Comment

DAVID R. BRILLINGER*

At the outset I would like to congratulate Ker-Chau Li for developing a new data analytic tool and for providing such a substantial study of it. Regarding the technique itself, an issue that arises quickly is just how robust is it to departures from the fundamental assumption of Section 3? To address this somewhat I attempted to employ the technique on some spatial-temporal global meteorological data. I was interested in studying motion present in this data. Suppose that $Y(x, y, t)$ denotes the measurement made at location (x, y) at time t. Suppose that energy is propagating as a plane wave. The motion may then be represented as

$$f(\alpha x + \beta y + \gamma t)$$

for a function f. This corresponds to movement in direction ϕ, given by $\tan \phi = \beta/\alpha$, with speed $\gamma/\sqrt{\alpha^2 + \beta^2}$. If two such waves are present then one can comtemplate a model

$$Y(x, y, t) = f_1(\alpha_1 X + \beta_1 y + \gamma_1 t) + f_2(\alpha_2 x + \beta_2 y + \gamma_2 t) + \textit{noise}.$$

* David R. Brillinger is Professor, Department of Statistics, University of California, Berkeley, CA 94720.

In the circumstance of concern, the measurements were made for (x, y, t) on a lattice. Such data values, (x, y, t), do not satisfy the critical assumption of Section 3. In an attempt to have such a condition obtain I proceeded as follows. Consider a normal distribution $\phi(x)\phi(y)\phi(t)$ centered on the domain of measurements. Obtain realizations of this distribution and let (x_j, y_j, t_j) denote the location on the lattice closest to the jth realization. Now subject the data $(x_j, y_j, t_j, Y(x_j, y_j, t_j))(j = 1, \ldots, n)$ to the analysis of the article. I wondered if Ker-Chau Li's technique would detect motion of weather fronts. I have to report that I was not successful. However the technique certainly did work with corresponding simulated data. (In the simulations, there was a single wave, and f was the cosine function.) There are clearly many things going on in the example, so the failure is not discouraging. The analysis was carried out in *S*. This package is so widely used now, so one thing I recommend to Ker-Chau Li is that he prepare versions of his programs in *S*.

I would like to end by mentioning how satisfied Ker-Chau Li's thesis supervisor, Jack Kiefer, would surely have been to see how Ker-Chau Li's work has become such a fine blend of theory and practice.

Comment

W. HÄRDLE AND A. B. TSYBAKOV*

The article by Li proposes a new and very useful approach to dimensionality reduction in multivariate nonparametric regression. The advantage of this approach as compared to others is the exceptional simplicity both of the idea and of the computational tools. We suppose that this would give rise to a wide implementation of sliced inverse regression (SIR).

As with many simple ideas, of course, SIR will also have its pitfalls in "nonsimple" situations. In particular, SIR depends very much on the probability structure of the $\mathbf{x}$ variables described by the following:

* W. Härdle is Professor, CORE, Université Catholique de Louvain, 34, Voie du Roman Pays, B-1348 Louvain-la-Neuve, Belgium. A. B. Tsybakov is Senior Researcher, Institute for Problems of Information Transmission, Academy of Sciences, U.S.S.R., 101447 Moscow, GSP-4, Ermolovoystr., 19 U.S.S.R.

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

$$E(b\mathbf{x} \mid \beta_1\mathbf{x}, \ldots, \beta_K\mathbf{x}) = c_0 + c_1\beta_1\mathbf{x} + \cdots + c_K\beta_K\mathbf{x}. \quad (3.1)$$

A nonsimple situation might be where the distribution of $\mathbf{x}$ is a mixture of two normal distributions or has a more complicated nonelliptical structure. In this case, a nonparametric technique based on estimating the multivariate density of $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ might be reasonable to check the assumption (3.1). We discuss an approach based on this (more complicated) technique later.

There are at least two questions that are important for a practitioner: How to choose the number of principal direc-

Journal of the American Statistical Association
June 1991, Vol. 86, No. 414, Theory and Methods

tions K and how to choose the number of slices H? These questions are addressed to some extent, but we feel that they deserve some more comments.

It is said that the root n consistency property in estimation of directions holds no matter how H is chosen and that it even holds when each slice contains only two observations. This is probably somewhat misleading. If H can be chosen arbitrarily, then it seems possible to use the simplest estimate, that is, to put $H = 1$. But this is, of course, bizarre, since in this case $p_h = 1$ and the estimate will be close to $m_h = E(Z) = 0$. When H increases, the number of nontrivial eigenvectors of the matrix V will also increase, although it will not be evident for what H all the K principal eigenvectors are present. This could suggest, rather, that H should be chosen large to make sure that we catch all the principal directions. Thus one might incline to the other extreme, that is, choosing only two observations per slice. To understand this extreme, let us think of one observation per slice, then $\hat{V} = \sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T$. Thus the principal directions are chosen from the covariance structure of $\mathbf{x}$, as in principal components analysis. Thus, between these two extremes of SIR, there is a lot of freedom, which makes alternative approaches interesting. One of them is based on a different method of identifying the e.d.r. space, another is based on average derivative estimation (ADE). Finally, we propose a nonparametric version of factor analysis.

