

$$\begin{aligned} &\Sigma_{\mathbf{b}}\gamma_i = \lambda_i \Sigma_{\mathbf{x}}\gamma_i, \qquad i = 1, \ldots, p, \\ &\lambda_1 \geq \cdots \geq \lambda_p. \end{aligned} \tag{2.9}$$

LEMMA 2.1. *The nonlinearity measure $\kappa_{\mathbf{b}}$ is equal to the largest eigenvalue $\lambda_1$ given in* (2.9) *and the first eigenvector $\gamma_1$ is a most nonlinear direction against* $\mathbf{b}$.

To prove this lemma, observe first that the residual for linear regression of $\mathbf{v}'\mathbf{x}$ on $\mathbf{b}'\mathbf{x}$ is equal to $\mathbf{v}'\mathbf{r}$. This shows that

$$\kappa_{\mathbf{v}'\mathbf{x} | \mathbf{b}'\mathbf{x}} = \frac{\operatorname{var}(\mathbf{v}'\mathscr{K}_{\mathbf{b}}(\mathbf{b}'\mathbf{x}))}{\operatorname{var}(\mathbf{v}'\mathbf{r})} = \frac{\mathbf{v}'\Sigma_{\mathbf{b}}\mathbf{v}}{\operatorname{var}(\mathbf{v}'\mathbf{r})}$$

Now the condition $\rho(\mathbf{v}'\mathbf{x}, \mathbf{b}'\mathbf{x}) = 0$ implies $\mathbf{v}'\mathbf{r} = \mathbf{v}'\mathbf{x}$. Therefore using (2.4), we have

$$\kappa_{\mathbf{b}} = \max_{\rho(\mathbf{v}'\mathbf{x}, \mathbf{b}'\mathbf{x})=0} \frac{\mathbf{v}'\Sigma_{\mathbf{b}}\mathbf{v}}{\mathbf{v}'\Sigma_{\mathbf{x}}\mathbf{v}} \leq \max_{\mathbf{v}\in R^p} \frac{\mathbf{v}'\Sigma_{\mathbf{b}}\mathbf{v}}{\mathbf{v}'\Sigma_{\mathbf{x}}\mathbf{v}} \tag{2.10}$$

The maximization problem on the right-hand side of the inequality in (2.10) is solved by the eigenvalue decomposition (2.7). We still have to show that the solution $\mathbf{v} = \gamma_1$ satisfies the orthogonality condition $\rho(\mathbf{b}'\mathbf{x}, \mathbf{v}'\mathbf{x}) = 0$. To continue, observe that $\Sigma_{\mathbf{b}}$ is degenerate in the direction of $\mathbf{b}$ (because $\mathbf{b}'\mathbf{r} = 0$). With this, we can derive the orthogonality condition

$$\rho(\mathbf{b}'\mathbf{x}, \gamma_1'\mathbf{x}) = \mathbf{b}'\Sigma_{\mathbf{x}}\gamma_1 = \lambda_1^{-1}\mathbf{b}'\Sigma_{\mathbf{b}}\gamma_1 = \lambda^{-1}0\gamma_1 = 0.$$

The proof of Lemma 2.1 is now complete. □

Next, we turn to a weighted version of the $\kappa$ measure. Suppose when defining the linear part $l(x_1)$ in (2.1), we take a weighted least squares procedure:

$$\min_{a,b} E\big[(x_2 - a - bx_1)^2 w(x_1)\big],$$

where $w(\cdot)$ is a nonnegative weight function. Use this weighted least squares linear regression to obtain the residual $r$ and then carry out the decomposition (2.2) as before to find the nonlinear component $k(x_1)$. The weighted $\kappa$ measure is defined as

$$\kappa_{x_2|x_1;w} = \frac{E\big(k(x_1)^2 w(x_1)\big)}{E\big(r^2 w(x_1)\big)}.$$

The multivariate version becomes

$$\kappa_{\mathbf{b};w} = \max_{\mathbf{v}\in R^p} \kappa_{\mathbf{v}'\mathbf{x}|\mathbf{b}'\mathbf{x};w}.$$

Weighted linear regression should be applied to obtain (2.5) and (2.6). To carry out the eigenvalue decomposition (2.9), we should replace (2.8) with

$$\Sigma_{\mathbf{b};w} = E\big[w(\mathbf{b}'x)E(\mathbf{r}|\mathbf{b}'\mathbf{x})E(\mathbf{r}|\mathbf{b}'\mathbf{x})'\big]/Ew(\mathbf{b}'\mathbf{x}) \tag{2.11}$$

and replace the covariance $\Sigma_{\mathbf{x}}$ with a weighted version

$$\Sigma_{\mathbf{x};w} = \big(Ew(\mathbf{b}'\mathbf{x})\big)^{-1} \times E\big[(\mathbf{x} - E\mathbf{x}w(\mathbf{b}'\mathbf{x})/Ew(\mathbf{b}'\mathbf{x}))(\mathbf{x} - E\mathbf{x}w(\mathbf{b}'\mathbf{x})/Ew(\mathbf{b}'\mathbf{x}))'w(\mathbf{b}'\mathbf{x})\big]$$

With these modifications, Lemma 2.1 still holds.

REMARK 2.1. Take $\mathbf{x} = (x_1, x_2)'$, $\mathbf{b}_1 = (1, 0)'$, $\mathbf{b}_2 = (0, 1)'$. We can write $\kappa_{x_2|x_1} = \kappa_{\mathbf{b}_1}$ and $\kappa_{x_1|x_2} = \kappa_{\mathbf{b}_2}$. The $\kappa$ measure defined in Definition 2.1 is asymmetric; $\kappa_{x_1|x_2}$ is not the same as $\kappa_{x_2|x_1}$. Consider an extreme case where $x_1 = x_2^2$ and $x_2$ is symmetric about 0. It is easy to see that $x_1$ and $x_2$ are uncorrelated and $\kappa_{x_1|x_2} = 1$. But $\kappa_{x_2|x_1} = 0$ because $E(x_2|x_1) = 0$. In this case,

we have $x_2 = +- \sqrt{x_1}$; the plot of $x_2$ against $x_1$ shows a bifurcation pattern and the nonlinearity comes from the component $e$ in (2.2). Although it is not the primary concern in this article, nonlinearity in $e$ surely deserves attention in further study.

REMARK 2.2. The eigenvalue decomposition (2.9) is similar to the one used in SIR [Li (1991), Duan and Li (1991a)] if we treat $\mathbf{b}'\mathbf{x}$ as $y$ and $\mathbf{r}$ as $\mathbf{x}$.

REMARK 2.3. The linearity condition (1.6) implies $\mathbf{r} = 0$ in (2.5), forcing $\kappa_\beta = 0$.

REMARK 2.4. Suppose $\mathbf{b}$ falls into a $d$-dimensional subspace spanned by vectors $\beta_1, \ldots, \beta_d$ for which (1.8) holds. Then for any vector $c$ such that $\operatorname{cov}(c'\mathbf{x}, \beta_i'\mathbf{x}) = 0$, $i = 1, \ldots, d$, it can be shown that $c'\mathbf{r} = 0$. This implies that the matrix $\Sigma_{\mathbf{b}}$ of (2.8) must degenerate along such $c$ directions (as well as $\mathbf{b}$ itself because $\mathbf{b}'\mathbf{r} = 0$). Therefore, the number of nonzero eigenvalues in the decomposition (2.9) should not be greater than $d - 1$. Although in this article we concentrate only on the first eigenvector, any other eigenvector associated with a nonzero eigenvalue also deserves some attention.

REMARK 2.5. By treating $\mathbf{b}$ as a $p$ by $d$ matrix, we can extend the notion of $\kappa$ measure to the case where the starting projection $\mathbf{b}'\mathbf{x}$ has a larger dimension. Equations (2.5) to (2.9) are well defined with this extension. The eigenvalue decomposition (2.8) can be performed to find the most nonlinear direction against the column space of $\mathbf{b}$.

**3. Exposing quasi-helices.** We can easily implement the procedure suggested by Lemma 2.1 for finding the most nonlinear direction against a given direction $\mathbf{b}$. The residual $\mathbf{r}$ and the covariance matrix $\Sigma_{\mathbf{x}}$ can be replaced by their natural sample estimates, $\hat{\mathbf{r}}$, $\hat{\Sigma}_{\mathbf{x}}$. For $\Sigma_{\mathbf{b}}$, a simple estimate $\hat{\Sigma}_{\mathbf{b}}$ can be formed by slicing.

1. Divide the range of $\mathbf{b}'\mathbf{x}$ into $H$ slices.
2. For each slice $h$, find the sample mean $\bar{\mathbf{r}}_h$ or $\hat{\mathbf{r}}$.
3. Find the covariance of the slice means, weighted by the proportion of cases $\hat{p}_h$ in each slice:

$$\hat{\Sigma}_{\mathbf{b}} = \frac{1}{H} \sum_{h=1}^{H} \hat{p}_h \bar{\mathbf{r}}_h \bar{\mathbf{r}}_h'.$$

Now eigenvalue decomposition (2.9) can be carried out to obtain estimates $\hat{\gamma}_i$, $\hat{\lambda}_i$. If the number of slices is large so that the number of cases in each slice is small, then the variance in estimating the sliced mean becomes sizable and $\hat{\lambda}_i$ tends to over-estimate $\lambda_i$. A simple modification is to take

$$\tilde{\lambda}_i = \max\bigl(\bigl(n\hat{\lambda}_i - H\bigr)/(n - H), 0\bigr).$$

This is motivated from the ANOVA identity used to justify the slice-two method of SIR [Li (1991)].

3.1. *Multiple linear regression and nonlinear confounding.* Empirical model building often begins with multiple linear regression; see, for example, Box and Draper (1987), and Cox and Snell (1981). The linear model (1.5) can be viewed as a first order approximation. After model-fitting, various diagnostic procedures can be applied to detect possible deficiency and to seek ways of remedy; for a recent account on a number of such methods, see Chapters 13 and 14 in Cook and Weisburg (1994). In this connection, we recommend a three-dimensional plot for detecting the presence of nonlinear confounding. This is the scatterplot of $y$ against $\hat{\beta}_{ls}'\mathbf{x}$ and $\hat{\gamma}_1'\mathbf{x}$, where $\hat{\beta}_{ls}$ is the least squares estimate of $\beta$ and $\hat{\gamma}_1$ is the most nonlinear direction against $\hat{\beta}_{ls}$.

To illustrate how it works, we begin with a simulation study which extends the bivariate example in Section 1.1 to a higher-dimensional situation.

