Estimating Optimal Transformations for Multiple Regression and Correlation

LEO BREIMAN and JEROME H. FRIEDMAN*

In regression analysis the response variable Y and the predictor variables $X_1, \ldots, X_p$ are often replaced by functions $\theta(Y)$ and $\phi_1(X_1), \ldots, \phi_p(X_p)$. We discuss a procedure for estimating those functions θ^* and $\phi_1^*, \ldots, \phi_p^*$ that minimize $e^2 = E\{[\theta(Y) - \sum_{j=1}^p \phi_j(X_j)]^2\}/\text{var}[\theta(Y)]$, given only a sample $\{(y_k, x_{k1}, \ldots, x_{kp}), 1 \le k \le N\}$ and making minimal assumptions concerning the data distribution or the form of the solution functions. For the bivariate case, $p = 1$, θ^* and ϕ^* satisfy $\rho^* = \rho(\theta^*, \phi^*) = \max_{\theta,\phi}\rho[\theta(Y), \phi(X)]$, where ρ is the product moment correlation coefficient and ρ^* is the maximal correlation between X and Y. Our procedure thus also provides a method for estimating the maximal correlation between two variables.

KEY WORDS: Smoothing; ACE.

1. INTRODUCTION

Nonlinear transformation of variables is a commonly used practice in regression problems. Two common goals are stabilization of error variance and symmetrization/normalization of error distribution. A more comprehensive goal, and the one we adopt, is to find those transformations that produce the best-fitting additive model. Knowledge of such transformations aids in the interpretation and understanding of the relationship between the response and predictors.

Let $Y, X_1, \ldots, X_p$ be random variables with Y the response and $X_1, \ldots, X_p$ the predictors. Let $\theta(Y), \phi_1(X_1), \ldots, \phi_p(X_p)$ be arbitrary measurable mean-zero functions of the corresponding random variables. The fraction of variance not explained (e^2) by a regression of $\theta(Y)$ on $\sum_{i=1}^p \phi_i(X_i)$ is

$$e^2(\theta, \phi_1, \ldots, \phi_p) = \frac{E\left\{\left[\theta(Y) - \sum_{i=1}^p \phi_i(X_i)\right]^2\right\}}{E\theta^2(Y)}. \quad (1.1)$$

Then define *optimal transformations* as functions $\theta^*, \phi_1^*, \ldots, \phi_p^*$ that minimize (1.1); that is,

$$e^2(\theta^*, \phi_1^*, \ldots, \phi_p^*) = \min_{\theta,\phi_1,\ldots,\phi_p} e^2(\theta, \phi_1, \ldots, \phi_p). \quad (1.2)$$

We show in Section 5 that optimal transformations exist and satisfy a complex system of integral equations. The heart of our approach is that there is a simple iterative algorithm using only bivariate conditional expectations, which converges to an optimal solution. When the conditional expectations are estimated from a finite data set, then use of the algorithm results in estimates of the optimal transformations.

This method has some powerful characteristics. It can be applied in situations where the response or the predictors involve arbitrary mixtures of continuous ordered variables and categorical variables (ordered or unordered). The functions θ, $\phi_1, \ldots, \phi_p$ are real-valued. If the original variable is categorical, the application of θ or ϕ_i assigns a real-valued score to each of its categorical values.

The procedure is nonparametric. The optimal transformation estimates are based solely on the data sample $\{(y_k, x_{k1}, \ldots, x_{kp}), 1 \le k \le N\}$ with minimal assumptions concerning the data distribution and the form of the optimal transformations. In particular, we do not require the transformation functions to be from a particular parameterized family or even monotone. (Later we illustrate situations in which the optimal transformations are not monotone.)

It is applicable to at least three situations:

1. random designs in regression
2. autoregressive schemes in stationary ergodic time series
3. controlled designs in regression.

In the first of these, we assume the data $(y_k, \mathbf{x}_k)$, $k = 1, \ldots, N$, are independent samples from the distribution of $Y, X_1, \ldots, X_p$. In the second, a stationary mean-zero ergodic time series $X_1, X_2, \ldots$ is assumed, the optimal transformations are defined to be the functions that minimize

$$e^2 = \frac{E\left\{\left[\theta(X_{p+1}) - \sum_{j=1}^p \phi_j(X_j)\right]^2\right\}}{E\theta^2(X_{p+1})},$$

and the data consist of $N + p$ consecutive observations $x_1, \ldots, x_{N+p}$. This is put in a standard data form by defining

$$y_k = x_{k+p}, \quad \mathbf{x}_k = (x_{k+p-1}, \ldots, x_k), \quad k = 1, \ldots, N.$$

In the controlled design situation, a distribution $P(dy \mid \mathbf{x})$ for the response variable Y is specified for every point $\mathbf{x} = (x_1, \ldots, x_p)$ in the design space. The Nth-order design consists of a specification of N points $x_1, \ldots, \mathbf{x}_N$ in the design space, and the data consist of these points together with measurements on the response variables $y_1, \ldots, y_N$. The $\{y_k\}$ are assumed independent with y_k drawn from the distribution $P(dy \mid \mathbf{x}_k)$.

Denote by $\hat{P}_N(d\mathbf{x})$ the empirical distribution that gives mass $1/N$ to each of the points $\mathbf{x}_1, \ldots, \mathbf{x}_N$. Assume further that $\hat{P}_N \xrightarrow{w} P$, where $P(d\mathbf{x})$ is a probability measure on the design space. Then $P(dy \mid \mathbf{x})$ and $P(d\mathbf{x})$ determine the distribution of random variables $Y, X_1, \ldots, X_p$, and the optimal transformations are defined as in (1.2).

For the bivariate case, $p = 1$, the optimal transformations $\theta^*(Y)$, $\phi^*(X)$ satisfy

$$\rho^*(X, Y) = \rho(\theta^*, \phi^*) = \max_{\theta,\phi} \rho[\theta(Y), \phi(X)], \quad (1.3)$$

* Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, CA 94720. Jerome H. Friedman is Professor, Department of Statistics and Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94305. This work was supported by Office of Naval Research Contracts N00014-82-K-0054 and N00014-81-K-0340.

Journal of the American Statistical Association
September 1985, Vol. 80, No. 391, Theory and Methods

where ρ is the product-moment-correlation coefficient. The quantity $\rho^*(X, Y)$ is known as the *maximal correlation* between X and Y, and it is used as a general measure of dependence (Gebelein 1947; also see Renyi 1959, Sarmanov 1958a, b, and Lancaster 1958). The maximal correlation has the following properties (Renyi 1959):

1. $0 \leq \rho^*(X, Y) \leq 1$.
2. $\rho^*(X, Y) = 0$ if and only if X and Y are independent.
3. If there exists a relation of the form $u(X) = v(Y)$, where u and v are Borel-measurable functions with $\text{var}[u(X)] > 0$, then $\rho^*(X, Y) = 1$.

Therefore, in the bivariate case our procedure can also be regarded as a method for estimating the maximal correlation between two variables, providing as a by-product estimates of the functions θ^*, ϕ^*, that achieve the maximum.

In the next section, we describe our procedure for finding optimal transformations using algorithmic notation, deferring mathematical justifications to Section 5 and Appendix A. We next illustrate the procedure in Section 3 by applying it to a simulated data set in which the optimal transformations are known. The estimates are surprisingly good. Our algorithm is also applied to the Boston housing data of Harrison and Rubinfeld (1978) as listed in Belsley et al. (1980). The transformations found by the algorithm generally differ from those applied in the original analysis. Finally, we apply the procedure to a multiple time series arising from an air pollution study. A FORTRAN implementation of our algorithm is available from either author. Section 4 presents a general discussion and relates this procedure to other empirical methods for finding transformations.

Section 5 and Appendix A provide some theoretical framework for the algorithm. In Section 5, under weak conditions on the joint distribution of $Y, X_1, \ldots, X_p$, it is shown that optimal transformations exist and are generally unique up to a change of sign. The optimal transformations are characterized as the eigenfunctions of a set of linear integral equations whose kernels involve bivariate distributions. We then show that our procedure converges to optimal transformations.

Appendix A discusses the algorithm as applied to finite data sets. The results are dependent on the type of data smooth employed to estimate the bivariate conditional expectations. Convergence of the algorithm is proven only for a restricted class of data smooths. However, in more than 1,000 applications of the algorithm on a variety of data sets using three different types of data smoothers, only one (very contrived) instance of nonconvergence has been found.

Appendix A also contains proof of a consistency result. Under fairly general conditions, as the sample size increases the finite data transformations converge in a "weak" sense to the distributional space optimal transformations. The essential condition of the theorem involves the asymptotic consistency of a sequence of data smooths. In the case of iid data there are known results concerning the consistency of various smooths. Stone's (1977) pioneering paper established consistency for k-nearest-neighbor smoothing. Devroye and Wagner (1980) and, independently, Spiegelman and Sacks (1980) gave weak conditions for consistency of kernel smooths. See Stone (1977) and Devroye (1981) for a review of the literature.

There are no analogous results, however, for stationary ergodic series or controlled designs. To remedy this we show that there are sequences of data smooths that have the requisite properties in all three cases.

This article is presented in two distinct parts. Sections 1–4 give a fairly nontechnical overview of the method and discuss its application to data. Section 5 and Appendix A are, of necessity, more technical, presenting the theoretical foundation for the procedure.

There is relevant previous work. Closest in spirit to the ACE algorithm we develop is the MORALS algorithm of Young et al. (1976) (also see de Leeuw et al. 1976). It uses an alternating least squares fit, but it restricts transformations on discrete ordered variables to be monotonic and transformations on continuous variables to be linear or polynomial. No theoretical framework for MORALS is given.

Renyi (1959) gave a proof of the existence of optimal transformations in the bivariate case under conditions similar to ours in the general case. He also derived integral equations satisfied by θ^* and φ^* with kernels depending on the bivariate density of X and Y and concentrated on finding solutions assuming this density known. The equations seem generally intractable with only a few known solutions. He did not consider the problem of estimating θ^*, φ^* from data.

Kolmogorov (see Sarmanov and Zaharov 1960 and Lancaster 1969) proved that if $Y_1, \ldots, Y_q, X_1, \ldots, X_p$ have a joint normal distribution, then the functions $\theta(Y_1, \ldots, Y_q)$, $\phi(X_1, \ldots, X_p)$ having maximum correlation are linear. It follows from this that in the regression model

$$\theta(Y) = \sum_{i=1}^{p} \phi_i(X_i) + Z, \tag{1.4}$$

if the $\phi_i(X_i)$, $i = 1, \ldots, p$, have a joint normal distribution and Z is an independent $N(0, \sigma^2)$, then the optimal transformations as defined in (1.2) are $\theta, \phi_1, \ldots, \phi_p$. Generally, for a model of the form (1.4) with Z independent of $(X_1, \ldots, X_p)$, the optimal transformations are not equal to $\theta, \phi_1, \ldots, \phi_p$. But in examples with simulated data generated from models of the form (1.4), with non-normal $\{\phi_i(X_i)\}$, the estimated optimal transformations were always close to $\theta, \phi_1, \ldots, \phi_p$.

Finally, we note the work in a different direction by Kimeldorf et al. (1982), who constructed a linear-programming-type algorithm to find the monotone transformations $\theta(Y)$, $\phi(X)$ that maximize the sample correlation coefficient in the bivariate case $p = 1$.

2. THE ALGORITHM

Our procedure for finding $\theta^*, \phi_1^*, \ldots, \phi_p^*$ is iterative. Assume a known distribution for the variables $Y, X_1, \ldots, X_p$. Without loss of generality, let $E\theta^2(Y) = 1$, and assume that all functions have expectation zero.

To illustrate, we first look at the bivariate case:

$$e^2(\theta, \phi) = E[\theta(Y) - \phi(X)]^2. \tag{2.1}$$

Consider the minimization of (2.1) with respect to $\theta(Y)$ for a given function $\phi(X)$, keeping $E\theta^2 = 1$. The solution is

$$\theta_1(Y) = E[\phi(X) \mid Y]/\|E[\phi(X) \mid Y]\| \tag{2.2}$$

with $\| \cdot \| \equiv [E(\cdot)^2]^{1/2}$. Next, consider the unrestricted minimization of (2.1) with respect to $\phi(X)$ for a given $\theta(Y)$. The solution is

$$\phi_1(X) = E[\theta(Y) \mid X]. \tag{2.3}$$

Equations (2.2) and (2.3) form the basis of an iterative optimization procedure involving *alternating conditional expectations* (ACE).

Basic ACE Algorithm

Set $\theta(Y) = Y/\|Y\|$;
Iterate until $e^2(\theta, \phi)$ fails to decrease:
  $\phi_1(X) = E[\theta(Y) \mid X]$;
  replace $\phi(X)$ with $\phi_1(X)$;
  $\theta_1(Y) = E[\phi(X) \mid Y]/\|E[\phi(X) \mid Y]\|$;
  replace $\theta(Y)$ with $\theta_1(Y)$;
End Iteration Loop;
θ and ϕ are the solutions θ^* and ϕ^*;
End Algorithm.