Let us consider instead of V the matrix

$$B = E_Y[E(\mathbf{x} \mid y)E(\mathbf{x}^T \mid y)]$$

(assume here that $\mathbf{x}$ is already standardized). Elements of B can be expressed as

$$b_{jk} = \int m_j(y)m_k(y)F(dy),$$

where $m_j(y)$ is the regression function of y on the jth component of $\mathbf{x}$ and F is the marginal distribution of y. To estimate b_{jk}, replace F by the empirical distribution F_n, and m_j, m_k by the nonparametric regression estimates $\hat{m}_j$, $\hat{m}_k$. Thus

$$\hat{b}_{jk} = \int \hat{m}_j(y)\hat{m}_k(y)F_n(dy) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_j(y_i)\hat{m}_k(y_i).$$

The functions $\hat{m}_j$, $\hat{m}_k$ may be kernel, orthogonal series, or any other estimates. If $\hat{m}$ is a regressogram, then we get something very similar to SIR, namely,

$$\hat{B} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(y_i)\hat{m}^T(y_i),$$

where

$$\hat{m}(y) = \frac{1}{np_h}\sum_{h=1}^{H}\sum_{s=1}^{n} I\{y_s \in I_h, y \in I_h\}\tilde{\mathbf{x}}_s.$$

This estimate will of course have a bias decreasing as $H \to \infty$. Similar functionals, like the average derivative, have a variance proportional to $1/n$. We suspect, therefore, that a careful choice of H will yield a $\sqrt{n}$ convergence of $\hat{B}$ to B.

All the eigenvectors of B that correspond to nonzero eigenvalues are contained in the e.d.r. space. In fact, it follows from Corollary 3.1 that

$$E(\mathbf{x} \mid y) = c_1(y)\beta_1 + \cdots + c_K(y)\beta_K,$$

where $c_j(y)$ are some functions. Therefore $B = \sum_{j,m=1}^{K} \tilde{c}_{jm}\beta_j\beta_m^T$, where $\tilde{c}_{jm} = E[c_j(y)c_m(y)]$. Thus if b is not in the e.d.r. space, that is, $b \perp \{\beta_1, \ldots, \beta_K\}$, then $Bb = 0$.

In the simplest case, where $K = 1$, one gets

$$B = \tilde{c}_{11}\beta_1\beta_1^T, \qquad \tilde{c}_{11} = E[c_1^2(y)].$$

Assume that β_1 is normalized, so that $\|\beta_1\| = 1$. Then β_1 is the eigenvector of B corresponding to the maximal eigenvalue $\tilde{c}_{11}$:

$$B\beta_1 = \tilde{c}_{11}\beta_1; \qquad \tilde{c}_{11} \geq b^TBb \qquad \forall\, b : \|b\| = 1.$$

Another approach, first developed for the case $K = 1$, is ADE; see Härdle and Stoker (1989), Härdle, Hart, Marron, and Tsybakov (1989). The *average derivative* is defined by

$$\int \nabla m(\mathbf{x}) f_X(\mathbf{x})\, d\mathbf{x},$$

where $\nabla m(x)$ is the gradient of the unknown regression function $m(\mathbf{x}) = E(Y \mid X = \mathbf{x})$ and $f_X(\mathbf{x})$ is the marginal density of $\mathbf{x}$. The average derivative can be estimated $\sqrt{n}$ consistently. Although all the previous work on ADE was concerned with the case of $K = 1$, its extension to the more general model $y = m(\beta_1^T\mathbf{x}, \ldots, \beta_K^T\mathbf{x}, \varepsilon)$ is straightforward. In fact, the average derivative is then

$$AD = E[\nabla_{\mathbf{x}} m(\beta_1^T\mathbf{x}, \ldots, \beta_K^T\mathbf{x}, \varepsilon)]$$

$$= c_1\beta_1 + \cdots + c_K\beta_K,$$

where

$$c_j = E\left[\frac{\partial}{\partial t} m(\beta_1^T\mathbf{x}, \ldots, \beta_{j-1}^T\mathbf{x}, t, \beta_{j+1}^T\mathbf{x}, \ldots, \beta_K^T\mathbf{x}, \varepsilon)\Big|_{t=\beta_j^T\mathbf{x}}\right].$$

Define the matrix $B_1 = AD \cdot AD^T$. This matrix is an analog of B, defined earlier, since all the eigenvectors of B that correspond to nonzero eigenvalues are in the e.d.r. space. Thus, in the same way as earlier, we can choose the estimates $\hat{\beta}_1, \ldots, \hat{\beta}_K$ of the principal directions as the first K eigenvectors of

$$\hat{B}_1 = \hat{AD}\ \hat{AD}^T,$$

where $\hat{AD}$ is an average derivative estimator.