EXAMPLE 3.1. First we describe how the data are generated. Let random variables $u_1, u_2$ follow the same joint distributions as the random variables $x_1, x_2$ in (1.3) with $\delta^* = 0.5$. Expand the set of input variables by generating $u_3, u_4, u_5$ from the standard normal distribution. Then take the following linear transformation:

$$x_1 = u_1 + u_3, \quad x_2 = u_2 + u_4 + u_5, \quad x_3 = u_3 - u_4, \quad x_4 = u_4, \quad x_5 = u_5$$

The $y$ variable is related to the regressor $\mathbf{x} = (x_1, \ldots, x_5)'$ via $u_1$:

$$y = \log u_1 + \varepsilon,$$

where, as in (1.4), $\varepsilon$ is normal with mean 0 and standard deviation $\sigma = 0.1$. The sample size is $n = 100$. The transformation in regressor is to hide nonlinearity from being detected by scatterplots of $\mathbf{x}$ coordinate variables. The quasi-helical data pattern like Example 1.1b is hidden in the projections:

$$u_1 = (1, 0, -1, 1, 0)'\mathbf{x}; \qquad u_2 = (0, 1, 0, -1, -1)'\mathbf{x}.$$

Suppose we apply multiple linear regression to this data set as usual. This gives

$$\hat{\beta}_{ls} = (0.667, 0.756, -0.729, -1.49, -0.699)'.$$

A clear linear trend is seen in Figure 3(a) which plots $y$ against $\hat{\beta}_{ls}'\mathbf{x}$ (the variable in the $x$-axis). To find the most nonlinear direction against $\hat{\beta}_{ls}$, we carry out the searching procedure with $H = 15$ slices. The first eigenvalue is $\tilde{\lambda}_1 = 0.578$, and all other eigenvalues are nearly zero. The most nonlinear direction is found to be $\hat{\gamma}_1 = (-1.36, -0.553, 1.41, 1.96, 0.509)'$. Figure 3(e) shows the scatterplot of $\gamma_1'\mathbf{x}$ (the $z$-axis) against $\hat{\beta}_{ls}'\mathbf{x}$ (the $x$-axis). To reveal the quasi-helical structure, rotate the three-dimensional scatterplot of $y$ against $\hat{\beta}_{ls}'\mathbf{x}$ and $\hat{\gamma}_1'\mathbf{x}$; Figure 3(a)–(e) shows what is seen from several angles.

Projections of $\mathbf{x}$ along the least squares direction and the most nonlinear direction are both highly correlated with $u_1$ and $u_2$; in fact, they are

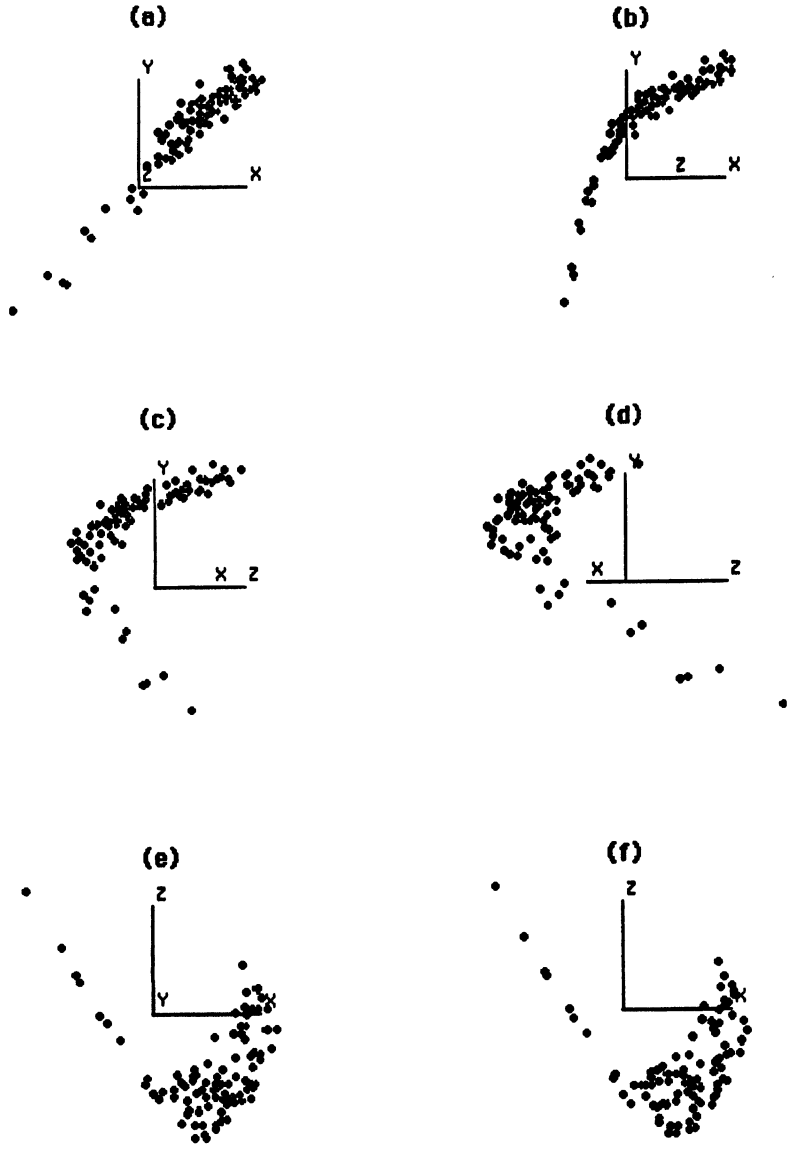


FIG. 3. *(a)–(e) are views of the three-dimensional plot for the quasi-helix found in Example* 3.1; *(f ) is an approximation of (e).*

approximated very well by linear combinations of $u_1$ and $u_2$:

$$\hat{\beta}_{ls}'\mathbf{x} \approx -0.64 + 0.70u_1 + 0.75u_2,$$
$$\hat{\gamma}_1'\mathbf{x} \approx 5.59u_1 - 1.30u_2,$$

each with a correlation coefficient of more than 0.99. The scatterplot of the

two approximated random variables is given in Figure 3(f), which looks almost identical to the original one, Figure 3(e).

Readers familiar with Brillinger's result, as extended by Li and Duan (1989), Li (1991), Duan and Li (1991a, b), Cook (1994) and Cook and Weisberg (1994), can easily see that by the way the data are generated, $\hat{\beta}_{ls}$ should asymptotically fall into the space spanned by $u_1$ and $u_2$. Following the arguments in these articles, it can be proved that under (1.1), if $\beta$ falls into a $d$-dimensional subspace spanned by $\beta_1, \ldots, \beta_d$ for which (1.8) holds, then $\hat{\beta}_{ls}$ will be consistent for some vector in this subspace. Furthermore, according to Remark 2.4, with $d = 2$, we can also see why there is only one large eigenvalue in our example. Further discussion on this example is to be continued in Section 3.2.

EXAMPLE 3.2. We analyze the Boston Housing Data [Harrison and Rubinfeld (1978); Breiman and Friedman (1985)] for a low-crime rate group which consists of 374 cases. Take $y$ to be the housing price and the remaining 13 variables as the regressor $\mathbf{x}$. We start with the multiple linear regression of $y$ on $\mathbf{x}$. The scatterplot of $y$ against $\hat{\beta}_{ls}'\mathbf{x}$, Figure 4(a), shows a clear linear trend. Further analysis based on residual plots does not yield clear evidence of severe departure from linearity in the regression function; a linear model appears fine for summarizing the data.

Our nonlinear confounding diagnosis offers a further look at the adequacy of the linear model. After carrying out the search for the most nonlinear direction $\hat{\gamma}_1$ against $\hat{\beta}_{ls}$, the first eigenvalue is found to be much larger than others:

$$(0.44, 0.08, 0.05, 0.03, 0.00, \ldots)$$

Figure 4(e) shows a clear nonlinear trend between the two projections, $\hat{\beta}_{ls}'\mathbf{x}$ ($x$-axis) and $\hat{\gamma}_1'\mathbf{x}$ ($z$-axis). A quasi-helix is exhibited; see Figure 1 and Figure 2(a)–(d). We use $H = 15$ slices here but other values also give similar results.

Both projections, $\hat{\beta}_{ls}'\mathbf{x}$ and $\hat{\gamma}_1'\mathbf{x}$, are linear combinations of the original 13 regressors. This requires a total of 26 loading coefficients to describe them, which may not be simple to interpret. Fortunately, we are able to find a small number of most significant regressors contributing to each combination. As it turns out, the correlation coefficient between $\hat{\beta}_{ls}'\mathbf{x}$ and $x_6$ is about 0.95 and the correlation coefficient between $\hat{\gamma}_1'\mathbf{x}$ and $0.32x_1 + 1.43x_6 + 19.6x_{13}$ is 0.94. Figure 4(f) shows the scatterplot of the two simplified projections, which appears very similar to the original one, Figure 4(e). It is interesting to notice that $x_6$, the number of rooms, is a physical measurement variable, while $x_1$, $x_{13}$ are socioeconomic variables. Typical users of multiple linear regression who rely on $\hat{\beta}_{ls}$ because of the strong linear trend exhibited by Figure 4(a), could end by paying more than enough attention to the physical size aspect, unduly downgrading the importance of other socioeconomic variables. But with the quasi-helix plot found here, such an attitude is easy to rectify. Even if the regression surface as perceived from the least squares direction

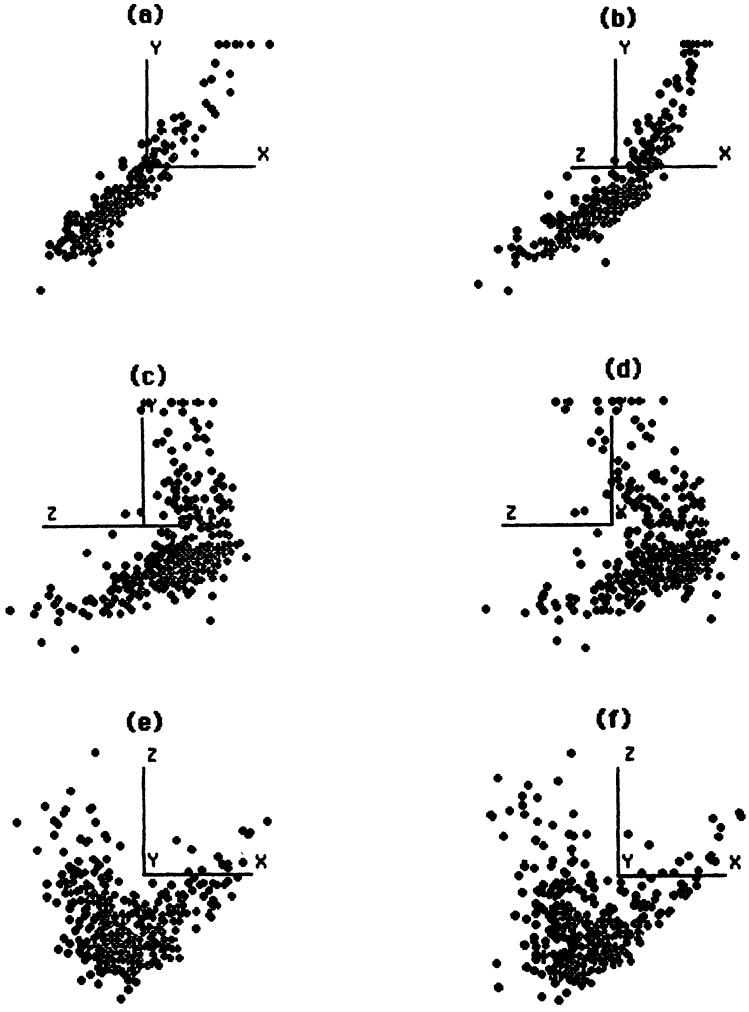


FIG. 4. *(a)–(e) are views of the quasi-helix found in Boston Housing Data; (f) is an approximation of (e) based on three regressors, x1, x6 and x13.*

appears to be approximately linear, the true function can be quite nonlinear and the least squares direction can have a serious bias; see Section 4 for the theory supporting this statement.