This algorithm decreases (2.1) at each step by alternatingly minimizing with respect to one function and holding the other fixed at its previous evaluation. Each iteration (execution of the iteration loop) performs one pair of these single-function minimizations. The process begins with an initial guess for one of the functions ($\theta = Y/\|Y\|$) and ends when a complete iteration pass fails to decrease e^2. In Section 5, we prove that the algorithm converges to optimal transformations θ^*, ϕ^*.

Now consider the more general case of multiple predictors $X_1, \ldots, X_p$. We proceed in direct analogy with the basic ACE algorithm. We minimize

$$e^2(\theta, \phi_1, \ldots, \phi_p) = E\left[\theta(Y) - \sum_{j=1}^{p} \phi_j(X_j)\right]^2, \tag{2.4}$$

holding $E\theta^2 = 1$, $E\theta = E\phi_1 = \cdots = E\phi_p = 0$, through a series of single-function minimizations involving bivariate conditional expectations. For a given set of functions $\phi_1(X_1), \ldots, \phi_p(X_p)$, minimization of (2.4) with respect to $\phi(Y)$ yields

$$\theta_1(Y) = E\left[\sum_{i=1}^{p} \phi_i(X_i) \mid Y\right] \Big/ \left\| E\left[\sum_{i=1}^{p} \phi_i(X_i) \mid Y\right] \right\|. \tag{2.5}$$

The next step is to minimize (2.4) with respect to $\phi_1(X_1), \ldots, \phi_p(X_p)$, given $\theta(Y)$. This is obtained through another iterative algorithm. Consider the minimization of (2.4) with respect to a single function $\phi_k(X_k)$ for given $\theta(Y)$ and a given set $\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_p$. The solution is

$$\phi_{k,1}(X_k) = E\left[\theta(Y) - \sum_{i \neq k} \phi_i(X_i) \mid X_k\right]. \tag{2.6}$$

The corresponding iterative algorithm is as follows:

Set $\phi_1(X_1), \ldots, \phi_p(X_p) = 0$;
Iterate until $e^2(\theta, \phi_1, \ldots, \phi_p)$ fails to decrease;
For $k = 1$ to p Do:
  $\phi_{k,1}(X_k) = E[\theta(Y) - \Sigma_{i \neq k} \phi_i(X_i) \mid X_k]$;
  replace $\phi_k(X_k)$ with $\phi_{k,1}(X_k)$;
End For Loop;
End Iteration Loop;
$\phi_1, \ldots, \phi_p$ are the solution functions.

Each iteration of the inner For loop minimizes e^2 (2.4) with respect to the function $\phi_k(X_k)$, $k = 1, \ldots, p$, with all other functions fixed at their previous evaluations (execution of the For loop). The outer loop is iterated until one complete pass over the predictor variables (inner For loop) fails to decrease e^2 (2.4).

Substituting this procedure for the corresponding single function optimization in the bivariate ACE algorithm gives rise to the full ACE algorithm for minimizing the (2.4) e^2.

ACE Algorithm

Set $\theta(Y) = Y/\|Y\|$ and $\phi_1(X_1), \ldots, \phi_p(X_p) = 0$;
Iterate until $e^2(\theta, \phi_1, \ldots, \phi_p)$ fails to decrease;
Iterate until $e^2(\theta, \phi_1, \ldots, \phi_p)$ fails to decrease;
For $k = 1$ to p Do:
  $\phi_{k,1}(X_k) = E[\theta(Y) - \Sigma_{i \neq k} \phi_i(X_i) \mid X_k]$;
  replace $\phi_k(X_k)$ with $\phi_{k,1}(X_k)$;
End For Loop;
End Inner Iteration Loop;
  $\theta_1(Y) = E[\Sigma_{i=1} \phi_i(X_i) \mid Y]/\|E[\Sigma_{i=1}^{p} \phi_i(X_i) \mid Y]\|$;
  replace $\theta(Y)$ with $\theta_1(Y)$;
End Outer Iteration Loop;
  $\theta, \phi_1, \ldots, \phi_p$ are the solutions $\theta^*, \phi_1^*, \ldots, \phi_p^*$;
End ACE Algorithm.

In Section 5, we prove that the ACE algorithm converges to optimal transformations.

3. APPLICATIONS

In the previous section, the ACE algorithm was developed in the context of known distributions. In practice, data distributions are seldom known. Instead, one has a data set $\{(y_k, x_{k1}, \ldots, x_{kp}), 1 \leq k \leq N\}$ that is presumed to be a sample from $Y, X_1, \ldots, X_p$. The goal is to estimate the optimal transformation functions $\theta(Y), \phi_1(X_1), \ldots, \phi_p(X_p)$ from the data. This can be accomplished by applying the ACE algorithm to the data with the quantities e^2, $\| \ \|$, and the conditional expectations replaced by suitable estimates. The resulting functions $\hat{\theta}^*, \hat{\phi}_1^*, \ldots, \hat{\phi}_p^*$ are then taken as estimates of the corresponding optimal transformations.

The estimate for e^2 is the usual mean squared error for regression:

$$e^2(\theta, \phi_1, \ldots, \phi_p) = \frac{1}{N} \sum_{k=1}^{N} \left[\theta(y_k) - \sum_{j=1}^{p} \phi_j(x_{kj})\right]^2.$$

If $g(y, x_1, \ldots, x_p)$ is a function defined for all data values, then $\|g\|^2$ is replaced by

$$\|g\|_N^2 = \frac{1}{N} \sum_{k=1}^{N} g^2(y_k, x_{k1}, \ldots, x_{kp}).$$

For the case of categorical variables, the conditional expectation estimates are straightforward: If the data are $\{(x_k, z_k)\}$, $k = 1, \ldots, N$, and Z is categorical, then

$$\hat{E}[X \mid Z = z] = \sum_{z_k = z} x_k \Big/ \sum_{z_k = z} 1,$$

where X is real-valued and the sums are over the subset of observations having (categorical) value $Z = z$. For variables that can assume many ordered values, the estimation is based

on smoothing techniques. Such procedures have been the subject of considerable study (e.g., see Gasser and Rosenblatt 1979, Cleveland 1979, and Craven and Wahba 1979). Since the smoother is repeatedly applied in the algorithm, high speed is desirable, as well as adaptability to local curvature. We use a smoother employing local linear fits with varying window width determined by local cross-validation (the "supersmoother"; see Friedman and Stuetzle 1982).

The algorithm evaluates $\hat{\theta}^*$, $\hat{\phi}_1^*$, . . . , $\hat{\phi}_p^*$ at all the corresponding data values; that is, $\hat{\theta}^*(y)$ is evaluated at the set of data values $\{y_k\}$, $k = 1, \ldots, N$. The simplest way to understand the shape of the transformations is by means of a plot of the function versus the corresponding data values—that is, through the plots of $\hat{\theta}^*(y_k)$ versus y_k and $\hat{\phi}_1^*$, . . . , $\hat{\phi}_p^*$ versus the data values of x_1, . . . , x_p, respectively.

In this section, we illustrate the ACE procedure by applying it to various data sets. In order to evaluate performance on finite samples, the procedure is first applied to simulated data for which the optimal transformations are known. We next apply it to the Boston housing data of Harrison and Rubinfeld (1978) as listed in Belsley et al. (1980), contrasting the ACE transformations with those used in the original analysis. For our last example, we apply the ACE procedure to a multiple time series to study the relation between air pollution (ozone) and various meteorological quantities.

Our first example consists of 200 bivariate observations $\{(y_k, x_k), 1 \leq k \leq 200\}$ generated from the model

$$y_k = \exp[x_k^3 + \varepsilon_k],$$

with the x_k^3 and the ε_k drawn independently from a standard normal distribution $N(0, 1)$. Figure 1(a) shows a scatterplot of these data. Figures 1(b)–1(d) show the results of applying the ACE algorithm to the data. The estimated optimal transformation $\hat{\theta}^*(y)$ is shown in Figure 1(b)'s plot of $\hat{\theta}^*(y_k)$ versus y_k, $1 \leq k \leq 200$. Figure 1(c) is a plot of $\hat{\phi}^*(x_k)$ versus x_k. These plots suggest the transformations $\theta(y) = \log(y)$ and $\phi(x) = x^3$, which are optimal for the parent distribution. Figure 1(d) is a plot of $\hat{\theta}^*(y_k)$ versus $\hat{\phi}^*(x_k)$. This plot indicates a more linear relation between the transformed variables than that between the untransformed ones.

The next issue we address is how much the algorithm overfits the data due to the repeated smoothings, resulting in inflated estimates of the maximal correlation ρ^* and of $R^{*2} = 1 - e^{*2}$. The answer, on the simulated data sets we have generated, is surprisingly little.

To illustrate this, we contrast two estimates of ρ^* and R^{*2}

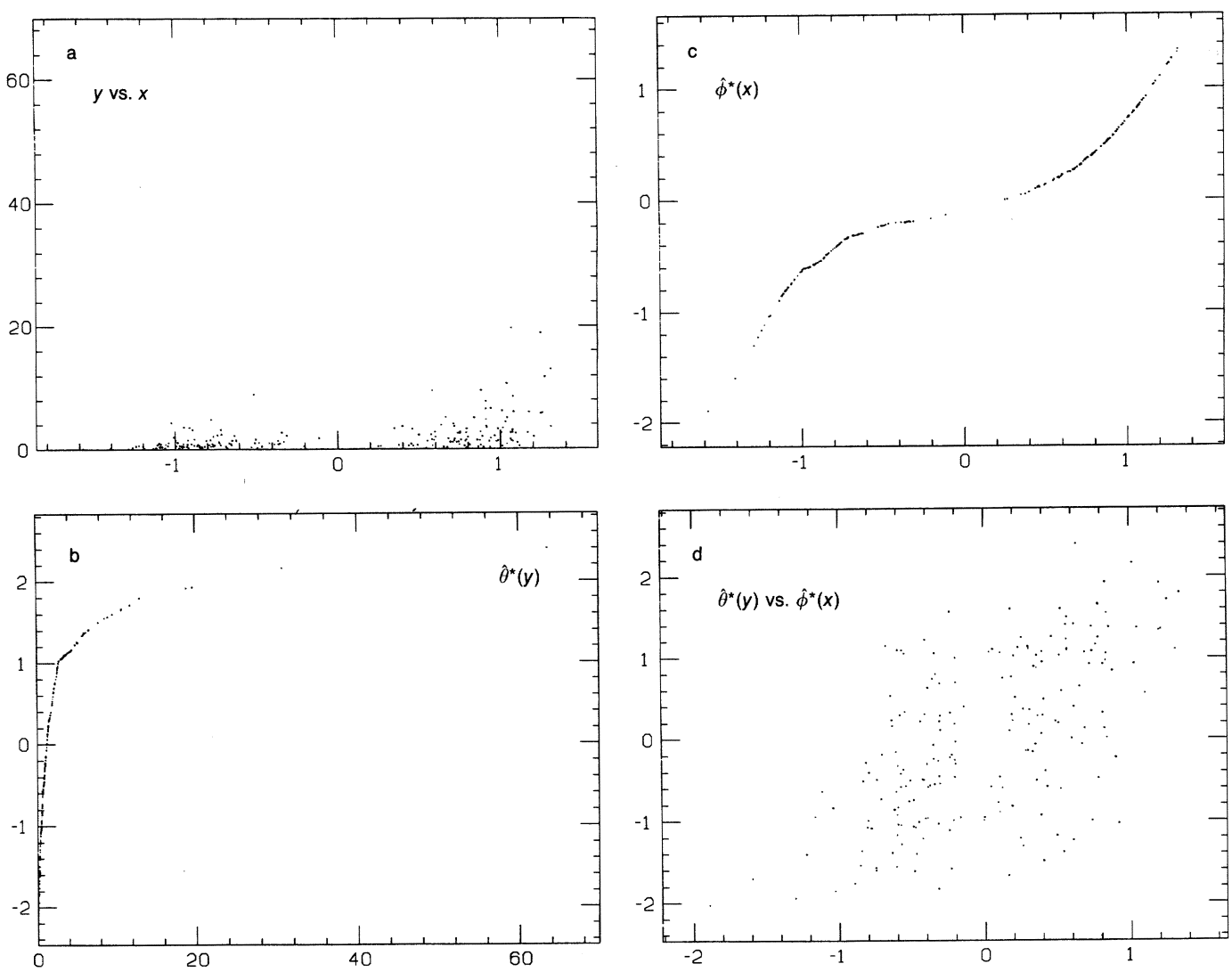

Figure 1. First Example: (a) Original Data; (b) Transform on y; (c) Transform on x; (d) Transformed Data.

Table 1. Comparison of ρ^* Estimates

Estimate	Mean	Standard Deviation
ρ^* direct	.700	.034
ACE	.709	.036

Table 3. Estimate Differences

Estimate	Mean	Standard Deviation
$\hat{\rho}^* - \hat{\rho}$	.001	.015
$\hat{R}^{*2} - \hat{R}^2$	.012	.022

using the above model. The known optimal transformations are $\theta(Y) = \log Y$, $\phi(X) = X^3$. Therefore, we define the *direct* estimate $\hat{\rho}$ for ρ^*, given any data set generated as above by the sample correlation between $\log y_k$ and x_k^3 and set $\hat{R}^2 = \hat{\rho}^2$. The ACE algorithm produces the estimates

$$\hat{\rho}^* = \frac{1}{N}\sum_{k=1}^{N} \hat{\theta}^*(y_k)\hat{\phi}^*(x_k)$$

and $\hat{R}^{*2} = 1 - \hat{e}^{*2} = \hat{\rho}^{*2}$. In this model $\rho^* = .707$ and $R^{*2} = .5$.