The choice of the number of principal directions K can be addressed in at least three different ways.

1. The candidates for principal directions are known and ordered; the first K directions are principal; K must be estimated.
2. The candidates for principal directions are known; the number K of principal directions and their positions are unknown; these directions must be estimated.
3. The candidates for principal directions are unknown; their number is also unknown.

Li proposes an interesting way of treating the problem in case (3) for normally distributed $\mathbf{x}$. His approach is based on the correlation structure of $\mathbf{x}$ only. This can be viewed as an analog to sequential hypothesis testing techniques in

linear regression. However, the extension to the case of non-Gaussian $\mathbf{x}$ seems to be somewhat difficult.

Note that (1) is solved if one has a solution of (2). Under (2) we can assume, in general, that possible candidates for the principal directions are all the coordinate axes. For example, this assumption is quite reasonable if one thinks of a nonparametric version of factor analysis. Thus the unknown regression function $m(\mathbf{x})$ ($\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathbf{R}^p$) is of the form

$$m(\mathbf{x}) = \sum_{K=1}^{K} g_{j_k}(\mathbf{x}_{j_k}), \qquad j_k \in \{1, \ldots, p\},$$

where $K < p$ is some integer and $K \geq 1$. The problem is to estimate the set $J = \{j_1, \ldots, j_K\}$. Given a sample $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$ define

$$r_{nj}(\mathbf{x}_j) = \frac{1}{nh_n} \sum_{i=1}^{n} Y_i \tilde{K}\left(\frac{\mathbf{x}_{ij} - \mathbf{x}_j}{h_n}\right),$$

$$f_{nj}(\mathbf{x}_j) = \frac{1}{nh_n} \sum_{i=1}^{n} \tilde{K}\left(\frac{\mathbf{x}_{ij} - \mathbf{x}_j}{h_n}\right).$$

Here, f_{nj} is the kernel estimate of the marginal density f_j of jth component, the $\mathbf{x}_{ij}$'s are the components of the vectors $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})$, $\tilde{K}$ is a kernel, and $h_n > 0$ is a bandwidth. Consider the following procedure of estimating J.

1. Calculate the quantities

$$S_{nj} = \frac{1}{n} \sum_{i=1}^{n} r_{nj}^2(\mathbf{x}_{ij}), \qquad j = 1, \ldots, p.$$

2. Arrange the S_{nj} in the decreasing order

$$S_n^{(1)} \geq S_n^{(2)} \geq \cdots \geq S_n^{(p)}.$$

Let $(1)_n$ be the integer that equals j, with maximal value $S_{nj} = S_n^{(1)}$; let $(2)_n$ be the integer that equals j with $S_{nj} = S_n^{(2)}$. Thus

$$(K)_n = j \in \{1, \ldots, p\} : S_{nj} = S_n^{(K)}.$$

Without loss of generality, assume that all $S_n^{(k)}$ are different [thus $(K)_n$ is uniquely defined]. In particular, we have

$$S_n^{(K)} = \frac{1}{n} \sum_{i=1}^{n} r_{n(K)_n}^2(\mathbf{x}_{i(K)_n}).$$

3. Choose K_n as the minimizer of the following statistic:

$$K_n = [\arg \min_{K \leq p}(S_n^{(p)} + Kb_n)] - 1,$$

where b_n is a sequence that tends to zero as $n \to \infty$ and $nb_n^2 \to \infty$.

The estimate of the set $\{j_1, \ldots, j_K\}$ is defined as $J_n = \{(1)_n, \ldots, (K_n)_n\}$, and the corresponding estimate of the regression function is

$$m_n(\mathbf{x}) = \sum_{K \in J_n} g_{nj_k}(\mathbf{x}_{j_k}),$$

where

$$g_{nj}(x_j) = \frac{r_{nj}(\mathbf{x}_j)}{f_{nj}(\mathbf{x}_j)}.$$

It can be proved that under suitable assumptions $P\{J_n = J\} \to 1$, $n \to \infty$ (Härdle and Tsybakov 1990). Moreover, the estimate $m_n(\mathbf{x})$ is pointwise asymptotically normal and converges to $m(\mathbf{x})$ with the rate that is achievable for the case of univariate regression function estimation.

This idea of estimating "principal components" can be viewed as a modification of AIC-BIC criteria, with the additional reordering of components according to some stochastic criterion. Note that, instead of S_{nj}'s, we could take for reordering any other data-dependent quantities that are asymptotically nonzero for principal components and are zero for negligible components.

REFERENCES

Härdle, W., Hart, J., Marron, J. S., and Tsybakov, A. B. (1989), "Bandwidth Choice for Average Derivative Estimation," unpublished manuscript.

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

Härdle, W., and Tsybakov, A. B. (1990), "How Many Terms Should be Added Into an Additive Model?" unpublished manuscript.