In each of the above two examples, the starting direction **b** was taken to be the slope vector $\hat{\beta}_{ls}$ obtained from linear least squares regression. Our searching procedure finds the most nonlinear direction $\hat{\gamma}_1$ against $\hat{\beta}_{ls}$, yielding a three-dimensional scatterplot, $y$ against projections along these two

directions, for scrutiny. By rotating this three-dimensional plot, we anticipate to find a quasi-helixlike pattern if the first eigenvalue is large.

We choose to start with the least squares estimate $\hat{\beta}_{ls}$ merely because of its popularity. Other regression estimates including those from generalized linear models should also be considered. Since such estimates are often obtained via weighted least squares, weighted versions of the $\kappa$ measure may be preferable. Ideally, it would be even better to carry out more extensive search by computing $\kappa_{\mathbf{b}}$ for as many $\mathbf{b}$ as possible. But the total amount of computing in large dimension could be quite heavy and efficient ways of implementation have not yet been developed.

We also remind the reader that standard regression diagnostic tools should still be consulted in connection with the quasi-helix diagnostics offered in this section. As one referee pointed out, in Example 3.1, if the distribution of regressor variables have heavy tails, then $\hat{\beta}_{ls}$ may be unduly influenced by outliers. Robust regression estimates are natural alternatives to use in such cases.

3.2. *Further consideration.* Perhaps the most appropriate time to study our three-dimensional plot for nonlinear confounding is after the linear model (1.5) has already passed several rounds of diagnostic checking, including at least an inspection on the plot of $y$ (and/or the residuals) against the fitted values. We are nearly ready to accept (1.5) as a reasonable model. At this moment, we may want to check the validity of the linearity condition (1.6) [with $\beta$ being estimated by $\hat{\beta}_{ls}$ according to the working model (1.5)]. This can be done by finding $\kappa_{\hat{\beta}_{ls}}$. According to Remark 2.3, if the $\kappa$ value turns out small, then there is nothing much to worry about. We have not found any evidence indicating that (1.6) is seriously violated. Such a conclusion supports the applicability of Brillinger's result, which in turn suggests that the scatterplot of $y$ (and/or the residuals) against the fitted values offers information about the form of $f$ in (1.1). A cycle is now reached because in the beginning, we have already used the suggested plot (among others) to verify (1.5). We are in a more confident position to accept (1.5) if after completing this cycle of checking, we find nothing contradictory.

But we do have a lot to worry about if on the other hand, the $\kappa$ measure turns out sizable. As will be explored further in Section 4, the perceived linear input–output pattern from the earlier diagnosis may be misleading. In this situation, exposing the quasi-helical structure helps visualize the potential weakness in summarizing the data solely by the linear model. As illustrated in Section 1.1, one such danger is in prediction, which is even harder to see in the higher dimensional situation without dimension reduction. Validity of prediction can be visually reassessed by examining the shape and the boundary of the sampled points in our three-dimensional plots.

Now a question about model building arises. Besides providing some cautionary notes about linear fitting, what would be the most appropriate step to take next? This is a harder question to answer and more work is clearly called for. Nevertheless, some tentative options are discussed here.

We begin with the observation that for the two examples from Section 3.1, all eigenvalues other than the largest one, obtained during the search of the most nonlinear direction, are very small. For a situation like this, we may want to consider a two-dimensional version of (1.7), $d = 2$, for these reasons.

1. The additional dimension helps a bit in sidestepping the model ambiguity issue discussed in Section 1. In the case of total confounding (Section 1.2), the minimum dimension reduction space (Section 1.4) is not unique; hence it makes sense to find a space containing all of them.
2. It is also not hard to generate three-dimensional data patterns similar to Example 1 from models with a minimum dimension of 2. For instance, the model can be taken as the average of (1.2) and (1.4). This further points to the difficulty in reducing dimension to just one for such data.
3. For the purpose of prediction, the ambiguity in determining the true dimension may not be a real concern. As we pointed out in Section 1.2, it is still better to examine the entire two-dimensional projection even if the true minimum dimension is 1.

Now suppose we are willing to increase the dimension $d$ from 1 to 2 in (1.7). Then the next model building step is to use $\hat{\beta}_{ls}$ and the associated most nonlinear direction $\hat{\gamma}_1$ as e.d.r. directions $\beta_1, \beta_2$ for (1.7). A rigorous justification of this recommendation requires condition (1.8). With (1.8), it can be proved that asymptotically, $\hat{\beta}_{ls}$ falls into the e.d.r. space and so does $\hat{\gamma}_1$.

The requirement of (1.8) calls for an accompanied diagnostic checking procedure, which is similar to our earlier use of $\kappa$ measure for checking (1.6). This time we need a two-dimensional extension of $\kappa$ measure as discussed in Remark 2.5. We can compute the $\kappa$-measure along the proposed two-dimensional e.d.r. space. If the value turns out to be small, then (1.6) appears reasonable and we can settle the case with $d = 2$. Otherwise, we are led to an even higher minimum dimension.

Up to now, we only focus on the first eigenvector in (2.9). If the minimum dimension is no more than two and (1.8) holds, then all eigenvalues except for the first one have to be 0. Other cases are more complicated and may lead to more than one nonzero eigenvalue. We can examine all large eigenvectors, but how to pursue this in a systematic way is still open for investigation. An alternative thought would be to divorce the analysis from the multiple linear regression approach. It may be worthwhile to start the model-building process with SIR or other methods of dimension reduction. The nonlinear confounding diagnosis can be carried out in a similar way. The method of Section 2 allows the flexibility of setting the starting direction $\mathbf{b}$ to whatever is needed, including the directions suggested by SIR and others.

**4. Overlinearization in multiple linear regression.** The population version of multiple linear regression concerns the minimization

$$\min_{a,\mathbf{b}} E(y - a - \mathbf{b}'\mathbf{x})^2,$$

which is equivalent to

$$\min_{a,\mathbf{b}} E(G(\mathbf{x}) - a - \mathbf{b}'\mathbf{x})^2, \tag{4.1}$$

where $G(\mathbf{x}) = E(y|\mathbf{x})$ is the regression surface. Denote the solution of (4.1) by $a_{ls}$, $\beta_{ls}$. When the regression surface $G(\mathbf{x})$ is nonlinear, the plane $a_{ls} + \beta_{ls}'\mathbf{x}$ serves as a linear approximation to $G(\mathbf{x})$. The size of the approximation error can be measured by $EH_{ls}(\mathbf{x})^2$, where

$$H_{ls}(\mathbf{x}) = G(\mathbf{x}) - a_{ls} - \beta_{ls}'\mathbf{x}. \tag{4.2}$$

Now suppose under (1.1), the regression function $G(\mathbf{x})$ can be expressed as

$$G(\mathbf{x}) = g(\beta'\mathbf{x}) \tag{4.3}$$

for some nonlinear function $g(\cdot)$. If the $\beta$ direction is available, then the linear approximation to $g(\cdot)$:

$$\min_{c_0, c_1} E(g(u) - c_0 - c_1 u)^2, \quad u = \beta'\mathbf{x} \tag{4.4}$$

should have an error of size $Eh(u)^2$, where

$$h(u) = g(u) - c_0^* - c_1^* u, \tag{4.5}$$

and $(c_0^*, c_1^*)$ denotes the least squares solution of (4.4).

The least squares direction $\beta_{ls}$ is in general different from the true direction $\beta$. Thus the term, $EH_{ls}(\mathbf{x})^2$, reflects only the error in approximating $g(\cdot)$ along a wrong direction. This wrong direction makes $g(\cdot)$ appear more linear than it really is. The term $EH_{ls}(\mathbf{x})^2$ is smaller than $Eh(u)^2$. The following definition is introduced to quantify the discrepancy between these two terms.

DEFINITION 4.1. For a nonlinear function of the form (4.3), define the overlinearization of multiple linear regression to be the ratio:

$$OL = \frac{\text{var}(h(u)) - \text{var}(H_{ls}(\mathbf{x}))}{\text{var}(h(u))}$$

If $\beta_{ls}$ is in the same direction as $\beta$, then (4.2) and (4.5) are the same, implying $OL = 0$. For elliptically contoured distributions, this is indeed the case in view of Brillinger's result in Section 1.4.

The following theorem shows that the $OL$ ratio is strongly affected by nonlinearity in the regressor. The proof is given in Section A.1.

THEOREM 4.1. *For a nonlinear regression function of the form* (4.3), *the OL ratio is bounded by the* $\kappa$ *measure of nonlinearity*:

$$OL \le \kappa_{\beta_{ls}'\mathbf{x}|\beta'\mathbf{x}} \le \kappa_\beta.$$

Theorem 4.1 has an interesting implication in residual analysis. A popular notion in this area is that model departure from linearity can be detected from various residual plots. This is in line with the result of Theorem 4.1 if $\kappa_\beta$ is small. The $OL$ ratio will be forced small so that most part of the nonlinear-

ity in $g(\cdot)$ will be retained in the residual of multiple linear regression. But if $\mathbf{x}$ is highly nonlinear, then overlinearization might be serious and not many nonlinear signals may be left in the residual plots. The following example illustrates such a situation.

EXAMPLE 4.1. We generate the data again as described in Example 3.1, but with a higher noise level, $\sigma = 0.5$. Nevertheless, if we know the true direction and regress $y$ against the correct projection $u_1$, then the residual plot, Figure 5(a), still shows a clear nonlinear trend. However, such a trend cannot be found in the standard residual plot, residuals versus predicted values, both from multiple linear regression of $y$ on $\mathbf{x}$; see Figure 5(b). The response surface is overlinearized to an extent that model violation cannot be detected by examining the residuals. Standard regression diagnostics would not find any problems and a multiple linear regression model seems acceptable.

Better diagnostics can be performed by carrying out the searching procedure as in Example 3.1. The first eigenvalue is large, about 0.45, a sign for possible nonlinear confounding. Figure 5(c)–(e) show some angles from the rotation plot of $y$ against $\hat{\beta}_{ls}'\mathbf{x}$ ($x$-axis) and $\hat{\gamma}_1'\mathbf{x}$ ($z$-axis). As expected, they suggest different regression models. In fact, $u_1$ has a 0.99 correlation coefficient with a special projection $0.27\hat{\beta}_{ls}'\mathbf{x} + 0.15\hat{\gamma}_1'\mathbf{x}$. If we remove the linear trend from the plot of $y$ against this special projection, then the residual pattern, Figure 5(f) is almost identical to Figure 5(a). Of course, we cannot directly observe $u_1$ and we cannot tell which angle of the three-dimensional object would suggest the most appropriate model. But at least after revealing such weakness, we should be less willing to accept the linear model now.