For 100 data sets, each of size 200, generated from the above model, the means and standard deviations of the ρ^* estimates are in Table 1. The means and standard deviations of the R^{*2} estimates are in Table 2.

We also computed the differences $\hat{\rho}^* - \hat{\rho}$ and $\hat{R}^{*2} - \hat{R}^2$ for the 100 data sets. The means and standard deviations are in Table 3.

The preceding experiment was duplicated for smaller sample size $N = 100$. In this case we obtained the differences in Table 4.

We next show an application of the procedure to simulated data generated from the model

$$y_k = \exp[\sin(2\pi x_k) + \varepsilon_k/2], \qquad 1 \le k \le 200,$$

with the x_k sampled from a uniform distribution $U(0, 1)$ and the ε_k drawn independently of the x_k from a standard normal distribution $N(0, 1)$. Figure 2(a) shows a scatterplot of these data. Figures 2(b) and 2(c) show the optimal transformation estimates $\hat{\theta}^*(y)$ and $\hat{\phi}^*(x)$. Although $\log(y)$ and $\sin(2\pi x)$ are not the optimal transformations for this model [owing to the non-normal distribution of $\sin(2\pi x)$], these transformations are still clearly suggested by the resulting estimates.

Our next example consists of a sample of 200 triples $\{y_k, x_{k1}, x_{k2})$, $1 \le k \le 200\}$ drawn from the model $Y = X_1X_2$, with X_1 and X_2 generated independently from a uniform distribution $U(-1, 1)$. Note that $\theta(Y) = \log(Y)$ and $\phi_j(X_j) = \log X_j$ $(j = 1, 2)$ cannot be solutions here, since Y, X_1, and X_2 all assume negative values. Figure 3(a) shows a plot of $\hat{\theta}^*(y_k)$ versus y_k, and Figures 3(b) and 3(c) show corresponding plots of $\hat{\phi}_1^*(x_{k1})$ and $\hat{\phi}_2^*(x_{k2})$ $(1 \le k \le 200)$. All three solution transformation functions are seen to be double-valued. The optimal transformations for this problem are $\theta^*(Y) = \log|Y|$ and $\phi_j^*(X_j) = \log|X_j|$ $(j = 1, 2)$. The estimates clearly reflect this structure except near the origin, where the smoother cannot reproduce the infinite discontinuity in the derivative.

This example illustrates that the ACE algorithm is able to produce nonmonotonic estimates for both response and predictor transformations.

For our next example, we apply the ACE algorithm to the Boston housing market data of Harrison and Rubinfeld (1978). A complete listing of these data appears in Belsley et al. (1980). Harrison and Rubinfeld used these data to estimate marginal air pollution damages as revealed in the housing market. Central to their analysis was a housing value equation that relates the median value of owner-occupied homes in each of the 506 census tracts in the Boston Standard Metropolitan Statistical Area to air pollution (as reflected in concentration of nitrogen oxides) and to 12 other variables that are thought to affect housing prices. This equation was estimated by trying to determine the best-fitting functional form of housing price on these 13 variables. By experimenting with a number of possible transformations of the 14 variables (response and 13 predictors), Harrison and Rubinfeld settled on an equation of the form

$$\begin{aligned}\log(MV) = {} & a_1 + a_2(RM)^2 + a_3 AGE \\ & + a_4\log(DIS) + a_5\log(RAD) + a_6 TAX \\ & + a_7 PTRATIO + a_8(B - .63)^2 \\ & + a_9\log(LSTAT) + a_{10} CRIM + a_{11} ZN \\ & + a_{12} INDUS + a_{13} CHAS + a_{14}(NOX)^p + \varepsilon.\end{aligned}$$

A brief description of each variable is given in Appendix B. (For a more complete description, see Harrison and Rubinfeld 1978, table 4.) The coefficients $a_1, \ldots, a_{14}$ were determined by a least squares fit to measurements of the 14 variables for the 506 census tracts. The best value for the exponent p was found to be 2.0, by a numerical optimization (grid search). This "basic equation" was used to generate estimates for the willingness to pay for and the marginal benefits of clean air. Harrison and Rubinfeld (1978) noted that the results are highly sensitive to the particular specification of the form of the housing price equation.

We applied the ACE algorithm to the transformed measurements $(y', x_1' \cdots x_{13}')$ (using $p = 2$ for NOX) appearing in the basic equation. To the extent that these transformations are close to the optimal ones, the algorithm will produce almost linear functions. Departures from linearity indicate transformations that can improve the quality of the fit.

In this (and the following) example we apply the procedure in a forward stepwise manner. For the first pass we consider

Table 2. Comparison of R^{*2} Estimates

Estimate	Mean	Standard Deviation
R^{*2} direct	.492	.047
ACE	.503	.050

Table 4. Estimate Differences, Sample Size 100

Estimate	Mean	Standard Deviation
$\hat{\rho}^* - \hat{\rho}$	.029	.034
$\hat{R}^{*2} - \hat{R}^2$	.042	.051

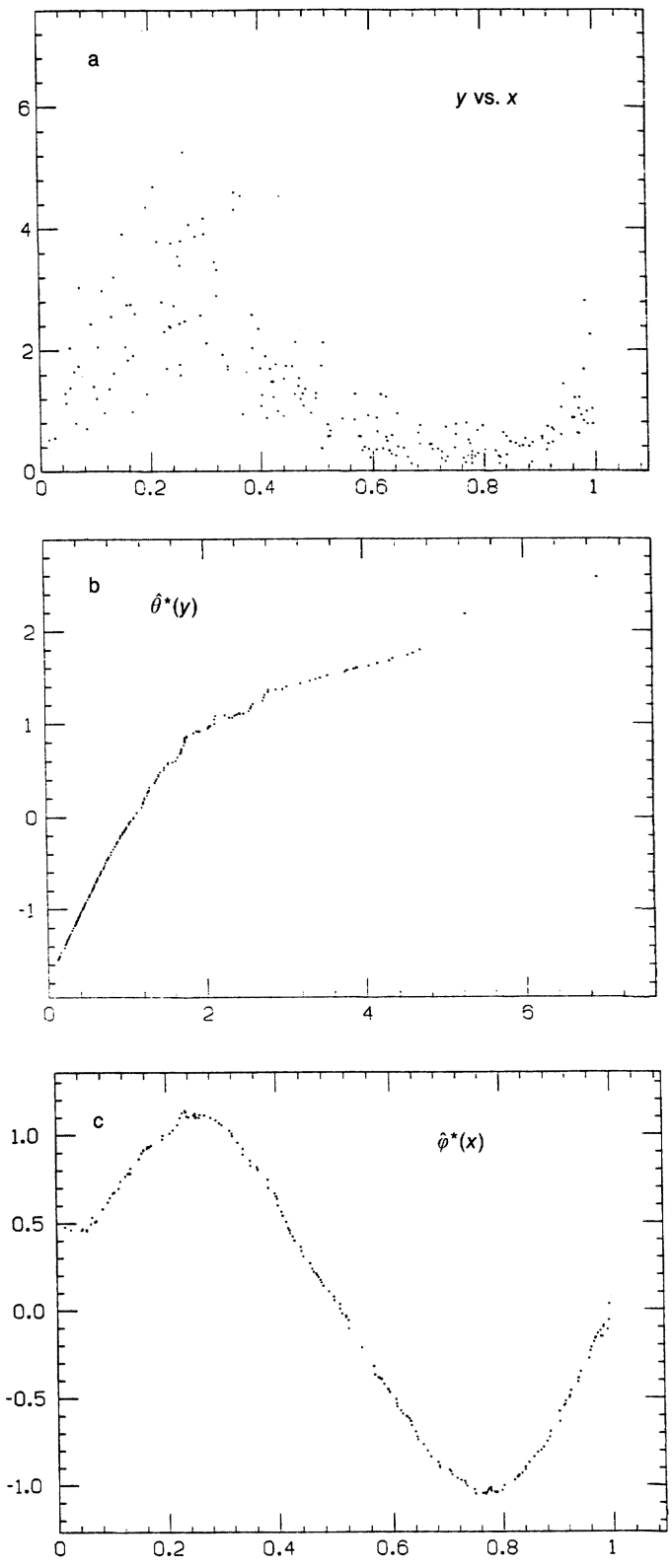

Figure 2. Second Example: (a) Original Data; (b) Transformed y; (c) Transformed x.

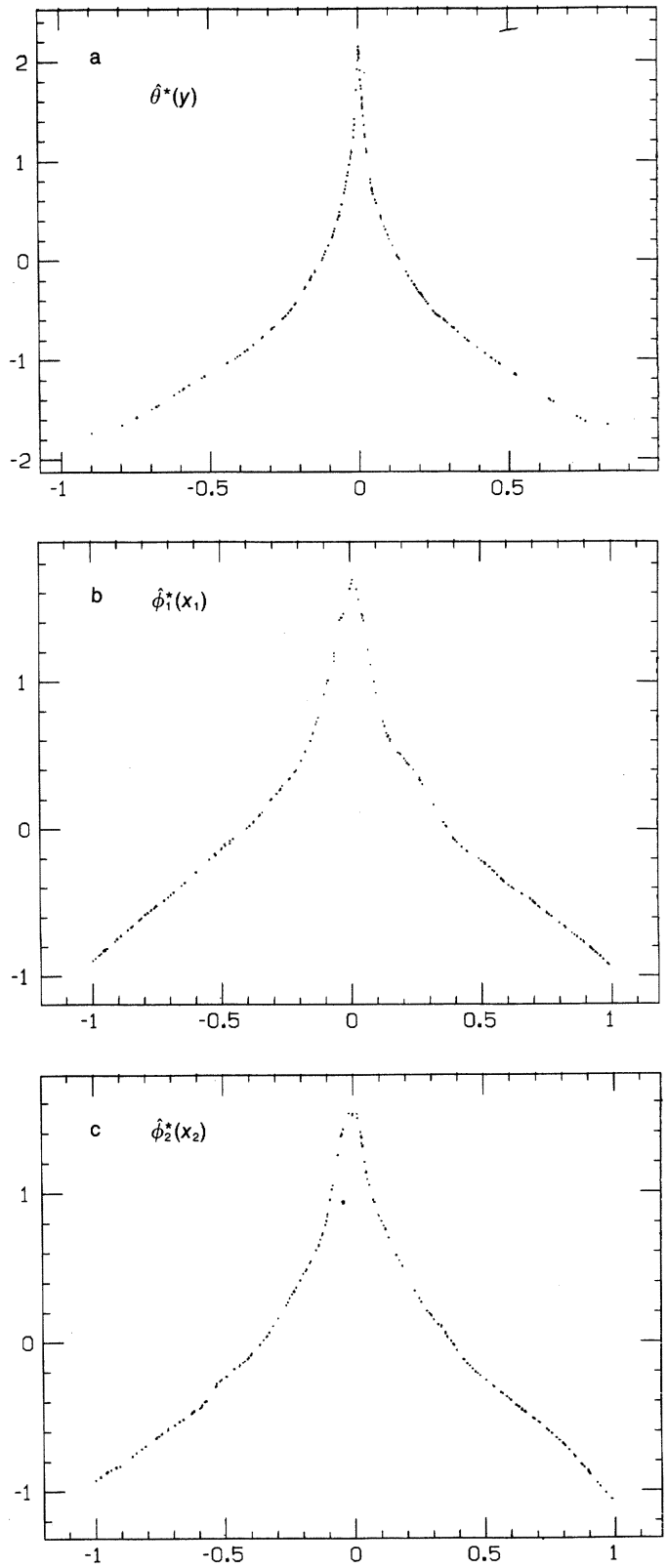

Figure 3. Third Example: (a) Transformed y; (b) Transformed x; (c) Transformed x_2.

the 13 bivariate problems ($p = 1$) involving the response y' with each of the predictor variables x_k' ($1 \leq k \leq 13$) in turn. The predictor k_1 that maximizes $\hat{R}^2[\hat{\theta}_1(y'), \hat{\phi}_{1,k}(x_k')]$ is included in the model. The second pass (over the remaining 12 predictors) includes the 12 trivariate problems ($p = 2$) involving y', x_{k_1}', x_k' ($k \neq k_1$). The predictor that maximizes $\hat{R}^2[\hat{\theta}_2(y'), \hat{\phi}_{2,k_1}(x_{k_1}'), \hat{\phi}_{2,k}(x_k')]$ is included in the model. This forward selection procedure is continued until the best predictor of the next pass increases the $\hat{R}^2$ of the previous pass by less than .01.

The resulting final model involved four predictors and had an $\hat{R}^2$ of .89. Applying ACE simultaneously to all 13 predictors results in an increase in $\hat{R}^2$ of only .02.