We turn to a related question: how close is the least squares direction $\beta_{ls}$ to the true direction $\beta$? We measure closeness by the square of the correlation coefficient between $\beta'\mathbf{x}$ and $\beta_{ls}'\mathbf{x}$. The quantity $1 - \rho(\beta'\mathbf{x}, \beta_{ls}'\mathbf{x})^2$ shows the size of bias in using $\beta_{ls}$ to estimate the direction of $\beta$.

THEOREM 4.2 (Maximum bias for linear least squares direction). *For any $\delta > 0$, consider the class $\mathcal{F}_\delta$ of regression functions with the form* (4.3), *such that the ratio* $\operatorname{var} h(u)/\operatorname{var} g(u)$ *does not exceed $\delta$. Then the maximum bias of the linear least squares direction over this class is given by*

$$\max_{G(\cdot)\in\mathcal{F}_\delta} 1 - \rho(\beta_{ls}'\mathbf{x}, \beta'\mathbf{x})^2 = \frac{\delta\kappa_\beta}{1 - \delta(1 - \kappa_\beta)}. \tag{4.6}$$

The proof of this theorem is given in Section A.2. The right-hand side of (4.6) vanishes when either $\delta$ or $\kappa_\beta$ equals 0. The former case, $\delta = 0$, confirms the unbiasedness of the least squares when the linear model assumption is exact while the latter case, $\kappa = 0$, reestablishes Brillinger's result (Section 1.4).

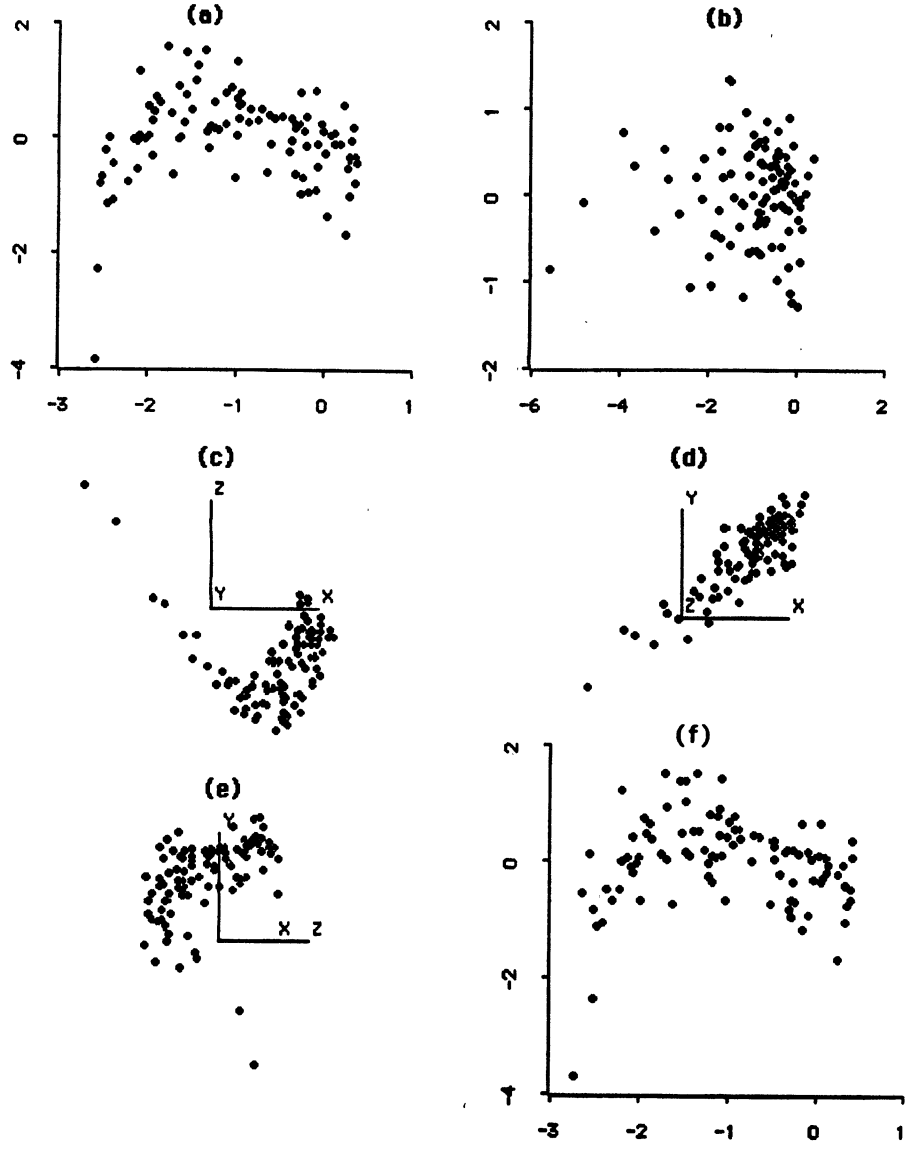


FIG. 5. *Overlinearization*: (*a*) *is the residual plot from regressing y against the true projection. The nonlinearity cannot be found in the standard residual plot* (*b*) *from regressing y on x*; (*c*)–(*e*) *are views on the quasi-helix*; (*f*) *is the residual plot from regressing y on a special projection in the quasi-helix.*

Abbreviate $\rho = \rho(\beta'\mathbf{x}, \beta_{ls}'\mathbf{x})$. We can also relate OL, $\rho$ to the $R$-square of the multiple linear regression, $R^2 = \text{var}(\beta_{ls}'\mathbf{x})/\text{var}\, y$:

$$OL = \frac{(1-\rho^2)R^2}{(1-\rho^2R^2 - (\text{var}(\varepsilon)/\text{var}(y)))} \geq \frac{(1-\rho^2)R^2}{1-\rho^2R^2}, \tag{4.7}$$

$$\rho^2 = \frac{R^2 - OL(1-(\text{var}(\varepsilon)/\text{var}(y)))}{R^2(1-OL)} \geq \frac{R^2-OL}{R^2(1-OL)}. \tag{4.8}$$

The identity in (4.7) can be verified directly from the definition of OL. The identity in (4.8) is obtained by rearranging terms on both sides of the identity in (4.7). Now, using Theorem 4.1, we can obtain

$$\rho^2 \geq \frac{R^2-\kappa_\beta}{R^2(1-\kappa_\beta)}.$$

This is the same as the bound derived in Duan and Li (1991b) where a different argument without the overlinearization measure was used.

**5. Overfit in nonlinear approximation.** Besides linear functions, certain classes of nonlinear functions are also frequently used in approximation. Sigmoidal functions such as the logistic map $g_0(t) = (1+e^{-t})^{-1}$, for example, are popular in neural network modeling [e.g., White (1989)]. Analogous to (4.1), we consider

$$\min_{a,v,c_0,c_1} E\big(G(\mathbf{x}) - [c_0 + c_1(g_0(a+v'\mathbf{x}))]\big)^2, \tag{5.1}$$

where $g_0$ is a given nonlinear function. For the regression function of the form (4.3), we consider the best approximation in the true direction:

$$\min_{c_0,c_1,c_2,c_3} E\big(g(\beta'\mathbf{x}) - [c_0 + c_1(g_0(c_2+c_3\beta'\mathbf{x}))]\big)^2. \tag{5.2}$$

Analogous to the definition of overlinearization, the overfit measure, denoted by $OF$, can be defined by the ratio

$$OF = \frac{(II)-(I)}{(II)}$$

where $(I)$ [respectively $(II)$] denotes the minimum of (5.1) [respectively (5.2)].

In Section A.3, we shall derive the following upper bound for $OF$:

$$OF \leq \max_v \tau_{v'\mathbf{x}|\beta'\mathbf{x}}, \tag{5.3}$$

where the definition of $\tau_{x_2|x_1}$, given by (A.3.3) in Section A.3, is a generalization of $\kappa_{x_2|x_1}$ involving $g_0(\cdot)$. When $g_0(\cdot)$ is linear, $\tau_{x_2|x_1}$ is reduced to $\kappa_{x_2|x_1}$.

**6. Model uncertainty and information loss.** The least squares estimate $\hat{\beta}_{ls}$ is efficient when the true regression function is linear and the errors are i.i.d. normal. The Fisher information matrix for the slope vector $\beta$ per

observation is equal to $\mathcal{I} = \sigma^{-2}\Sigma_{\mathbf{x}}$ and the covariance matrix of $\hat{\beta}_{ls}$ is $n^{-1}\mathcal{I}^{-1}$. If the true function is only approximately linear, how much information will be lost? To answer this question, we may parametrize any reasonable patterns of departure from linearity, treating the additional parameters as the nuisance parameters. The presence of nuisance parameters typically reduces the information for $\beta$. A least favorable parametric model is the one which yields the smallest information $\mathcal{I}_{\min}$. The information loss, $\mathcal{I} - \mathcal{I}_{\min}$, will be examined in this section. As it turns out, a major source of the loss comes from nonlinearity in the regressor. If the distribution of $\mathbf{x}$ is elliptically symmetric, there will be no loss at all in estimating $\beta$ up to a proportionality constant. On the other hand, the loss may be grave if serious nonlinear confounding is present.

Consideration of least favorable parametric models is essential in semiparametrics and adaptiveness studies; useful background can be found in Bickel, Klassen, Ritov and Wellner (1992). A semiparametric model suitable for our discussion has the following form:

$$y = g(\alpha + \beta'\mathbf{x}) + \varepsilon, \tag{6.1}$$

where $\varepsilon$ is normal with mean 0 and variance $\sigma^2$. We shall treat $g(\cdot)$ as the nonparametric component and $\alpha, \beta$ as the parametric component of the semiparametric model. In Section 6.1, we shall first construct a least favorable parametric model by assuming that the true regression function falls within a small neighborhood of a given function $g_0(\cdot)$. After that, information loss under approximate linearity will be discussed under the special case that $g_0(\cdot)$ is an identity function.

6.1. *Least favorable parametric model.* For any fixed direction $\beta_0$, first construct the vector-valued function

$$\eta(u) = E(\mathbf{x}|\beta_0'\mathbf{x} = u). \tag{6.2}$$

Note that this function is determined by the joint distribution of $\mathbf{x}$. After this function is constructed, the least favorable parametric model can be expressed as

$$y = g_0(\alpha + \beta'\mathbf{x} - \delta'\eta(\beta'\mathbf{x})) + \varepsilon \tag{6.3}$$

with $\delta$ being a $p$-dimensional nuisance parameter. When $\delta = 0$, (6.3) is reduced to (6.1) with $g(\cdot) = g_0(\cdot)$.

THEOREM 6.1. *For the parametric model* (6.3), *the Fisher information of* $\beta$ *per observation is equal to*

$$\mathcal{I}_{\min} = \sigma^{-2}E\dot{g}_0(\alpha_0 + \beta_0'\mathbf{x})^2(\mathbf{x} - \eta(\beta_0'\mathbf{x}))(\mathbf{x} - \eta(\beta_0'\mathbf{x}))'. \tag{6.4}$$

*This Fisher information is no greater than the information matrix for* $\beta$ *derived from any parametric model*

$$y = g(\alpha + \beta'\mathbf{x}; \tilde{\delta}) + \varepsilon \tag{6.5}$$

*at* $\beta = \beta_0$, $\tilde{\delta} = 0$, $\alpha = \alpha_0$, *where* $\tilde{\delta}$ *is any finite-dimensional nuisance parameter with* $g(\alpha_0 + \beta_0'\mathbf{x}; 0) = g_0(\alpha_0 + \beta_0'\mathbf{x})$.

The proof is given in the Section A.4. We briefly describe how our least favorable model is found. Consider a direction $\beta$ in a small neighborhood of $\beta_0$. The principle of least favorable model is to find a $g$ function so that $g(\alpha + \beta'\mathbf{x})$ can be as close to $g_0(\alpha + \beta_0'\mathbf{x})$ as possible because this would make the task of discriminating $\beta$ from $\beta_0$ the most difficult. For the extreme case that $\beta_0'\mathbf{x}$ is a function of $\beta'\mathbf{x}$, say $\beta_0'\mathbf{x} = h(\beta'\mathbf{x})$, the least favorable $g$ is obvious. We need only to set $g(\alpha + u) = g_0(\alpha + h(u))$, in which case we cannot distinguish between $\beta$ and $\beta_0$ at all, a completely confounded case as already mentioned in Section 1. For the general case, the idea is to replace $h(u)$ by the conditional expectation $E(\beta_0'\mathbf{x}|\beta'\mathbf{x} = h)$. This leads to

$$g(\alpha + u) = g_0(\alpha + E(\beta_0'\mathbf{x}|\beta'\mathbf{x} = u)).$$

Observe that

$$\begin{aligned} E(\beta_0'\mathbf{x}|\beta'\mathbf{x} = u) &= E(\beta'\mathbf{x}|\beta'\mathbf{x} = u) - (\beta - \beta_0)'E(\mathbf{x}|\beta'\mathbf{x} = u) \\ &= u - (\beta - \beta_0)'E(\mathbf{x}|\beta_0'\mathbf{x} = u) + o(\|\beta - \beta_0\|) \end{aligned} \tag{6.6}$$

Dropping the little $o$ term and taking $\beta - \beta_0$ as $\delta$, we have

$$g(\alpha + u) \approx g_0(\alpha + u - \delta'\eta(u)).$$

This gives the least favorable model (6.3).

6.2. *Information loss for nearly linear regression.* The information loss when the regression function is only approximately linear can be assessed by taking

$$g_0(u) = u$$

together with small nuisance parameters in Theorem 6.1. The minimum Fisher information is found to be

$$\mathcal{I}_{\min} = \sigma^{-2}E(\mathbf{x} - \eta(\beta_0'\mathbf{x}))(\mathbf{x} - \eta(\beta_0'\mathbf{x}))' = \sigma^{-2}\operatorname{cov}\mathbf{e}, \tag{6.7}$$

where $\mathbf{e} = \mathbf{x} - \eta(\beta_0'\mathbf{x})$. To compare this with the Fisher information $\mathcal{I}$ from the linear model theory, we recall the notations in (2.5)–(2.7) with $\mathbf{b} = \beta_0$ to obtain the decomposition

$$\mathbf{x} = L_{\beta_0}(\beta_0'\mathbf{x}) + \mathcal{K}_{\beta_0}(\beta_0'\mathbf{x}) + \mathbf{e}.$$

Now $\sigma^2\mathcal{I} = \operatorname{cov}(\mathbf{x})$ can be decomposed as

$$\begin{aligned} \sigma^2\mathcal{I} &= \operatorname{cov}L_{\beta_0}(\beta_0'\mathbf{x}) + \operatorname{cov}\mathcal{K}_{\beta_0}(\beta_0'\mathbf{x}) + \operatorname{cov}\mathbf{e} \\ &= (\operatorname{var}(\beta_0'\mathbf{x}))^{-1}\Sigma_{\mathbf{x}}\beta_0\beta_0'\Sigma_{\mathbf{x}} + \Sigma_{\beta_0} + \sigma^2\mathcal{I}_{\min}. \end{aligned} \tag{6.8}$$

The first term in (6.8) is a rank one matrix. The loss of this term is because we cannot distinguish $\beta$ from its scalar multiple under (6.1); any scalar multiple of $\beta$ can be absorbed into the $g$. The least favorable model (6.3) also

reflects this limitation. This is because for any $\delta$ which is proportional to $\beta_0$, the function $\delta'\eta(\cdot)$ becomes a multiple of the identity function. We can only identify $\beta$ up to a constant of proportionality.

The second and the third terms in (6.8) constitute the information for estimating $\beta$ up to a proportionality constant in the linear model. To see this, consider any homogeneous function of $\beta$:

$$\phi(c\beta) = \phi(\beta) \quad \text{for any } c \neq 0.$$

One such example is $\phi(\beta) = (\beta/\|\beta\|)\text{sign}(\beta)$, where sign, taking values $+1$ or $-1$, can be set in any convenient way so that the function will be invariant under the sign change of the argument. The asymptotic covariance matrix for the least squares estimate $\phi(\hat{\beta}_{ls})$ equals

$$\nabla\phi(\beta_0)\mathcal{I}^{-1}\nabla\phi(\beta_0)' = \nabla\phi(\beta_0)(\Sigma_{\beta_0} + \mathcal{I}_{\min})^{-}\nabla\phi(\beta_0)',$$

where the superscript minus denotes a generalized inverse matrix, and the gradient operator yields a row vector. The identity comes from the fact that $\nabla\phi(\beta_0)$ is orthogonal to $\beta_0$.

We are ready to claim the following observation.

OBSERVATION 6.1. When $g$ is approximately linear, the information loss in estimating the regression slope vector up to a proportionality constant is equal to $\sigma^{-2}\Sigma_{\beta_0}$.

The most nonlinear direction $\gamma_1$ against $\beta_0$ given by Lemma 2.1 (with $\mathbf{b} = \beta_0$) is the direction where the relative information loss is the greatest. This observation depicts well the relationship between the nonlinear confounding and the information loss. If $\kappa_{\beta_0}$ is zero then there is no information loss. But when the nonlinear confounding is serious, $\kappa_{\beta_0}$ is large and so is the information loss. The scatterplot of the projection $\gamma_1'\mathbf{x}$ against $\beta_0'\mathbf{x}$ reveals the nonlinear pattern that causes the most serious information loss. Suppose we are allowed to collect more data. Then information on this direction can be fortified if the projections of the added data points on this plot are distributed in a way to flatten out the curvilinearity as much as possible. This opens up a new aspect in guiding the selection of a future sample from any given set of possible sites.

6.3. *Information loss in nonlinear regression.* Suppose (6.1) is known to hold exactly for a specified nonlinear function $g_0$. The Fisher information $\mathcal{I}$ for $\beta$ is easy to find. Using the notations of weighted expectation and covariance from Section 2, we have

$$\mathcal{I} = \sigma^{-2}Ew(\beta_0'\mathbf{x})(\Sigma_{\mathbf{x};w(\beta_0'\mathbf{x})})$$

with the weight

$$w(\beta_0'\mathbf{x}) = \dot{g}_0(\alpha_0 + \beta_0'\mathbf{x})^2.$$

To compare with the minimum information matrix from Theorem 6.1, we can use the weighted version of (6.8). We can attribute the loss of the first term (with rank 1) to the nonidentifiability of the length of $\beta$ again. Subject to this constraint, the information loss is found to be

$$\sigma^{-2} Ew(\beta_0' \mathbf{x})(\Sigma_{\beta_0; w})$$

Similarly to the case of linear regression, if the weighted $\kappa$ measure in $\mathbf{x}$ is large, then the information loss can be serious.

**7. Hypothesis testing for nearly linear regression.** This section discusses the impact of nonlinear confounding on the practice of the $F$-test in linear models. Again, since linear models are rarely exact, the popularity of the $F$-test is based on the common notion that it is still valid in some approximation sense as long as linear models are acceptable. We shall examine this notion carefully.

Specifically, for a given direction $\beta_0$, consider

$$H_0\colon \beta = c\beta_0 \quad \text{for some } c \in R \quad \text{vs.} \quad H_1\colon \beta \neq c\beta_0 \quad \text{for any } c.$$

We shall study the significance level and the power of $F$-test when the regression model (6.1) is only approximately linear. It turns out that they depend critically on the $\kappa$ measure of nonlinearity in $\mathbf{x}$. Our results provide some support for the use of the $F$-test if the $\kappa$ measure is small. But if the $\kappa$ measure is large, then the level of the $F$-test can be blown up quickly and the power of the test may be seriously impaired as well. To better accommodate nonlinearity in $\mathbf{x}$, an alternative testing procedure will be proposed.

7.1. *Significance level of the F-test.* For a given sample, $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, the usual $F$-test is based on the ratio

$$F = \frac{(SSE_0 - SSE_1)/(p-1)}{SSE_1/(n-p-1)} \tag{7.1}$$

$SSE_0$ = residual sum of squares under $H_0$

$SSE_1$ = residual sum of squares under $H_1$

Under the exact linear model assumption, $SSE_0/\sigma^2$ and $SSE_1/\sigma^2$ are chi-square random variables with $(n-2)$ and $(n-p-1)$ degrees of freedom respectively when the null hypothesis is true. When the sample size is large, $SSE_1/(n-p-1)$ converges to $\sigma^2$, and we can approximate the $F$-test by a chi-square test with $p-1$ degrees of freedom. The null hypothesis is rejected if $(p-1)F > C_{p-1,\alpha}$ where $C_{p-1,\alpha}$ denotes the $(1-\alpha)$ quantile of the chi-square distribution with $p-1$ degrees of freedom.

It may be questionable to apply the $F$-test unless we can verify that the linear model assumption is acceptable. This can be done by going through various residual diagnostic checking and even conducting series of formal model testing procedures. But accepting a linear model at most implies that the true model is approximately linear. To define a class $C_\delta$ of approximately

linear functions $g(\cdot)$ for our study, consider the nonlinear part $h(\cdot)$ in $g(\cdot)$ as defined by (4.5) with $u = \beta_0'\mathbf{x}$. The class $C_\delta$ consists of all functions $g(\cdot)$ with the variance of the nonlinear part $\operatorname{var} h(\beta_0'\mathbf{x})$ bounded by $\delta\sigma^2$ for a fixed positive number $\delta$. The smaller the $\delta$ index is set, the stricter linearity on the regression function is required.