Figure 4(a) shows a plot of the solution response transfor-

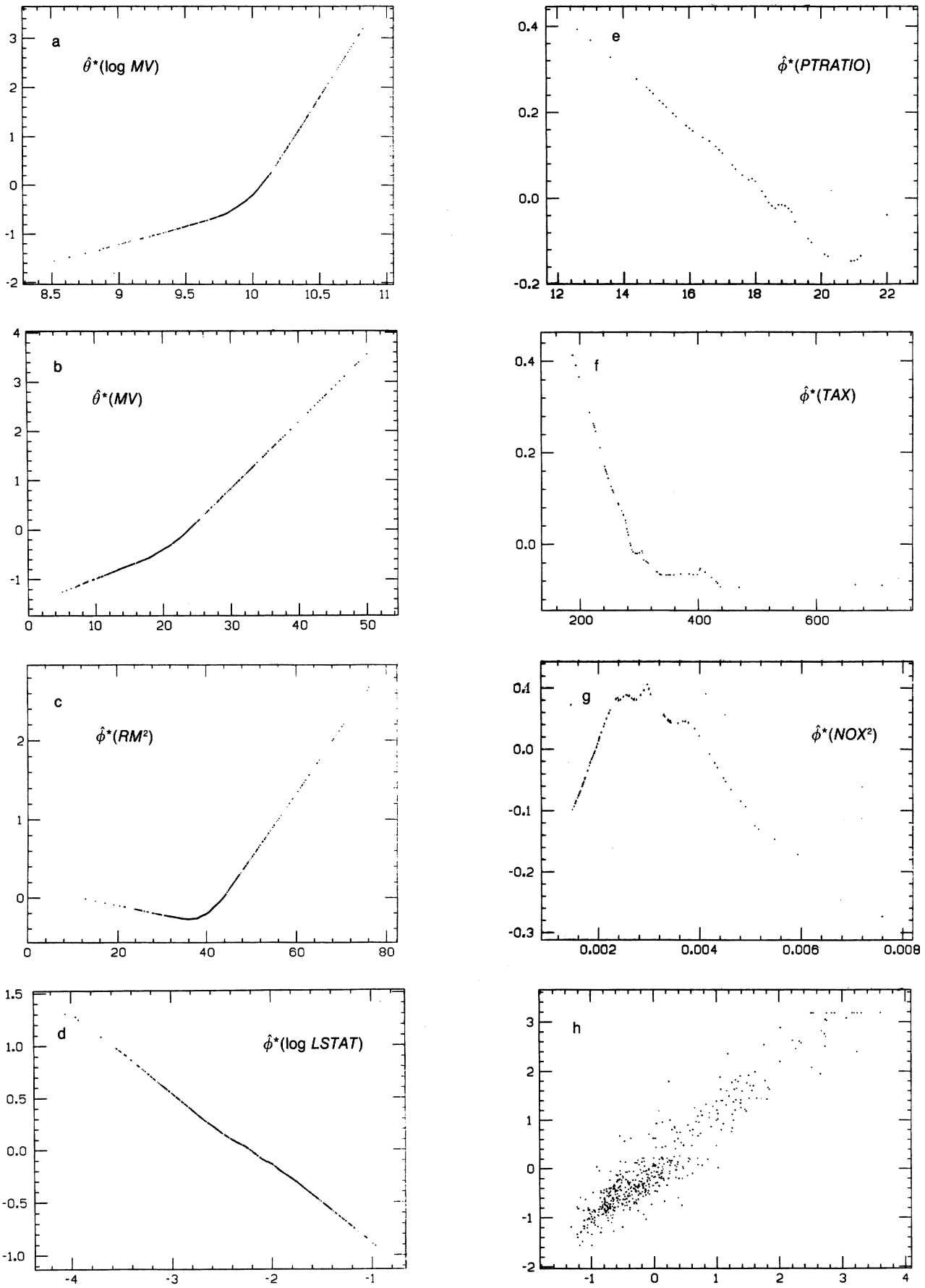

Figure 4. Boston Housing Data: (a) Transformed log(MV); (b) Transformed MV; (c) Transformed RM² (σ = .492); (d) Transformed log(LSTAT) (σ = .417); (e) Transformed PT Ratio (σ = .147); (f) Transformed Tax (σ = .122); (g) Transformed NOX² (σ = .09); (h) Transformed y Versus Predictor of Transformed y.

mation $\hat{\theta}(y')$. This function is seen to have a positive curvature for central values of y', connecting two straight line segments of different slope in either side. This suggests that the logarithmic transformation may be too severe. Figure 4(b) shows the transformation $\hat{\theta}(y)$ resulting when the (forward stepwise) ACE algorithm is applied to the original *untransformed* census measurements. (The same predictor variable set appears in this model.) This analysis indicates that, if anything, a mild transformation, involving *positive* curvature, is most appropriate for the response variable.

Figures 4(c)–4(f) show the ACE transformations $\hat{\phi}_{k_1}(x'_{k_1}) \cdots \hat{\phi}_{k_4}(x'_{k_4})$ for the (transformed) predictor variables x' appearing in the final model. The standard deviation $\sigma(\hat{\phi}_j^*)$ is indicated in each graph. This provides a measure of how strongly each $\hat{\phi}_j^*(x_j)$ enters into the model for $\hat{\theta}^*(y')$. [Note that $\sigma(\hat{\theta}) = 1$.] The two terms that enter most strongly involve the number of rooms squared [Figure 4(c)] and the logarithm of the fraction of population that is of lower status [Figure 4(d)]. The nearly linear shape of the latter transformation suggests that the original logarithmic transformation was appropriate for this variable. The transformation on the number of rooms squared variable is far from linear, however, indicating that a simple quadratic does not adequately capture its relationship to housing value. For fewer than six rooms, housing value is roughly independent of room number, whereas for larger values there is a strong increasing linear dependence. The remaining two variables that enter into this model are pupil-teacher ratio and property tax rate. The solution transformation for the former, Figure 4(e), is seen to be approximately linear whereas that for the latter, Figure 4(f), has considerable nonlinear structure. For tax rates of up to \$320, housing price seems to fall rapidly with increasing tax, whereas for larger rates the association is roughly constant.

Although the variable $(NOX)^2$ was not selected by our stepwise procedure, we can try to estimate its marginal effect on median home value by including it with the four selected variables and running ACE with the resulting five predictor variables. The increase in $\hat{R}^2$ over the four-predictor model was .006. The solution transformations on the response and original four predictors changed very little. The solution transformation for $(NOX)^2$ is shown in Figure 4(g). This curve is a nonmonotonic function of NOX^2, not well approximated by a linear (or monotone) function. This makes it difficult to formulate a simple interpretation of the willingness to pay for clean air from these data. For low concentration values, housing prices seem to increase with increasing $(NOX)^2$, whereas for higher values this trend is substantially reversed.

Figure 4(h) shows a scatterplot of $\hat{\theta}^*(y_k)$ verus $\sum_{j=1}^4 \hat{\phi}_j^*(x_{kj})$ for the four-predictor model. This plot shows no evidence of additional structure not captured in the model

$$\hat{\theta}^*(y) = \sum_{j=1}^{4} \hat{\phi}_j^*(x_j) + e.$$

The $\hat{e}^{*2}$ resulting from the use of the ACE transformations was .11, as compared to the e^2 value of .20 produced by the Harrison and Rubinfeld (1978) transformations involving all 14 variables.

For our final example, we use the ACE algorithm to study the relationship between atmospheric ozone concentration and meteorology in the Los Angeles basin. The data consist of daily measurements of ozone concentration (maximum one hour average) and eight meteorological quantities for 330 days of 1976. Appendix C lists the variables used in the study. The ACE algorithm was applied here in the same forward stepwise manner as in the previous (housing data) example. Four variables were selected. These are the first four listed in Appendix C. The resulting $\hat{R}^2$ was .78. Running the ACE algorithm with all eight predictor variables produces an $\hat{R}^2$ of .79.

In order to assess the extent to which these meteorological variables capture the daily variation of the ozone level, the variable *day-of-the-year* was added and the ACE algorithm was run with it and the four selected meteorological variables. This can detect possible seasonal effects not captured by the meteorological variables. The resulting $\hat{R}^2$ was .82. Figures 5(a)–5(f) show the optimal transformation estimates.

The solution for the response transformation, Figure 5(a), shows that, at most, a very mild transformation with negative curvature is indicated. Similarly, Figure 5(b) indicates that there is no compelling necessity to consider a transformation on the most influential predictor variable, Sandburg Air Force Base Temperature. The solution transformation estimates for the remaining variables, however, are all highly nonlinear (and nonmonotonic). For example, Figure 5(d) suggests that the ozone concentration is much more influenced by the magnitude than the sign of the pressure gradient.

The solution for the day-of-the-year variable, Figure 5(f), indicates a substantial seasonal effect after accounting for the meteorological variables. This effect is minimum at the year boundaries and has a broad maximum peaking at about May 1. This can be compared with the dependence of ozone pollution on day-of-the-year alone, without taking into account the meteorological variables. Figure 5(g) shows a smooth of ozone concentration on day-of-the-year. This smooth has an $\hat{R}^2$ of .38 and is seen to peak about three months later (August 3).

The fact that the day-of-the-year transformation peaked at the beginning of May was initially puzzling to us, since the highest pollution days occur from July to September. This latter fact is confirmed by the day-of-the-year transformation with the meteorological variables removed. Our current belief is that with the meteorological variables entered, day-of-the-year becomes a partial surrogate for hours of daylight before and during the morning commuter rush. The decline past May 1 may then be explained by the fact that daylight saving time goes into effect in Los Angeles on the last Sunday in April.

These data illustrate that ACE is useful in uncovering interesting and suggestive relationships. The form of the dependence on the Daggett pressure gradient and on the day-of-the-year would be extremely difficult to find by any previous methodology.

4. DISCUSSION

The ACE algorithm provides a fully automated method for estimating optimal transformations in multiple regression. It also provides a method for estimating maximal correlation between random variables. It differs from other empirical methods for finding transformations (Box and Tidwell 1962; Anscombe and Tukey 1963; Box and Cox 1964; Kruskal 1964, 1965; Fraser 1967; Box and Hill 1974; Linsey 1972, 1974; Wood

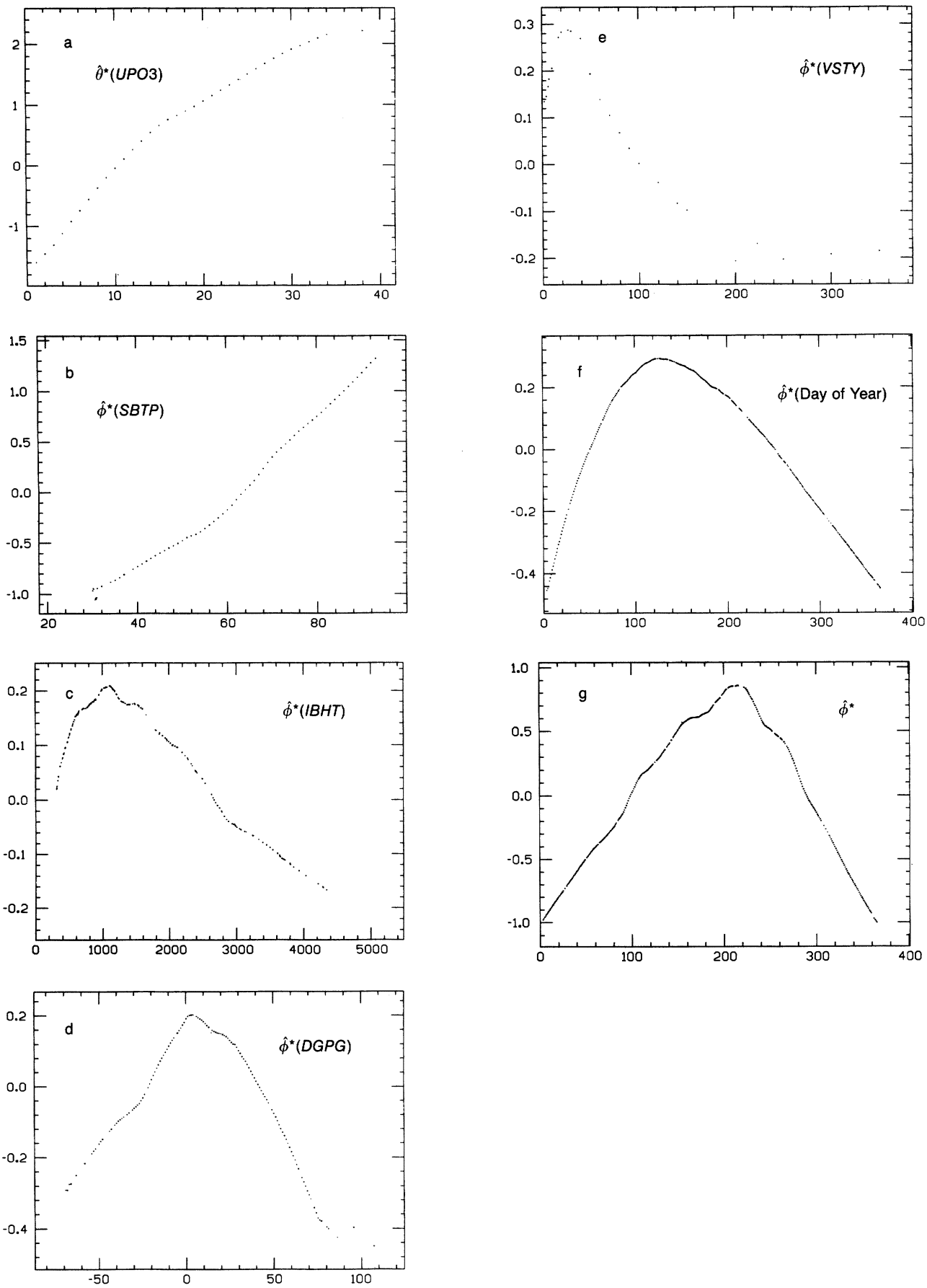

Figure 5. Air Pollution Data: (a) Transformed UPO3 (σ = 1.0); (b) Transformed SBTP (σ = .56); (c) Transformed IBHT (σ = .16); (d) Transformed DGPG (σ = .16); (e) Transformed VSTY (σ = .16); (f) Transformed Day of the Year (σ = .23); (g) Day of the Year as a Single Independent Variable.