Denote the level of the test at $g(\cdot)$ by

$$\alpha_g = P\{(p-1)F > C_{p-1;\alpha} | H_0\}.$$

We shall show in Section A.5 that

$$\max_{g \in C_\delta} \alpha_g \approx P\left\{\chi^2_{p-1}(n\delta\kappa_{\beta_0}) > C_{p-1;\alpha}\left(1 + (1-\kappa_{\beta_0})\delta\right)\right\}, \tag{7.2}$$

where $\chi^2_{p-1}(\phi)$ denotes the noncentral chi-squared random variable of $p-1$ degrees of freedom with noncentrality parameter $\phi$.

Equation (7.2) shows that the chance of falsely rejecting $H_0$ is an increasing function of the $\kappa$ measure in $\mathbf{x}$ along the direction $\beta_0$. In order to keep the $\alpha$ level under control, the noncentrality parameter $n\delta\kappa_{\beta_0}$ must be small. This can be achieved either if $\kappa_{\beta_0}$ is small or if $\delta$ is small. Thus when nonlinear confounding is not serious, the standard F-test may not be too sensitive to mild violation of the linearity model assumption. But if $\kappa_{\beta_0}$ is large, a much more stringent linearity condition has to be imposed. Asymptotically, in order to keep the noncentrality parameter bounded, $\delta$ must be within the order of $n^{-1}$. This could be hard to guarantee under the standard linear model checking practice; see Example 7.1 in Section 7.3 for an illustration.

REMARK 7.1. A related result concerning the F-test can be found in Li and Duan (1989). Their results imply that if the distribution of $\mathbf{x}$ is normal, then asymptotically the standard F-test statistics still follows an $F$-distribution even if the model (1.1) is nonlinear.

7.2. *Loss of power in F-test.* The power of an $F$-test under an exact linear model assumption can be approximated by using noncentral chi-squared distributions. In our case, the power at $\beta$ is approximately equal to

$$P_\beta \approx P\{\chi^2_{p-1}(\phi) > C_{p-1;\alpha}\},$$

where the noncentrality parameter $\phi$ is just the residual sum of squares for regressing $\beta'\mathbf{x}$ on $\beta_0'\mathbf{x}$. It can be shown that

$$\phi \approx n\left(\beta'\Sigma_{\mathbf{x}}\beta - (\beta'\Sigma_{\mathbf{x}}\beta_0)^2/\beta_0'\Sigma_{\mathbf{x}}\beta_0\right)/\sigma^2,$$

which increases rapidly at rate $n$. However, if the true regression function $g(\alpha + \beta'\mathbf{x})$ deviates from linearity, then the noncentrality parameter can get either smaller or larger. The worst situation is when the (population version) least squares estimate $\beta_{ls}$ as defined in (4.1) with $G(\mathbf{x}) = g(\alpha + \beta'\mathbf{x})$, which is now biased, falls on the direction of $\beta_0$ (instead of the true direction $\beta$):

$$\beta_{ls} = c\beta_0 \quad \text{for some } c. \tag{7.3}$$

This is the case where the noncentrality parameter will not tend to infinity as $n$ increases. The power of the test remains bounded away from one even for very large sample sizes; in other words, the F-test will not be consistent.

Consider $g(\cdot)$ with the decomposition

$$\begin{aligned} g(\alpha + \beta'\mathbf{x}) &= \alpha + \beta'\mathbf{x} + h(\beta'\mathbf{x}), \\ \operatorname{var} h(\beta'\mathbf{x}) &\le \delta^0, \end{aligned} \tag{7.4}$$

where $h(\beta'\mathbf{x})$ is uncorrelated with $\beta'\mathbf{x}$. We want to find the smallest $\delta^0$, denoted by $\delta_\beta$, so that there exists such a $g$ for which the worst situation (7.3) happens. We may consider $\delta_\beta$ as the largest amount of nonlinearity that can be incorporated into the linear model to maintain the consistency of the F-test.

In the Section A.5, we shall show that

$$\delta_\beta \ge \kappa_\beta^{-1} \mathbf{v}_0' \Sigma_{\mathbf{x}} \mathbf{v}_0, \tag{7.5}$$

where $\mathbf{v}_0$ is the orthogonal complement of $\beta$, $\mathbf{v}_0' \Sigma_{\mathbf{x}} \beta = 0$, such that $\mathbf{v}_0 + \beta = c\beta_0$ for some constant $c$.

We see that for maintaining consistency of the F-test, larger deviation from linearity can be allowed if $\kappa_\beta$ is small.

7.3. *An augmented test.* We have shown that if $\kappa_{\beta_0}$ is large, a small departure from exact linearity for the true regression function can force the standard F-test to fail badly in preserving its nominal significance level. To find a remedy, our idea is to augment the exact linear model with a small number of nuisance parameters $\delta_i$, $i = 1, \ldots, k$ so that important small departure from linearity can be accommodated:

$$y = \alpha + \beta'\mathbf{x} + \sum_i^k \delta_i z_i + \varepsilon. \tag{7.6}$$

The auxiliary regressors $z_i$'s are constructed by incorporating the nonlinearity in $\mathbf{x}$ along the direction $\mathbf{b} = \beta_0$, (2.5)–(2.9):

$$z_i = \gamma_i' \mathscr{H}_{\beta_0}(\beta_0'\mathbf{x}), \qquad i = 1, \ldots, k \le p. \tag{7.7}$$

Now we can conduct a standard F-test under this augmented model to see if $H_0$: $\beta = c\beta_0$, for some constant $c$ holds.

$$F^* = \frac{(SSE_0^* - SSE_1^*)/(p-1)}{SSE_1^*/(n-p-k-1)}$$

$SSE_1^*$ = residual sum of squares under (7.6)

$SSE_0^*$ = residual sum of squares under (7.6) and $H_0$.

We then reject $H_0$ if $F^*$ exceeds the $(1-\alpha)$ quantile of an $F$-distribution with degrees of freedom $(p-1, n-p-k-1)$.

To carry out this test, we need to estimate $z_i$. This can be done by (1) carrying out the eigenvalue decomposition as given in Section 3 and keeping

only the first $k$ eigenvectors, $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$; (2) finding the conditional expectation of $\hat{\gamma}_i'\mathbf{x}$ against $\beta_0'\mathbf{x}$ by any nonparametric regression method.

EXAMPLE 7.1. For the data generated in Example 4.1, consider the hypothesis

$$H_0\colon \beta = c\beta_0 = c(1, 0, -1, -1, 0)' \quad \text{for some } c.$$

In this case, $\beta_0$ is the true direction that generates the data. The standard F-test (7.1) gives a large value of $F = 12.6$, with $R$-squares of 0.587 and 0.730, respectively, under the null and the alternative hypotheses. Thus $H_0$ is falsely rejected by the standard F-test.

We now carry out the procedure for finding the most nonlinear direction against the direction $\mathbf{b} = \beta_0$. Only one large eigenvalue is found, $\tilde{\lambda}_1 = 0.628$, $\hat{\gamma}_1 = (4.60, -1.47, -4.66, -3.02, 1.46)'$. This direction $\hat{\gamma}_1$ is used to carry out the new test. First, we apply the LOWESS function from Xlips.stat (Tierney 1991) for smoothing; Figure 6 shows the nonparametric fit for $\hat{\gamma}_1'\mathbf{x}$ against $\hat{\beta}_{ls}'\mathbf{x}$. The fitted values are used as the input for the additional regressor $z_1$ in the augmented linear model (7.6) with $k = 1$. Then the new $F$-test is conducted. The new $F$-value, $F^*$, is now reduced to 1.57, with the $R$-squares of 0.729 under $H_0$ and 0.746 under the alternative hypothesis. The $p$-value of the test is 0.19, big enough to accept $H_0$. The correct null hypothesis will not be rejected by our new test. Note that Figure 7.1 is produced using the default values provided by the software. Other values have been tried and yield similar results.

The nuisance parameter $\delta$ in the augmented linear model (7.6) serves as an outlet for absorbing the bias of multiple linear regression due to the nonlinearity in $g(\cdot)$. The immediate consequence is that the expectation of $SSE_0^* - SSE_1^*$ remains bounded as $n$ increases. The significance level is under much better protection. Under $H_0$, we can show that even if $g(\cdot)$ is

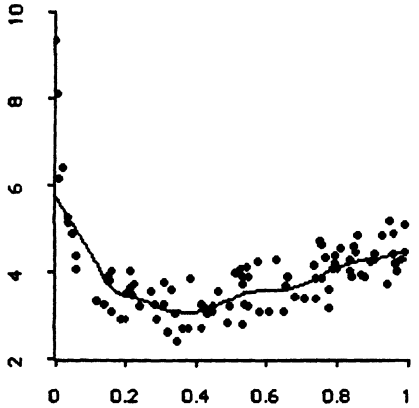


FIG. 6. *Smoothing by LOWESS for the most nonlinear direction against the null direction.*

nonlinear, the least squares fit under the augmented linear model (7.6) with $k$ equal to the number of nonzero eigenvalues in (2.9) always gives the correct $\beta$ direction; a solution for the following minimization is achieved at $\mathbf{b} = c^*\beta_0$ for some constant $c^*$:

$$\min_{a, \mathbf{b}, \delta} E\big(g(\alpha + \beta_0'\mathbf{x}) - a - \mathbf{b}'\mathbf{x} - \delta'\mathscr{K}_{\beta_0}(\beta_0'\mathbf{x})\big)^2. \tag{7.8}$$

An easy way to see this is to consider the decomposition of $\mathbf{x}$ into the linear part, the nonlinear part and $\mathbf{e} = \mathbf{x} - E(\mathbf{x}|\beta_0'\mathbf{x})$ again. We have

$$E\Big(g(\alpha + \beta_0'\mathbf{x}) - a - \mathbf{b}'\big(L_{\beta_0}(\beta_0'\mathbf{x})\big) - (\mathbf{b} + \delta)'\mathscr{K}_{\beta_0}(\beta_0'\mathbf{x}) - \mathbf{b}'\mathbf{e}\Big)^2$$

$$= E\Big(g(\alpha + \beta_0'\mathbf{x}) - a - \mathbf{b}'\big(L_{\beta_0}(\beta_0'\mathbf{x})\big) - (\mathbf{b} + \delta)'\mathscr{K}_{\beta_0}(\beta_0'\mathbf{x})\Big)^2 + E(\mathbf{b}'\mathbf{e})^2.$$

The last term $E(\mathbf{b}'\mathbf{e})^2$ vanishes if $\mathbf{b}$ is in the direction of $\beta_0$. This shows that the minimization (7.8) can be achieved by restricting $\mathbf{b}$ to the direction $\beta_0$, as claimed.

REMARK 7.2. A referee pointed out that the idea of augmenting a multiple linear model can be found in other places, especially in the context of discussing partial residual plots and its generalization by Cook (1993). Further connection is worth exploring. Another related front is in the general area of semiparametric models and adaptive estimation. An outline of how to use the model augmentation idea iteratively for constructing an adaptive estimate of $\beta$ was given in an earlier version of this article.