1974; Mosteller and Tukey 1977; and Tukey 1982) in that the "best" transformations of the response and predictor variables are unambiguously defined and estimated without use of ad hoc heuristics, restrictive distributional assumptions, or restriction of the transformation to a particular parametric family.

The algorithm is reasonably computer efficient. On the Boston housing data set comprising 506 data points with 14 variables each, the run took 12 seconds of central processing unit (CPU) time on an IBM 3081 computer. Our guess is that this translates into 2.5 minutes on a VAX 11/750 computer. To extrapolate to other problems, use the estimate that running time is proportional to (number of variables) × (sample size).

A strong advantage of the ACE procedure is the ability to incorporate variables of quite different type in terms of the set of values they can assume. The transformation functions $\theta(y)$, $\phi_1(x_1), \ldots, \phi_p(x_p)$ assume values on the real line. Their arguments can, however, assume values on any set. For example, ordered real, periodic (circularly valued) real, ordered, and unordered categorical variables can be incorporated in the same regression equation. For periodic variables, the smoother window need only wrap around the boundaries. For categorical variables, the procedure can be regarded as estimating optimal scores for each of their values. (The special case of a categorical response and a single categorical predictor variable is known as canonical analysis—see Kendall and Stuart 1967, p. 568—and the optimal scores can, in this case, also be obtained by solution of a matrix eigenvector problem.)

The ACE procedure can also handle variables of mixed type. For example, a variable indicating present marital status might take on an integer value (number of years married) or one of several categorical values (N = never, D = divorced, W = widowed, etc.). This presents no additional complication in estimating conditional expectations. This ability provides a straightforward way to handle missing data values (Young et al. 1976). In addition to the regular sets of values realized by a variable, it can also take on the value "missing."

In some situations the analyst, after running ACE, may want to estimate values of y rather than $\theta^*(y)$, given a specific value of $\mathbf{x}$. One method for doing this is to attempt to compute $\hat{\theta}^{*-1}(\sum_{j=1}^p \hat{\phi}_j^*(x_j))$. Letting $Z = \sum_{j=1}^p \hat{\phi}_j^*(X_j)$, however, we know that the best least squares predictor of Y of the form $\chi(Z)$ is given by $E(Y \mid Z)$. This is implemented in the current ACE program by predicting y as the function of $\sum_{j=1}^p \hat{\phi}_j^*(x_j)$, obtained by smoothing the data values of y on the data values of $\sum_{j=1}^p \hat{\phi}_j^*(x_j)$. We are grateful to Arthur Owens for suggesting this simple and elegant prediction procedure.

The solution functions $\hat{\theta}^*(y)$ and $\hat{\phi}_1^*(x_1), \ldots, \hat{\phi}_p^*(x_p)$ can be stored as a set of values associated with each observation $(y_k, x_{k1}, \ldots, x_{kp})$, $1 \le k \le N$. Since $\theta(y)$ and $\phi(x)$, however, are usually smooth (for continuous y, x), they can be easily approximated and stored as cubic spline functions (deBoor 1978) with a few knots.

As a tool for data analysis, the ACE procedure provides graphical output to indicate a need for transformations as well as to guide in their choice. If a particular plot suggests a familiar functional form for a transformation, then the data can be pretransformed using this functional form and the ACE algorithm can be rerun. The linearity (or nonlinearity) of the resulting ACE transformation on the variable in question gives an indication of how good the analyst's guess is. We have found that the plots themselves often give surprising new insights into the relationship between the response and predictor variables.

As with any regression procedure, a high degree of association between predictor variables can sometimes cause the individual transformation estimates to be highly variable, even though the complete model is reasonably stable. When this is suspected, running the algorithm on randomly selected subsets of the data, or on bootstrap samples (Efron 1979), can assist in assessing the variability.

The ACE method has generality beyond that exploited here. An immediate generalization would involve multiple response variables $Y_1, \ldots, Y_q$. The generalized algorithm would estimate optimal transformations $\theta_1^*, \ldots, \theta_q^*, \phi_1^*, \ldots, \phi_p^*$ that minimize

$$E\left[\sum_{l=1}^{q} \theta_l(Y_l) - \sum_{j=1}^{p} \phi_j(X_j)\right]^2$$

subject to $E\theta_l = 0$, $l = 1, \ldots, q$, $E\phi_j = 0$, $j = 1, \ldots, p$, and $\|\sum_{l=1}^q \theta_l(Y_l)\|^2 = 1$.

This extension generalizes the ACE procedure in a sense similar to that in which canonical correlation generalized linear regression.

The ACE algorithm (Section 2) is easily modified to incorporate this extension. An inner loop over the response variables, analogous to that for the predictor variables, replaces the single-function minimization.

5. OPTIMAL TRANSFORMATIONS IN FUNCTION SPACE

5.1 Introduction

In this section, we first prove the existence of optimal transformations (Theorem 5.2). Then we show that the ACE algorithm converges to an optimal transformation (Theorems 5.4 and 5.5).

Define random variables to take values either in the reals or in a finite or countable unordered set. Given a set of random variables $Y, X_1, \ldots, X_p$, a *transformation* is defined by a set of real-valued measurable functions $(\theta, \phi_1, \ldots, \phi_p) = (\theta, \boldsymbol{\phi})$, each function defined on the range of the corresponding random variables, such that

$$E\theta(Y) = 0, \quad E\phi_j(X_j) = 0, \quad j = 1, \ldots, p$$
$$E\theta^2(Y) < \infty, \quad E\phi_j^2(X_j) < \infty, \quad j = 1, \ldots, p. \tag{5.1}$$

Use the notation

$$\tilde{\phi}(\mathbf{X}) = \sum_{j}^{p} \phi_j(X_j). \tag{5.2}$$

Denote the set of all transformations by $\mathcal{F}$.

Definition 5.1. A transformation $(\theta^*, \boldsymbol{\phi}^*)$ is optimal for regression if $E(\theta^*)^2 = 1$ and

$$e^{*2} = E[\theta^*(Y) - \tilde{\phi}^*(\mathbf{X})]^2$$
$$= \inf_{\mathcal{F}} \{E[\theta(Y) - \tilde{\phi}(\mathbf{X})]^2; E\theta^2 = 1\}.$$

Definition 5.2. A transformation $(\theta^{**}, \boldsymbol{\phi}^{**})$ is optimal for

correlation if $E(\theta^{**})^2 = 1$, $E(\tilde{\phi}^{**})^2 = 1$, and

$$\rho^* = E[\theta^{**}(Y)\tilde{\phi}^{**}(\mathbf{X})] = \sup_{\mathcal{F}} \{E[\theta(Y)\tilde{\phi}(\mathbf{X})];\ E(\tilde{\phi})^2 = 1,\ E\theta^2 = 1\}.$$

Theorem 5.1. If $(\theta^{**}, \boldsymbol{\phi}^{**})$ is optimal for correlation, then $\theta^* = \theta^{**}$, $\boldsymbol{\phi}^* = \rho^*\boldsymbol{\phi}^{**}$ is optimal for regression, and the converse. Furthermore, $e^{*2} = 1 - \rho^{*2}$.

Proof. Write

$$E(\theta - \tilde{\phi})^2 = 1 - 2E\theta\tilde{\phi} + E\tilde{\phi}^2 = 1 - 2E(\theta\hat{\phi})\sqrt{E\tilde{\phi}^2} + E\tilde{\phi}^2,$$

where $\hat{\phi} = \tilde{\phi}/\sqrt{E\tilde{\phi}^2}$. Hence

$$E(\theta - \tilde{\phi})^2 \geq 1 - 2\rho^*\sqrt{E\tilde{\phi}^2} + E\tilde{\phi}^2 \qquad (5.3)$$

with equality only if $E\theta\hat{\phi} = \rho^*$. The minimum of the right side of (5.3) over $E\tilde{\phi}^2$ is at $E\tilde{\phi}^2 = (\rho^*)^2$, where it is equal to $1 - (\rho^*)^2$. Then $(e^*)^2 = 1 - (\rho^*)^2$; and if $(\theta^{**}, \boldsymbol{\phi}^{**})$ is optimal for correlation, then $\theta^* = \theta^{**}$, $\boldsymbol{\phi}^* = \rho^*\boldsymbol{\phi}^{**}$ is optimal for regression. The argument is reversible. (A similar result appears in Csáki and Fisher 1963.)

5.2 Existence of Optimal Transformations

To show existence of optimal transformations, two additional assumptions are needed.

Assumption 5.1. The only set of functions satisfying (5.1) such that

$$\theta(Y) + \sum_j \phi_j(X_j) = 0 \text{ a.s.}$$

are individually a.s. zero.

To formulate the second assumption, we use Definition 5.3.

Definition 5.3. Define the Hilbert spaces $H_2(Y)$, $H_2(X_1)$, . . . , $H_2(X_p)$ as the sets of functions satisfying (5.1) with the usual inner product; that is, $H_2(X_j)$ is the set of all measurable ϕ_j such that $E\phi_j(X_j) = 0$, $E\phi_j^2(X_j) < \infty$ with $(\phi_j', \phi_j) = E[\phi_j'(X_j)\phi_j(X_j)]$.

Assumption 5.2. The conditional expectation operators

$$E(\phi_j(X_j) \mid Y) : H_2(X_j) \to H_2(Y),$$
$$E(\phi_j(X_j) \mid X_i) : H_2(X_j) \to H_2(X_i), \qquad i \neq j$$
$$E(\theta(Y) \mid X_j) : H_2(Y) \to H_2(X_j)$$

are all compact.

Assumption 5.2 is satisfied in most cases of interest. A sufficient condition is given by the following. Let X, Y be random variables with joint density $f_{X,Y}$ and marginals f_X, f_Y. Then the conditional expectation operator on $H_2(Y) \to H_2(X)$ is compact if

$$\int\int [f_{XY}^2/f_X f_Y]dxdy < \infty. \qquad (5.4)$$

Theorem 5.2. Under Assumptions 5.1 and 5.2, optimal transformations exist.

Some machinery is needed.

Proposition 5.1. The set of all functions f of the form

$$f(Y, \mathbf{X}) = \theta(Y) + \sum_j \phi_j(X_j), \qquad \theta \in H_2(Y),\ \phi_j \in H_2(X_j),$$

with the inner product and norm

$$(g, f) = E[gf], \qquad \|f\|^2 = Ef^2,$$

is a Hilbert space denoted by H_2. The subspace of all functions $\tilde{\phi}$ of the form

$$\tilde{\phi}(\mathbf{X}) = \sum_1^p \phi_j(X_j), \qquad \phi_j \in H_2(X_j),$$

is a closed linear subspace denoted by $H_2(X)$. So are $H_2(Y)$, $H_2(X_1)$, . . . , $H_2(X_p)$.

Proposition 5.1 follows from Proposition 5.2.

Proposition 5.2. Under Assumptions 5.1 and 5.2, there are constants $0 < c_1 \leq c_2 < \infty$ such that

$$c_1\left(\|\theta\|^2 + \sum_1^p \|\phi_j\|^2\right) \leq \left\|\theta + \sum_1^p \phi_j\right\|^2 \leq c_2\left(\|\theta\|^2 + \sum_1^p \|\phi_j\|^2\right).$$

Proof. The right-hand inequality is immediate. If the left side does not hold, we can find a sequence $f_n = \theta_n + \Sigma\, \phi_{n,j}$ such that $\|\theta_n\|^2 + \Sigma_1^p \|\phi_{nj}\|^2 = 1$, but $\|f_n\|^2 \to 0$. There is a subsequence n' such that $\theta_{n'} \xrightarrow{w} \theta$, $\phi_{n'j} \xrightarrow{w} \phi_j$ in the sense of weak convergence in $H_2(Y)$, $H_2(X_1)$, . . . , $H_2(X_p)$, respectively.

Write

$$E[\phi_{n'j}(X_j)\phi_{n'i}(X_i)] = E[\phi_{n'j}(X_j)E(\phi_{n'i}(X_i) \mid X_j)]$$

to see that Assumption 5.2 implies $E\phi_{n'j}\phi_{n'i} \to E\phi_j\phi_i$ $(i \neq j)$, and similarly for $E\theta_{n'}\phi_{n',j}$. Furthermore $\|\phi_i\| \leq \liminf \|\phi_{n'i}\|$, $\|\theta\| \leq \liminf \|\theta_{n'}\|$. Thus, defining $f = \theta + \Sigma_j\phi_j$,

$$\|f\|^2 = \|\theta + \sum_j \phi_j\|^2 \leq \liminf \|f_{n'}\|^2 = 0,$$

which implies, by Assumption 5.1, that $\theta = \phi_1 = \cdots = \phi_p = 0$. On the other hand,

$$\|f_{n'}\|^2 = \|\theta_{n'}\|^2 + \sum_j \|\phi_{n'j}\|^2 + 2\sum_j (\theta_{n'}, \phi_{n',j}) + 2\sum_{i\neq j} (\phi_{n'j}, \phi_{n',i}).$$

Hence, if $f = 0$, then $\liminf \|f_n\|^2 \geq 1$.

Corollary 5.1. If $f_n \xrightarrow{w} f$ in H_2, then $\theta_n \to \theta$ in $H_2(Y)$, $\phi_{nj} \xrightarrow{w} \phi_j$ in $H_2(X_j)$, $j = 1, \ldots, p$, and the converse.

Proof. If $f_n = \theta_n + \Sigma_j\, \phi_{nj} \xrightarrow{w} \theta + \Sigma_j\, \phi_j$, then by Proposition 5.2, $\limsup \|\theta_n\| < \infty$, $\limsup \|\phi_{nj}\| < \infty$. Take n' such that $\theta_{n'} \xrightarrow{w} \theta'$, $\phi_{n'j} \xrightarrow{w} \phi_j'$, and let $f' = \theta' + \Sigma_j\, \phi_j'$. Then for any $g \in H_2$, $(g, f_{n'}) \xrightarrow{w} (g, f')$, so $(g, f) = (g, f')$ all g. The converse is easier.

Definition 5.4. In H_2, let P_Y, P_j, and P_X denote the projection operators into $H_2(Y)$, $H_2(X_j)$, and $H_2(X)$, respectively.

On $H_2(X_i)$, P_j $(j \neq i)$ is the conditional expectation operator, and similarly for P_Y.

Proposition 5.3. P_Y is compact on $H_2(X) \to H_2(Y)$, and P_X is compact on $H_2(Y) \to H_2(X)$.

Proof. Take $\tilde{\phi}_n \in H_2(X)$, $\tilde{\phi}_n \xrightarrow{w} \tilde{\phi}$. This implies, by Corollary 5.1, that $\phi_{n,j} \xrightarrow{w} \phi_j$. By Assumption 5.2, $P_Y\phi_{n,j} \xrightarrow{s} P_Y\phi_j$ so that $P_Y\tilde{\phi}_n \xrightarrow{s} P_Y\tilde{\phi}$. Now take $\theta \in H_2(Y)$, $\tilde{\phi} \in H_2(X)$; then $(\theta, P_Y\tilde{\phi}) = (\theta, \tilde{\phi}) = (P_X\theta, \tilde{\phi})$. Thus $P_X : H_2(Y) \to H_2(X)$ is the adjoint of P_Y and hence compact.

Now to complete the proof of Theorem 5.2, consider the functional $\|\theta - \tilde{\phi}\|^2$ on the set of all $(\theta, \tilde{\phi})$ with $\|\theta\|^2 = 1$. For any θ, $\tilde{\phi}$,

$$\|\theta - \tilde{\phi}\|^2 \geq \|\theta - P_X\theta\|^2.$$

If there is a θ^* that achieves the minimum of $\|\theta - P_X\theta\|^2$ over $\|\theta\|^2 = 1$, then an optimal transformation is θ^*, $P_X\theta^*$. On $\|\theta\|^2 = 1$,

$$\|\theta - P_X\theta\|^2 = 1 - \|P_X\theta\|^2.$$

Let $\bar{s} = \{\sup\|P_X\theta\|; \|\theta\| = 1\}$. Take θ_n such that $\|\theta_n\|^2 = 1$, $\theta_n \xrightarrow{w} \theta$, and $\|P_X\theta_n\| \to \bar{s}$. By the compactness of P_X, $\|P_X\theta_n\| \to \|P_X\theta\| = \bar{s}$. Furthermore, $\|\theta\| \leq 1$. If $\|\theta\| < 1$, then for $\theta' = \theta/\|\theta\|$, we get the contradiction $\|P_X\theta'\| > \bar{s}$. Hence $\|\theta\| = 1$ and $(\theta, P_X\theta)$ is an optimal transformation. This argument assumes that $\bar{s} > 0$. If $\bar{s} = 0$, then $\|\theta - P_X\theta\| = 1$ for all θ with $\|\theta\| = 1$, and any $(\theta, 0)$ is optimal.

5.3 Characterization of Optimal Transformations

Define two operators, $U : H_2(Y) \to H_2(Y)$ and $V : H_2(X) \to H_2(X)$, by

$$U\theta = P_YP_X\theta, \qquad V\tilde{\phi} = P_XP_Y\tilde{\phi}.$$

Proposition 5.4. U and V are compact, self-adjoint, and non-negative definite. They have the same eigenvalues, and there is a 1–1 correspondence between eigenspaces for a given positive eigenvalue specified by

$$\tilde{\phi} = P_X\theta/\|P_X\theta\|, \qquad \theta = P_Y\tilde{\phi}/\|P_Y\tilde{\phi}\|.$$

Proof. Direct verification.

Let the largest eigenvalue be denoted by $\bar{\lambda}$, $\bar{\lambda} = \|U\| = \|V\|$. In the sequel we add the assumption that there is at least one $\theta(Y)$ such that $\|P_X\theta\| > 0$. Then $\bar{\lambda} > 0$ and Theorem 5.3 follows.

Theorem 5.3. If θ^*, $\tilde{\phi}^*$ is an optimal transformation for regression, then

$$\bar{\lambda}\theta^* = U\theta^*, \qquad \bar{\lambda}\tilde{\phi}^* = V\tilde{\phi}^*.$$

Conversely, if θ satisfies $\bar{\lambda}\theta = U\theta$, $\|\theta\| = 1$, then θ, $P_X\theta$ is optimal for regression. If $\tilde{\phi}$ satisfies $\bar{\lambda}\tilde{\phi} = V\tilde{\phi}$, then $\theta = P_Y\tilde{\phi}/\|P_Y\tilde{\phi}\|$, and $\bar{\lambda}\tilde{\phi}/\|P_Y\tilde{\phi}\|$ are optimal for regression. In addition,

$$(e^2) = 1 - \bar{\lambda}.$$

Proof. Let θ^*, $\tilde{\phi}^*$ be optimal. Then $\tilde{\phi}^* = P_X\theta^*$. Write

$$\|\theta^* - \tilde{\phi}^*\|^2 = 1 - 2(\theta^*, \tilde{\phi}^*) + \|\tilde{\phi}^*\|^2.$$

Note that $(\theta^*, \tilde{\phi}^*) = (\theta^*, P_Y\tilde{\phi}^*) \leq \|P_Y\tilde{\phi}^*\|$ with equality only if $\theta^* = cP_Y\tilde{\phi}^*$, c constant. Therefore, $\theta^* = P_Y\tilde{\phi}^*/\|P_Y\tilde{\phi}^*\|$. This implies

$$\|P_Y\tilde{\phi}^*\|\theta^* = U\theta^*, \qquad \|P_Y\tilde{\phi}^*\|\tilde{\phi}^* = V\tilde{\phi}^*,$$

so that $\|P_Y\tilde{\phi}^*\|$ is an eigenvalue λ^* of U, V. Computing gives $\|\theta^* - \tilde{\phi}^*\|^2 = 1 - \lambda^*$. Now take θ any eigenfunction of U corresponding to $\bar{\lambda}$, with $\|\theta\| = 1$. Let $\tilde{\phi} = P_X\theta$; then $\|\theta - \tilde{\phi}\|^2 = 1 - \bar{\lambda}$. This shows that θ^*, $\tilde{\phi}^*$ are not optimal unless $\lambda^* = \bar{\lambda}$. The rest of the theorem is straightforward verification.

Corollary 5.2. If $\bar{\lambda}$ has multiplicity one, then the optimal transformation is unique up to a sign change. In any case, the set of optimal transformations is finite dimensional.

5.4 Alternating Conditional Methods

Direct solution of the equations $\bar{\lambda}\theta = U\theta$ or $\bar{\lambda}\tilde{\phi} = V\tilde{\phi}$ is formidable. Attempting to use data to directly estimate the solutions is just as difficult. In the bivariate case, if X, Y are categorical, then $\tilde{\phi}\theta = U\theta$ becomes a matrix eigenvalue problem and is tractable. This is the case treated in Kendall and Stuart (1967).

The ACE algorithm is founded on the observation that there is an iterative method for finding optimal transformations. We illustrate this in the bivariate case. The goal is to minimize $\|\theta(Y) - \phi(X)\|^2$ with $\|\theta\|^2 = 1$. Denote $P_X\theta = E(\theta \mid X)$, $P_Y\phi = E(\phi \mid Y)$. Start with any first-guess function $\theta_0(Y)$ having a nonzero projection on the eigenspace of the largest eigenvalue of U. Then define a sequence of functions by

$$\phi_0 = P_X\theta_0$$

$$\theta_1 = P_Y\phi_0/\|P_Y\phi_0\|$$

$$\phi_1 = P_X\theta_1,$$

and in general $\phi_{n+1} = P_X\theta_n$, $\theta_{n+1} = P_Y\phi_{n+1}/\|P_Y\theta_{n+1}\|$. It is clear that at each step in the iteration $\|\theta - \phi\|^2$ is decreased. It is not hard to show that in general, θ_n, ϕ_n converge to an optimal transformation.

The preceding method of alternating conditionals extends to the general multivariate case. The analog is clear; given θ_n, $\tilde{\phi}_n$, the next iteration is

$$\tilde{\phi}_{n+1} = P_X\theta_n, \qquad \theta_{n+1} = P_Y\tilde{\phi}_{n+1}/\|P_Y\tilde{\phi}_{n+1}\|.$$

However, there is an additional issue: How can $P_X\theta$ be computed using only the conditional expectation operators P_j $(j = 1, \ldots, p)$? This is done by starting with some function $\tilde{\phi}_0$ and iteratively subtracting off the projections of $\theta - \tilde{\phi}_n$ on the subspaces $H_2(X_1), \ldots, H_2(X_p)$ until we get a function $\tilde{\phi}$ such that the projection of $\theta - \tilde{\phi}$ on each of $H_2(X_j)$ is zero. This leads to the *double-loop algorithm*.

The Double-Loop Algorithm

The Outer Loop. (a) Start with an initial guess $\theta_0(Y)$. (b) Put $\tilde{\phi}_{n+1} = P_X\theta_n$, $\theta_{n+1} = P_Y\tilde{\phi}_{n+1}/\|P_Y\tilde{\phi}_{n+1}\|$ and repeat until convergence.

Let $P_E\theta_0$ be the projection of θ_0 on the eigenspace E of U corresponding to $\bar{\lambda}$.

Theorem 5.4. If $\|P_E\theta_0\| \neq 0$, define an optimal transformation by $\theta^* = P_E\theta_0/\|P_E\theta_0\|$, $\tilde{\phi}^* = P_X\theta_0^*$. Then $\|\theta_n - \theta^*\| \to 0$, $\|\tilde{\phi}_n - \tilde{\phi}^*\| \to 0$.

Proof. Notice that $\theta_{n+1} = U\theta_n/\|U\theta_n\|$. For any n, $\theta_n = a_n\theta^* + g_n$, where $g_n \perp E$, because, if it is true for n, then

$$\theta_{n+1} = (a_n\bar{\lambda}\theta^* + Ug_n)/\|a_n\bar{\lambda}\theta^* + Ug_n\|$$

and Ug_n is $\perp$ to E. For any $g \perp E$, $\|Ug\| \le r\|g\|$, where $r < \bar{\lambda}$. Since $a_{n+1} = \bar{\lambda}a_n/\|U\theta_n\|$, $g_{n+1} = Ug_n/\|U\theta_n\|$; then

$$\|g_{n+1}\|/a_{n+1} = \|Ug_n\|/\bar{\lambda}a_n \le (r/\bar{\lambda})\|g_n\|/a_n.$$

Thus $\|g_n\|/a_n \le c(r/\bar{\lambda})^n$. But $\|\theta_n\| = 1$, $a_n^2 + \|g_n\|^2 = 1$, implying $a_n^2 \to 1$. Since $a_0 > 0$, then $a_n > 0$; so $a_n \to 1$. Now use $\|\theta_n - \theta^*\|^2 = (1 - a_n)^2 + \|g_n\|^2$ to reach the conclusion. Since $\|\tilde{\phi}_{n+1} - \tilde{\phi}^*\| = \|P_X\theta_n - P_X\theta^*\| \le \|\theta_n - \theta^*\|$, the theorem follows.