REMARK 7.3. Taking $\beta_0$ to be a $p$ by $d$ matrix, we can extend the discussion to the more general hypothesis:

$H_0$: $\beta$ belongs to the subspace spanned by column vectors of $\beta_0$.

The extended notion of $\kappa$ measure given in Remark 2.5 is needed.

**8. Conclusion.** This paper introduces a $\kappa$ measure for assessing nonlinearity in the regressor $\mathbf{x}$. A procedure based on this measure is proposed for finding projections that may help detect the presence of nonlinear confounding. The procedure starts with any tentative candidate $\mathbf{b}$ for the e.d.r. direction $\beta$ in (1.1). Given $\mathbf{b}$, a most nonlinear direction $\mathbf{v}$ is found so that the scatterplot of $\mathbf{v}'\mathbf{x}$ against $\mathbf{b}'\mathbf{x}$ has a strongest nonlinear trend. The three-dimensional plot of $y$ against these two projections of $\mathbf{x}$ may exhibit a quasi-helical pattern if the $\kappa$ measure is large. The presence of such structures in the data leads to several kinds of difficulties in regression analysis. The following is a summary of the issues we have addressed.

1. Exploratory study. Quasi-helical patterns are unusual features in the data that call for special attention. As such a three-dimensional data cloud is rotated on the computer screen, different curve patterns are observed from different projection angles.

2. Prediction. Because of (1), there is a lot of ambiguity in specifying a suitable prediction model. Examining the two-dimensional projection of $\mathbf{x}$ provided by the three-dimensional plot is helpful in locating regions where prediction of $y$ may be harder.
3. Overlinearization. Multiple linear regression offers a best linear approximation to $E(y|\mathbf{x})$ by the least squares method. But the nature of minimization can also make the regression surface appear more linear than it really is. Consequently, under serious nonlinear confounding, plots of residuals from multiple linear regression may become ineffective for detecting nonlinearity in the regression function. We use the $\kappa$ measure to set an upper bound on the size of overlinearization.
4. Information loss. Regression models are rarely exact. A small deviation from the assumed model can lead to a large information loss in estimating the direction of the true $\beta$ parameter. The main source of information loss is found to come from nonlinearity in $\mathbf{x}$.
6. Validity of inference. Nonlinearity in $\mathbf{x}$ can cause standard F-tests to exceed their nominal levels substantially. The loss in the power of the test can also be serious. An augmented F-test is introduced for tempering the problem.

Although our use of the $\kappa$ measure has primarily been set in the context of multiple linear regression analysis, it is worthwhile to extend it to other situations where methods such as sliced inverse regression (SIR), sliced average variance estimate (SAVE), principal Hessian direction (pHd), sliced inverse regression-II (SIR-II) are appropriate; see Carroll and Li (1992), Cook and Weisberg (1991), Duan and Li (1991a), Hsing and Carroll (1992), Li (1990, 1991, 1992a, b).

Nonlinear confounding is an intricate issue which has not received proper attention in the literature. Our work is just a beginning. There are still many unanswered problems and some of these are pointed out in Section 3.2.

In a context rather different from ours, nonlinear association between regressors has also received attention in the analysis of additive models. A concept of "concurvity" is brought up [Buja, Hastle and Tibshirani (1989)] and diagnostic methods are proposed [Gu (1992)]. The application of these ideas to the study of nonlinear confounding is not explored yet. One referee suggests that Aldrin, Bølviken and Schweder (1993) may be relevant.

## APPENDIX

**A.1. Proof of Theorem 4.1.** First, (4.4) can be viewed as fitting the model (4.1) with the constraint that $\mathbf{b}$ is proportional to $\beta$. Let $\mathbf{b}_r$ be the slope vector from regressing the residual $h(u)$ on $\mathbf{x}$:

$$\min_{a,\mathbf{v}} E\big(h(u) - a - \mathbf{v}'\mathbf{x}\big)^2. \tag{A.1.1}$$

It is easy to see that $H_{ls}(\mathbf{x}) = h(u) - \mathbf{b}_r'\mathbf{x}$, $\operatorname{var} H_{ls}(\mathbf{x}) = \operatorname{var} h(u) - \operatorname{var}(\mathbf{b}_r'\mathbf{x})$ and

$$\beta_{ls} = \mathbf{b}_r + c_1^*\beta. \tag{A.1.2}$$

The overlinearization measure $OL$ can be written as the

$$\begin{aligned} OL &= \frac{\operatorname{var}(\mathbf{b}_r'\mathbf{x})}{\operatorname{var}(h(u))} = \rho\big(h(u), \mathbf{b}_r'\mathbf{x}\big)^2 \\ &= \frac{\operatorname{cov}\big(h(u), \mathbf{b}_r'\mathbf{x}\big)^2}{\operatorname{var}(h(u))\operatorname{var}(\mathbf{b}_r'\mathbf{x})} \\ &= \frac{\operatorname{cov}\big(h(u), \mathbf{b}_r'\mathscr{I}_\beta(u)\big)^2}{\operatorname{var}(h(u))\operatorname{var}(\mathbf{b}_r'\mathbf{x})} \\ &= \rho\big(h(u), \mathbf{b}_r'\mathscr{I}_\beta(u)\big)^2 \cdot \kappa_{\mathbf{b}_r'\mathbf{x}|\beta'\mathbf{x}} \le \kappa_{\mathbf{b}_r'\mathbf{x}|\beta'\mathbf{x}}, \end{aligned}$$

where the second identity is due to the least squares property of (A.1.1); the fourth identity follows from the fact that the term $\mathbf{x} - E(\mathbf{x}|u)$ is uncorrelated with any function of $u$.

To complete the proof, we simply apply (2.3) and (A.1.1) to obtain $\kappa_{\mathbf{b}_r'\mathbf{x}|\beta'\mathbf{x}} = \kappa_{\beta_{ls}'\mathbf{x}|\beta'\mathbf{x}}$.

**A.2. Proof of Theorem 4.2.** From the decomposition (A.1.2), we can derive

$$\begin{aligned} \rho^2(\beta_{ls}'\mathbf{x}, \beta'\mathbf{x}) &= \frac{\operatorname{var}(c_1^*\beta'\mathbf{x})}{\operatorname{var}\beta_{ls}'\mathbf{x}} = \frac{\operatorname{var}(c_1^*u)}{\operatorname{var}(c_1^*u) + \operatorname{var} h(u) - \operatorname{var} H_{ls}(u)} \\ &= \frac{\operatorname{var}(c_1^*u)/\operatorname{var} h(u)}{\operatorname{var}(c_1^*u)/(\operatorname{var} h(u)) + OL}. \end{aligned}$$

Now since $\operatorname{var}(c_1^*u)/\operatorname{var} h(u) \le (1-\delta)/\delta$, the last expression is no greater:

$$\frac{(1-\delta)/\delta}{(1-\delta)\delta^{-1} + OL} \ge \frac{1-\delta}{1-\delta+\kappa_\beta\delta}.$$

This proves Theorem 4.2. □

**A.3. Proof of (5.3).** We first introduce an extension of the $\kappa$ measure for random variable $x_2$ against random variable $x_1$ involving a nonlinear transformation $g_0(\cdot)$. Consider the decomposition

$$\begin{aligned} g_0(a + x_2) &= E\big(g_0(a + x_2)|x_1\big) + e \\ &= \big[c_0^* + c_1^* g_0(c_2^* + c_3^* x_1)\big] + q(x_1) + e, \end{aligned} \tag{A.3.1}$$

where $(c_0^*, c_1^*, c_2^*, c_3^*)$ is the solution of

$$\text{(A.3.2)} \qquad \min_{c_0, c_1, c_2, c_3} E\big[E(g_0(a + x_2)|x_1) - c_0 - c_1 g_0(c_2 + c_3 x_1)\big]^2$$

and $e, q$ are given implicitly to satisfy the equation. Define

$$\text{(A.3.3)} \qquad \tau_{x_2|x_1;a} = \frac{\operatorname{var} q}{\operatorname{var} q + \operatorname{var} e},$$
$$\tau_{x_2|x_1} = \max_a \tau_{x_2|x_1;a}.$$

We begin to prove (5.3) now. Take $x_1 = \beta'\mathbf{x}$ and $x_2 = \mathbf{v}'\mathbf{x}$ in (A.3.1)–(A.3.3). Let

$$\text{(A.3.4)} \qquad A = \frac{SD(c_1^*(g_0(c_2^* + c_3^*\beta'\mathbf{x})))}{SD(g_0(a + \mathbf{v}'\mathbf{x}))} \le 1$$

$$\text{(A.3.5)} \qquad B = \frac{SD(q(\beta'\mathbf{x}))}{SD(g_0(a + \mathbf{v}'\mathbf{x}))} = \sqrt{(1 - A^2)\tau_{\mathbf{v}'\mathbf{x}|\beta'\mathbf{x};a}}$$

where SD$(\cdot)$ denotes standard deviation. Let $\rho_1$ be the correlation between $g(\beta'\mathbf{x})$ and $g_0(c_2^* + c_3^*\beta'\mathbf{x})$, and $\rho_2$ be the correlation between $g(\beta'\mathbf{x})$ and $q(\beta'\mathbf{x})$. Then due to the orthogonality of decomposition (A.3.1), we have

$$\text{(A.3.6)} \qquad \rho_1^2 + \rho_2^2 \le 1.$$

Now the $R$-squares between $g(\beta'\mathbf{x})$ and $g_0(a + \mathbf{v}'\mathbf{x})$ can be written as

$$\rho(g(\beta'\mathbf{x}), g_0(a + \mathbf{v}'\mathbf{x}))^2 = \frac{\operatorname{cov}(g(\beta'\mathbf{x}), E(g_0(a + v'\mathbf{x})|\beta'\mathbf{x}))^2}{\operatorname{var}(g(\beta'\mathbf{x}))\operatorname{var}(g_0(a + \mathbf{v}'\mathbf{x}))}$$
$$= (\rho_1 \operatorname{sgn}(c_1^*)A + \rho_2 B)^2,$$

where $\operatorname{sgn}(c_1^*)$ denotes the sign of $c_1^*$. The minimum for (5.1) can be written as

$$(I) = \operatorname{var}(g(\beta'\mathbf{x})) \min_{a,\mathbf{v}} (1 - \rho(g(\beta'\mathbf{x}), g_0(a + \mathbf{v}'\mathbf{x}))^2$$
$$= \operatorname{var}(g(\beta'\mathbf{x}))\Big[1 - \max_{a,\mathbf{v}} (\rho_1 \operatorname{sgn}(c_1^*)A + \rho_2 B)^2\Big]$$

The minimum for (5.2) can be bounded by

$$(II) \le \operatorname{var}(g(\beta'\mathbf{x})) \cdot (1 - \rho_1^2)$$

Thus we can bound OF by

$$OF \le \max_{\mathbf{v},a} \frac{(\rho_1 \operatorname{sgn}(c_1^*)A + \rho_2 B)^2 - \rho_1^2}{1 - \rho_1^2}$$
$$\le \max_{\mathbf{v},a} \tau_{\mathbf{v}'\mathbf{x}|\beta'\mathbf{x};a} = \max_{\mathbf{v}} \tau_{\mathbf{v}'\mathbf{x}|\beta'\mathbf{x}},$$

where the last inequality can be derived by maximization over the range of $A$ first, using (A.3.4) and (A.3.5), and then over $\rho_1$, using (A.3.6). The proof is now complete. □

**A.4. Proof of Theorem 6.1.** To show (6.4), we follow a generic procedure for calculating the Fisher information of the main parameter $\beta$ in presence of the nuisance parameter $\delta$ when data come from the family $\{f(y; \beta, \delta)\}$. The first step is to compute the Fisher scores,

$$S_\beta = \frac{\partial \log f(y; \beta, \delta)}{\partial \beta}; \qquad S_\delta = \frac{\partial \log f(y; \beta, \delta)}{\partial \delta}.$$

The second step is to eliminate the influence of $S_\delta$ on $S_\beta$ by orthogonalization. A matrix $M$ can be found so that $S = S_\beta - MS_\delta$ is uncorrelated with $S_\delta$; in other words, $S$ is the residual for regressing $S_\beta$ on $S_\delta$ linearly. The final step is to find the covariance of $S$, which is the Fisher information for $\beta$.