The Inner Loop. (a) Start with functions θ, $\tilde{\phi}_0$. (b) If, after m stages of iteration, the functions are $\phi_j^{(m)}$, then define, for $j = 1, 2, \ldots, p$,

$$\phi_j^{(m+1)} = P_j\left(\theta - \sum_{i>j}\phi_j^{(m)} - \sum_{i<j}\phi_i^{(m+1)}\right).$$

Theorem 5.5. Let $\tilde{\phi}_m = \sum_j \phi_j^{(m)}$. Then $\|P_X\theta - \tilde{\phi}_m\| \to 0$.

Proof. Define the operator T by

$$T = (I - P_p)(I - P_{p-1}) \cdots (I - P_1).$$

Then the iteration in the inner loop is expressed as

$$\begin{aligned}\theta - \tilde{\phi}_{m+1} &= T(\theta - \tilde{\phi}_m) \\ &= T^{m+1}(\theta - \tilde{\phi}_0). \qquad (5.5)\end{aligned}$$

Write $\theta - \tilde{\phi}_0 = \theta - P_X\theta + P_X\theta - \tilde{\phi}_0$. Noting that $T(\theta - P_X\theta) = \theta - P_X\theta$, (5.5) becomes

$$\tilde{\phi}_{m+1} = P_X\theta - T^{m+1}(P_X\theta - \tilde{\phi}_0).$$

The theorem is then proven by Proposition 5.5.

Proposition 5.5. For any $\tilde{\phi} \in H_2(X)$, $\|T^m\tilde{\phi}\| \to 0$.

Proof. $\|(I - P_j)\tilde{\phi}\|^2 = \|\tilde{\phi}\|^2 - \|P_j\tilde{\phi}\|^2 \le \|\tilde{\phi}\|^2$. Thus $\|T\| \le 1$. There is no $\tilde{\phi} \ne 0$ such that $\|T\tilde{\phi}\| = \|\tilde{\phi}\|$. If there were, then $\|P_j\tilde{\phi}\| = 0$, all j. Then for $\tilde{\phi}' = \sum \tilde{\phi}_j'$,

$$(\tilde{\phi}, \tilde{\phi}') = \sum_j (\tilde{\phi}, \tilde{\phi}_j') = \sum_j (P_j\tilde{\phi}, \tilde{\phi}_j') = 0.$$

The operator T can be decomposed as $I + W$, where W is compact. Now we claim that $\|T^mW\| \to 0$ on $H_2(X)$. To prove this, let $\gamma > 0$ and define

$$G(\gamma) = \sup_{\tilde{\phi}} \{\|TW\tilde{\phi}\|/\|W\tilde{\phi}\|;\ \|\tilde{\phi}\| \le 1,\ \|W\tilde{\phi}\| \ge \gamma\}.$$

Take $\tilde{\phi}_n \xrightarrow{w} \tilde{\phi}$, $\|\tilde{\phi}_n\| \le 1$, $\|W\tilde{\phi}_n\| \ge \gamma$ so that $\|TW\tilde{\phi}_n\|/\|W\tilde{\phi}_n\| \to G(\gamma)$. Then $\|\tilde{\phi}\| \le 1$, $\|W\tilde{\phi}\| \ge \gamma$, and $G(\gamma) = \|TW\tilde{\phi}\|/\|W\tilde{\phi}\|$. Thus $G(\gamma) < 1$ for all $\gamma > 0$ and is clearly nonincreasing in γ. Then

$$\|T^mW\tilde{\phi}\| = \|TWT^{m-1}\tilde{\phi}\| \le G(\|T^{m-1}W\tilde{\phi}\|)\|T^{m-1}W\tilde{\phi}\|.$$

Put $\gamma_0 = \|W\|$, $\gamma_m = G^m(\gamma_m)\gamma_0$; then $\|T^mW\| \le \gamma_m$. But clearly $\gamma_m \to 0$.

The range of W is dense in $H_2(X)$. Otherwise, there is a $\tilde{\phi}' \ne 0$ such that $(\tilde{\phi}', W\tilde{\phi}) = 0$, all $\tilde{\phi}$. This implies $(W^*\tilde{\phi}', \tilde{\phi}) = 0$ or $W^*\tilde{\phi}' = 0$. Then $\|T^*\tilde{\phi}'\| = \|\tilde{\phi}'\|$, and a repetition of the argument given before leads to $\tilde{\phi}' = 0$. For any $\tilde{\phi}$ and $\varepsilon > 0$, take $W\tilde{\phi}_1$ so that $\|\tilde{\phi} - W\tilde{\phi}_1\| \le \varepsilon$. Then $\|T^m\tilde{\phi}\| \le \varepsilon + \|T^mW\tilde{\phi}_1\|$, which completes the proof.

There are two versions of the double loop. In the first, the initial functions $\tilde{\phi}_0$ are the limiting functions produced by the preceding inner loop. This is called the *restart* version. In the second, the initial functions are $\tilde{\phi}_0 \equiv 0$. This is the *fresh start* version. The main theoretical difference is that a stronger consistency result holds for the fresh start. Restart is a faster-running algorithm, and it is embodied in the ACE code.

The Single-Loop Algorithm

The original implementation of ACE combined a single iteration of the inner loop with an iteration of the outer loop. Thus it is summarized by the following.

1. Start with θ_0, $\tilde{\phi}_0 = 0$.
2. If the current functions are θ_n, $\tilde{\phi}_n$, define $\tilde{\phi}_{n+1}$ by
$$\theta_n - \tilde{\phi}_{n+1} = T(\theta_n - \tilde{\phi}_n).$$
3. Let $\theta_{n+1} = P_Y\tilde{\phi}_{n+1}/\|P_Y\tilde{\phi}_{n+1}\|$. Run to convergence.

This is a cleaner algorithm than the double loop, and its implementation on data runs at least twice as fast as the double loop and requires only a single convergence test. Unfortunately, we have been unable to prove that it converges in function space. Assuming convergence, it can be shown that the limiting θ is an eigenfunction of U. But giving conditions for θ to correspond to $\bar{\lambda}$, or even showing that θ will correspond to $\bar{\lambda}$, "almost always" seems difficult. For this reason, we adopted the double-loop algorithm instead.

APPENDIX A: THE ACE ALGORITHM ON FINITE DATA SETS

A.1 Introduction

The ACE algorithm is implemented on finite data sets by replacing conditional expectations, given continuous variables, by data smooths. In the theoretical results concerning the convergence and consistency properties of the ACE algorithm, the critical element is the properties of the data smooth used. The results are fragmentary. Convergence of the algorithm is proven only for a restricted class of smooths. In practice, in more than 1,000 runs of ACE on a wide variety of data sets and using three different types of smooths, we have seen only one instance of failure to converge. A fairly general, but weak, consistency proof is given. We conjecture the form of a stronger consistency result.

A.2 Data Smooths

Define a data set D to be a set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of N points in p-dimensional space; that is, $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})$. Let $\mathfrak{D}_N$ be the collection of all such data sets. For fixed D, define $F(\mathbf{x})$ as the space of all real-valued functions ϕ defined on D; that is, $\phi \in F(\mathbf{x})$ is defined by the N real numbers $\{\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_N)\}$. Define $F(x_j)$, $j = 1, \ldots, p$, as the space of all real-valued functions defined on the set $\{x_{1j}, x_{2j}, \ldots, x_{Nj}\}$.

Definition A.1. A data smooth S of $\mathbf{x}$ on x_j is a mapping $S : F(\mathbf{x}) \to F(x_j)$ defined for every D in $\mathfrak{D}_N$. If $\phi \in F(\mathbf{x})$, denote the corresponding element in $F(x_j)$ by $S(\phi \mid x_j)$ and its values by $S(\phi \mid x_{kj})$.

Let x be any one of $x_1, \ldots, x_p$. Some examples of data smooths are the following.

1. *Histogram.* Divide the real axis into disjoint intervals $\{I_l\}$. If $x_k \in I_l$, define

$$S(\phi \mid x_k) = \frac{1}{n_l} \sum_{\mathbf{x}_m : x_k \in I_l} \phi(\mathbf{x}_m).$$

2. *Nearest Neighbor.* Fix $M < N/2$. Order the x_k getting $x_1 < x_2 < \cdots < x_N$ (assume no ties) and corresponding $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)$. Put

$$S(\phi \mid x_k) = \frac{1}{2M} \sum_{\substack{m=-M \\ m \neq 0}}^{M} \phi(\mathbf{x}_{k+m}).$$

If M points are not available on one side, make up the deficiency on the other side.

3. *Kernel.* Take $K(x)$ defined on the reals with maximum at $x = 0$. Then

$$S(\phi \mid x_k) = \sum_m \phi(\mathbf{x}_m) K(x_m - x_k) \Big/ \sum_l K(x_l - x_k).$$

4. *Regression.* Fix M and order x_k as in example 2. At x_k, regress the values of $\phi(\mathbf{x}_{k+M}), \ldots, \phi(\mathbf{x}_{k+M})$, excluding $\phi(\mathbf{x}_k)$, on $x_{k-M}, \ldots, x_{k+M}$, excluding x_k, getting a regression line $L(x)$. Put $S(\phi \mid x_k) = L(x_k)$. If M points are not available on each side of x_k, make up the deficiency on the other side.

5. *Supersmoother.* See Friedman and Stuetzle (1982).

Some properties that are relevant to the behavior of smoothers are given next. These properties hold only if they are true for all $D \in \mathfrak{D}_n$.

1. *Linearity.* A smooth is linear if

$$S(a\phi_1 + \beta\phi_2) = aS\phi_1 + \beta S\phi_2$$

for all $\phi_1, \phi_2 \in F(\mathbf{x})$ and all constants a, β.

2. *Constant Preserving.* If $\phi \in F(\mathbf{x})$ is constant ($\phi \equiv c$), then $S\phi \equiv c$.

To give a further property, introduce the inner product $(\ )_N$ on $F(\mathbf{x})$ defined by

$$(\phi, \phi')_N = \frac{1}{N} \sum_k \phi(\mathbf{x}_k)\phi'(\mathbf{x}_k)$$

and the corresponding norm $\| \ \|_N$.

3. *Boundedness.* S is bounded by M if

$$\|S\phi\|_N \leq M\|\phi\|_N, \qquad \text{all } \phi \in F(\mathbf{x}),$$

where $\|S\phi\|_N$ is defined on $F(\mathbf{x}_j)$ exactly as $\|\phi\|_N$ is defined on $F(\mathbf{x})$.

In these examples of smooths, all are linear, except the supersmoother. This implies they can be represented as an $N \times N$ matrix operator varying with D. All are constant preserving. Histograms and the nearest neighbor are bounded by 2. Regression is unbounded due to end effects, but in the Section A.5 we introduce a modified regression smooth that is bounded by 2. The bound for kernel smooths is more complicated.

A.3 Convergence of ACE

Let the data be of the form $(y_k, \mathbf{x}_k) = (y_k, x_{k1}, \ldots, x_{kp})$, $k = 1, \ldots, N$. Assume that $\bar{y} = \bar{x}_1 = \cdots = \bar{x}_p = 0$. Define smooths S_y, $S_1, \ldots, S_p$, where $S_y : F(y, \mathbf{x}) \to F(y)$ and $S_j : F(y, \mathbf{x}) \to F(x_j)$. Let $H_2(y, \mathbf{x})$ be the set of all functions in $F(y, \mathbf{x})$ with zero mean, and let $H_2(y)$, $H_2(x_j)$ be the corresponding subspaces.

It is essential to modify the smooths so that the resulting functions have zero means. This is done by subtracting the mean; thus the modified S_j is defined by

$$S_j\phi = S_j\phi - Av(S_j\phi). \tag{A.1}$$

Henceforth, we use only *modified smooths* and assume the original smooth to be constant preserving so that the modified smooths take constants into zero.

The ACE algorithm is defined by the following.

1. $\theta^{(0)}(y_k) = y_k$, $\phi_j^{(0)}(x_{kj}) \equiv 0$.

(The inner loop)

2. At the n stage of the outer loop, start with $\theta^{(n)}$, $\phi_j^{(0)}$. For every $m \geq 1$ and $j = 1, \ldots, p$, define

$$\phi_j^{(m+1)} = S_j\left(\theta^{(n)} - \sum_{i<j} \phi_i^{(m+1)} - \sum_{i>j} \phi_i^{(m)}\right).$$

Keep increasing m until convergence to ϕ_j.

(The outer loop)

3. Set $\theta^{(n+1)} = S_y(\sum_j \phi_j)/\|S_y(\sum_j \phi_j)\|_N$. Go back to the inner loop with $\phi_j^{(0)} = \phi_j$ (restart) or $\phi_j^{(0)} = 0$ (fresh start). Continue until convergence.