Following the above recipe and treating both $\alpha$ and $\delta$ as nuisance parameters in model (6.3), at $\alpha = \alpha_0$, $\beta = \beta_0$, $\delta = 0$, we find

$$\begin{aligned} S_\alpha &= \sigma^{-2}\varepsilon \dot{g}_0(\alpha_0 + \beta_0'\mathbf{x}) \\ S_\beta &= \sigma^{-2}\varepsilon \dot{g}_0(\alpha_0 + \beta_0'\mathbf{x})\mathbf{x} \\ S_\delta &= \sigma^{-2}\varepsilon \dot{g}_0(\alpha_0 + \beta_0'\mathbf{x})\eta(\beta_0'\mathbf{x}) \\ S &= \sigma^{-2}\varepsilon \dot{g}_0(\alpha_0 + \beta_0'\mathbf{x})(\mathbf{x} - \eta(\beta_0'\mathbf{x})). \end{aligned}$$

Our calculation can be confirmed by checking that $S$ is uncorrelated with $S_\alpha, S_\delta$:

$$\operatorname{cov}(S, S_\alpha) = \sigma^{-2}E\left[\dot{g}_0(\alpha_0 + \beta_0'\mathbf{x})^2 E(\mathbf{x} - \eta(\beta_0'\mathbf{x})|\beta_0'\mathbf{x})\right] = 0$$

$$\operatorname{cov}(S, S_\delta) = \sigma^{-2}E\left[\dot{g}_0(\alpha_0 + \beta_0'\mathbf{x})^2 \eta(\beta_0'\mathbf{x}) E(\mathbf{x} - \eta(\beta_0'\mathbf{x})|\beta_0'\mathbf{x})\right] = 0.$$

The covariance matrix of $S$ is equal to (6.4). But we still have to show that it is the minimum. Let $\tilde{S}_\alpha, \tilde{S}_\beta, \tilde{S}_{\tilde{\delta}}$ be the Fisher scores for model (6.5). Then at $\beta = \beta_0$, $\tilde{\delta} = 0$, $\alpha = \alpha_0$, we have $\tilde{S}_\alpha = S_\alpha$, $\tilde{S}_\beta = S_\beta$. The term $\tilde{S}_{\tilde{\delta}}$ equals $\sigma^2 \varepsilon g_2(\alpha_0 + \beta_0'\mathbf{x}; 0)$, where $g_2(\cdot;\cdot)$ denotes the partial derivative of $g(\cdot;\cdot)$ with respect to the second argument. Regardless of the form of the partial derivative, $\tilde{S}_{\tilde{\delta}}$ is always a function of $\beta_0'\mathbf{x}$. Thus it must be uncorrelated with $S$. This is because $E(S_\beta|\beta_0'\mathbf{x}) = S_\delta$, which implies that $S$ $(= S_\beta - S_\delta)$ is uncorrelated with any function of $\beta_0'\mathbf{x}$.

Now the adjusted Fisher score, which takes the form of $\tilde{S}_\beta - M\tilde{S}_{\tilde{\delta}}$ for some matrix $M$ can be decomposed into the sum of two uncorrelated terms, $S$ and $[(S_\beta - S) - M\tilde{S}_{\tilde{\delta}}]$. This shows that the Fisher information is equal to $\operatorname{cov} S + \operatorname{cov}[(S_\beta - S) - M\tilde{S}_{\tilde{\delta}}]$, which is no less than $\operatorname{cov} S = I_{\min}$. This proves the theorem. □

**A.5. Derivation of (7.2).** For any $g$ in the class $C_\delta$, it is easy to see that $SSE_1/n$ converges to

$$\operatorname{var}(\varepsilon + H_{ls}(\beta_0'\mathbf{x})) = \sigma^2 + \operatorname{var} H_{ls}(\beta_0'\mathbf{x}) = \sigma^2(1 + (1 - OL)\delta^*).$$

Similarly, $SSE_0/n$ converges to $\sigma^2(1 + \delta^*)$. Thus the noncentrality parameter for $(SSE_0 - SSE_1)/\sigma^2$ is approximately equal to $(n\delta^*)OL$. It follows that

approximately

$$(p-1)F \sim (1+(1-OL)\delta^*)^{-1}\chi^2_{p-1}((n\delta)OL).$$

Therefore, we have

$$\alpha_g \approx P\{\chi^2_{p-1}((n\delta^*)OL) > (1+(1-OL)\delta)C_{p-1;\alpha}\}$$

Now (7.2) follows from Theorem 4.1. □

**A.6. Proof of (7.5).** First of all, it is enough to consider $h(\cdot)$ with the form

$$h(\beta'\mathbf{x}) = \mathbf{u}'\mathscr{X}_\beta(\beta'\mathbf{x})$$

for some vector $\mathbf{u}$, $\mathbf{u}'\Sigma_{\mathbf{x}}\beta = 0$. This is because for any $h(\beta'\mathbf{x})$, we can always reduce $\delta^0$ by considering its projection on the linear space spanned by $\beta'\mathbf{x}$ and $\mathscr{X}_\beta(\beta'\mathbf{x})$. This does not change the least squares estimate $\beta_{ls}$.

In order to get (7.3), we must have

$$\begin{aligned} c\beta_0 &= \Sigma_{\mathbf{x}}^{-1}\operatorname{cov}(\mathbf{x}, g(\alpha+\beta'\mathbf{x})) \\ &= \Sigma_{\mathbf{x}}^{-1}(\Sigma_{\mathbf{x}}\beta + \Sigma_\beta \mathbf{u}) \end{aligned}$$

which implies that

$$\Sigma_\beta \mathbf{u} = \Sigma_{\mathbf{x}}(c\beta_0 - \beta).$$

Since $\beta'\Sigma_\beta = 0$, multiplying $\beta'$ on both sides of the last expression, we see that $\mathbf{v}_0 = c\beta_0 - \beta$ is orthogonal to $\beta$ with respect to $\Sigma_{\mathbf{x}}$. Hence $\mathbf{u}$ must satisfy the equation $\Sigma_\beta \mathbf{u} = \Sigma_{\mathbf{x}}\mathbf{v}_0$. The variance of $h(\beta'\mathbf{x})$ can be written as

$$\begin{aligned} \mathbf{u}'\Sigma_\beta \mathbf{u} &= \mathbf{v}_0'\Sigma_{\mathbf{x}}\Sigma_\beta^-\Sigma_{\mathbf{x}}\mathbf{v}_0 \\ &\geq \kappa_\beta^{-1}\mathbf{v}_0'\Sigma_{\mathbf{x}}\mathbf{v}_0 \end{aligned}$$

The proof of (7.5) is now complete. □

**Acknowledgments.** I received helpful comments from Dennis Cook and benefited from discussions with audiences when this paper was presented on several occasions. Their input is very much appreciated. I am also grateful to an Associate Editor and two referees for valuable comments which lead to a better presentation of this paper.

## REFERENCES

ALDRIN, M., BØLVIKEN, E. and SCHWEDER, T. (1993). Projection pursuit regression for moderate non-linearities. *Comput. Statist. Data Anal.* **16** 379–403.

BICKEL, P. J., KLASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1992). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–246.

BOX, G. E. P. and DRAPER, N. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, New York.

BREIMAN, L. and FRIEDMAN, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–597.

BRILLINGER, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 97–114. Wadsworth,

BRILLINGER, D. R. (1991). Discussion of "Sliced inverse regression." *J. Amer. Statist. Assoc.* **86** 333–333.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555.

CARROLL, R. J. and LI, K. C. (1992). Measurement error regression with unknown link: dimension reduction and data visualization. *J. Amer. Statist. Assoc.* **87** 1040–1050.

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

COOK, R. D. (1993). Exploring partial residual plots. *Technometrics* **35** 351–362.

COOK, R. D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc.* **89** 177–189.

COOK, R. D. and NACHTSHEIM, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *J. Amer. Statist. Assoc.* **89** 592–599.

COOK, R. D. and WEISBERG, S. (1991). Discussion of "Sliced inverse regression" by K. C. Li. *J. Amer. Statist. Assoc.* **86** 328–332.

COOK, R. D. and WEISBERG, S. (1994). *An Introduction to Regression Graphics*. Wiley, New York.

COX, D. R. and SNELL, E. J. (1981). *Applied Statistics: Principles and Examples*. Chapman & Hall, New York.

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

DUAN, N. and LI, K. C. (1991b). A bias bound for applying linear regression to a general linear model. *Statist. Sinica* **1** 127–136.

FRIEDMAN, J. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GU, C. (1992). Diagnostics for nonparametric regression models with additive terms. *J. Amer. Statist. Assoc.* **87** 1051–1058.

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

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

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178.

HÄRDLE, W. and STOKER, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995.

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

HSING, T. and CARROLL, R. J. (1992). Asymptotic properties of sliced inverse regression. *Ann. Statist.* **20** 1040–1061.

LI, K. C. (1990). Data-visualization with SIR: a transformation-based projection pursuit method. Technical report.

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

LI, K. C. (1992a). Uncertainty analysis for mathematical models with SIR. In *Probability and Statistics* (J. Ze-Pei, Y. Shi-Jian, C. Ping and W. Rong, eds.) 138–162. World Scientific, Singapore.

LI, K. C. (1992b). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87** 1025–1039.

LI, K. C. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

SAMAROV, A. (1993). Exploring regression structure using functional estimation. *J. Amer. Statist. Assoc.* **88** 836–847.

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

WHITE, H. (1989). Some asymptotic results for learning in single hidden-layer feed-forward network models. *J. Amer. Statist. Assoc.* **84** 1003–1013.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA 90024
E-MAIL: kcli@math.ucla.edu