To formalize this algorithm, introduce the space $H_2(\theta, \boldsymbol{\phi})$ with elements $(\theta, \phi_1, \ldots, \phi_p)$, $\theta \in H_2(y)$, $\phi_j \in H_2(x_j)$, and subspaces $H_2(\theta)$ with elements $(\theta, 0, 0, \ldots, 0) = \boldsymbol{\theta}$ and $H_2(\boldsymbol{\phi})$ with elements $(0, \phi_1, \ldots, \phi_p) = \boldsymbol{\phi}$.

For $f = (f_0, f_1, \ldots, f_p)$ in $H_2(\theta, \boldsymbol{\phi})$, define $S_j : H_2(\theta, \boldsymbol{\phi}) \to H_2(\theta, \boldsymbol{\phi})$ by

$$(S_jf)_i = 0, \qquad i \neq j$$

$$= f_j + S_j\left(\sum_{i \neq j} f_i\right), \qquad j = i.$$

Starting with $\boldsymbol{\theta} = (\theta, 0, 0, \ldots, 0)$, $\boldsymbol{\phi}^{(m)} = (0, \phi_p^{(m)})$, one complete cycle in the inner loop is described by

$$\boldsymbol{\theta} - \boldsymbol{\phi}^{(m+1)} = (I - S_p)(I - S_{p-1}) \cdots (I - S_1)(\boldsymbol{\theta} - \boldsymbol{\phi}^{(m)}). \tag{A.2}$$

Define $\hat{T}$ on $H_2(\theta, \boldsymbol{\phi}) \to H_2(\theta, \boldsymbol{\phi})$ as the product operator in (A.2). Then

$$\boldsymbol{\phi}^{(m)} = \boldsymbol{\theta} - \hat{T}^m(\boldsymbol{\theta} - \boldsymbol{\phi}^{(0)}). \tag{A.3}$$

If, for a given $\boldsymbol{\theta}$, the inner loop converges, then the limiting $\boldsymbol{\phi}$ satisfies

$$S_j(\boldsymbol{\theta} - \boldsymbol{\phi}) = 0, \qquad j = 1, \ldots, p. \tag{A.4}$$

That is, the smooth of the residuals on any predictor variable is zero.

Adding

$$\boldsymbol{\theta} = S_y\boldsymbol{\phi}/\|S_y\boldsymbol{\phi}\|_N \tag{A.5}$$

to (A.4) gives a set of equations satisfied by the estimated optimal transformations.

Assume, for the remainder of this section, that the smooths are linear. The (A.4) can be written as

$$S_j\boldsymbol{\phi} = S_j\boldsymbol{\theta}, \qquad j = 1, \ldots, p. \tag{A.6}$$

Let $\text{sp}(S_j)$ denote the spectrum of the matrix S_j. Assume $1 \notin \text{sp}(S_j)$. (The number 1 is in the spectrum for constant preserving smooths but not for modified smooths.) Define matrices A_j by $A_j = S_j(I - S_j)^{-1}$ and the matrix A as $\sum_j A_j$. Assume further that $-1 \notin \text{sp}(A)$. Then (A.6) has the unique solution

$$\phi_j = A_j(I + A)^{-1}\theta, \qquad j = 1, \ldots, p. \tag{A.7}$$

The element $\boldsymbol{\phi} = (0, \phi_1, \ldots, \phi_p)$ given by (A.7) will be denoted by $\hat{P}\boldsymbol{\theta}$. Rewrite (A.3) using $(I - \hat{T})(\boldsymbol{\theta} - \hat{P}\boldsymbol{\theta}) = 0$ as

$$\boldsymbol{\phi}^{(m)} = \hat{P}\boldsymbol{\theta} - \hat{T}^m(\hat{P}\theta - \boldsymbol{\phi}^{(0)}). \tag{A.8}$$

Therefore, the inner loop converges if it can be shown that $\hat{T}^m f \to 0$ for all $f \in H_2(\boldsymbol{\phi})$. What we can show is Theorem A.1.

Theorem A.1. If $\det[I + A] \neq 0$ and if the spectral radii of $S_1, \ldots, S_p$ are all less than one, a necessary and sufficient condition for

$\hat{T}^m f \to 0$ for all $f \in H_2(\boldsymbol{\phi})$ is that

$$\det\left[\lambda I - \prod_1^p (I - S_j/\lambda)^{-1}(I - S_j)\right] \tag{A.9}$$

has no zeros in $|\lambda| \geq 1$ except $\lambda = 1$.

Proof. For $\hat{T}^m f \to 0$, all $f \in H_2(\boldsymbol{\phi})$, it is necessary and sufficient that the spectral radius of $\hat{T}$ be less than one. The equation $\hat{T}f = \lambda f$ in component form is

$$\lambda f_j = -S_j\left(\lambda_i \sum_{i<j} f_i + \sum_{i>j} f_i\right), \qquad j = 1, \ldots, p. \tag{A.10}$$

Let $s = \sum_i f_i$ and rewrite (A.10) as

$$(\lambda I - S_j)f_j = S_j\left((1 - \lambda)\sum_{i<j} f_i - s\right). \tag{A.11}$$

If $\lambda = 1$, (A.11) becomes $(I - S_j)f_j = -S_j s$ or $s = -As$. By assumption, this implies that $s = 0$, and hence $f_j = 0$, for all j. This rules out $\lambda = 1$ as an eigenvalue of $\hat{T}$. For $\lambda \neq 1$, but λ greater than the maximum of the spectral radii of the S_j $(j = 1, \ldots, p)$, define $g_j = (1 - \lambda)\sum_{i<j} f_i - s$. Then $f_j = (g_{j+1} - g_j)/(1 - \lambda)$, so

$$(\lambda I - S_j)(g_{j+1} - g_j) = (1 - \lambda)S_j g_j$$

or

$$g_{j+1} = (I - S_j/\lambda)^{-1}(I - S_j)g_j. \tag{A.12}$$

Since $g_{p+1} = -\lambda s$, $g_1 = -s$; then (A.12) leads to

$$\lambda s = (I - S_p/\lambda)^{-1}(I - S_p) \cdots (I - S_1/\lambda)^{-1}(I - S_1)s. \tag{A.13}$$

If (A.13) has no nonzero solutions, then $s = 0$, $g_j = 0$, and $j = 1, \ldots, p$, implying all $f_j = 0$. Conversely, if (A.13) has a solution $s \neq 0$, it leads to a solution of (A.10).

Unfortunately, condition (A.9) is difficult to verify for general linear smooths. If the S_j are self-adjoint, non-negative definite, such that all elements in the unmodified smooth matrix are non-negative, then all spectral radii of S_j are less than one and (A.9) can be shown to hold by verifying that

$$|\lambda| \leq \prod_1^p \|(I - S_j/\lambda)^{-1}(I - S_j)\|$$

has no solutions λ with $|\lambda| > 1$ and then ruling out solutions with $|\lambda| = 1$.

Assuming that the inner loop converges to $\hat{P}\theta$, the outer loop iteration is given by

$$\boldsymbol{\theta}^{(n+1)} = S_y\hat{P}\boldsymbol{\theta}^{(n)}/\|S_y\hat{P}\boldsymbol{\theta}^{(n)}\|_N.$$

Put the matrix $S_y\hat{P} = \hat{U}$ so that

$$\theta^{(n+1)} = \hat{U}\theta^{(n)}/\|\hat{U}\theta^{(n)}\|_N. \tag{A.14}$$

If the eigenvalue $\hat{\lambda}$ of $\hat{U}$ having largest absolute value is real and positive, then $\theta^{(n+1)}$ converges to the projection of $\theta^{(0)}$ on the eigenspace of $\hat{\lambda}$. The limiting θ, $\hat{P}\theta$ is a solution of (A.4) and (A.5). If $\hat{\lambda}$ is not real and positive, then $\theta^{(n)}$ oscillates and does not converge. If the smooths are self-adjoint and non-negative definite, then $S_y\hat{P}$ is the product of two self-adjoint non-negative definite matrices; hence it has only real non-negative eigenvalues. We are unable to find conditions guaranteeing this for more general smooths.

It can be easily shown that with modifications near the endpoints, the nearest neighbor smooth satisfies the preceding conditions. Our current research indicates a possibility that other types of common smooths can also be modified into self-adjoint, non-negative definite smooths with non-negative matrix elements. For these, ACE convergence is guaranteed by the preceding arguments.

ACE, however, has invariably converged using a variety of non-self-adjoint smooths (with one exception found using an odd type of kernel smooth). We conjecture that for most data sets, reasonable smooths are "close" enough to being self-adjoint so that their largest eigenvalue is real, positive, and less than one.

A.4 Consistency of ACE

For $\phi_0, \phi_1, \ldots, \phi_p$, any functions in $H_2(Y), H_2(X_1), \ldots, H_2(X_p)$, and any data set $D \in \mathcal{D}_N$, define functions $P_j(\phi_i \mid x_j)$ by

$$P_j(\phi_i \mid x_{kj}) = E(\phi_i(X_i) \mid X_j = x_{kj}). \tag{A.15}$$

Let ϕ_j in $H_2(x_j)$ be defined as the restriction of ϕ_j to the set of data values $\{x_{1j}, \ldots, x_{Nj}\}$ minus its mean value over the data values.

Assume that the N data vectors $(y_k, \mathbf{x}_k)$ are samples from the distribution of $(Y, X_1, \ldots, X_p)$, not necessarily independent or even random (see Section A.5).

Definition A.2. Let $S_y^{(N)}$, $S_j^{(N)}$ be any sequence of data smooths. They are mean squared consistent if

$$E\|S_j^{(N)}(\phi_i \mid x_j) - P_j(\phi_i \mid x_j)\|_N^2 \to 0$$

for all $\phi_0, \ldots, \phi_p$ as above, with the analogous definition for $S_y^{(N)}$.

Whether or not the algorithm converges, a weak consistency result can be given under general conditions for the fresh-start algorithm. Start with $\theta_0 \in H_2(Y)$. On each data set, run the inner-loop iteration m times; that is, define

$$\boldsymbol{\phi}_m^{(n+1)} = \boldsymbol{\theta}^{(n)} - \hat{T}^m(\boldsymbol{\theta}^{(n)}).$$

Then set

$$\boldsymbol{\theta}_m^{(n+1)} = S_y\boldsymbol{\phi}_m^{(n+1)}/\|S_y\boldsymbol{\phi}_m^{(n+1)}\|_N.$$

Repeat the outer loop l times, getting the final functions $\theta_N(y; m, l)$, $\phi_{jN}(x_j; m, l)$. Do the analogous thing in function space starting with θ_0, getting functions whose restriction to the data set D are denoted by $\theta(y; m, l)$, $\phi_j(x_j; m, l)$.

Theorem A.2. For the fresh-start algorithm, if the smooths $S_y^{(N)}$, $S_j^{(N)}$ are mean squared consistent, linear, and uniformly bounded as $N \to \infty$, and if for any $\theta \in L_2(Y)$, $\|\theta\|_N^2 \xrightarrow{P} \|\theta\|^2$, $E\|\theta\|_N^2 \to \|\theta\|^2$, then $E\|\theta_N(y; m, l) - \theta(y; m, l)\|_N^2 \to 0$,

$$E\|\phi_{jN}(x_j; m, l) - \phi_j(x_j; m, l)\|_N^2 \to 0.$$

If θ^* is the optimal transformation $P_E\theta_0/\|P_E\theta_0\|$, $\tilde{\phi}^* = P_X\theta^*$, then as $m, l \to \infty$ in any way,

$$\|\theta(\cdot; m, l) - \theta^*\| \to 0, \qquad \|\phi_j(\cdot; m, l) - \phi_j^*\| \to 0.$$

Proof. First note that for any product of smooths $S_{i_1}^{(N)} \cdots S_{i_h}^{(N)}$,

$$E\|S_{i_1}^{(N)} \cdots S_{i_h}^{(N)}\theta_0 - P_{i_1} \cdots P_{i_h}\theta_0\|_N^2 \to 0.$$

This is illustrated with $S_i^{(N)}S_j^{(N)}\theta_0$ $(i \neq j)$. Since $E\|S_j^{(N)}\theta_0 - P_j\theta_0\|_N^2 \to 0$, then $S_j^{(N)}\theta_0 = P_j\theta_0 + \phi_{j,N}$, where $E\|\phi_{j,N}\|_N^2 \to 0$. Therefore

$$S_i^{(N)}(S_j^{(N)}\theta_0) = S_i^{(N)}P_j\theta_0 + S_i^{(N)}\phi_{j,N}.$$

By assumption, $\|S_i^{(N)}\phi_{j,N}\|_N \leq M\|\phi_{j,N}\|_N$, where M does not depend on N. Therefore $E\|S_i^{(N)}\phi_{j,N}\|_N^2 \to 0$. By assumption, $E\|S_i^{(N)}P_j\theta_0 - P_iP_j\theta_0\|_N^2 \to 0$ so that $E\|S_i^{(N)}S_j^{(N)}\theta_0 - P_iP_j\theta_0\|_N^2 \to 0$.

Proposition A.1. If θ_N is defined in $H_2(y)$ for all data sets D, and $\theta \in H_2(Y)$ such that

$$E\|\theta_N(y) - \theta(y)\|_N^2 \to 0,$$

then

$$E\left\|\frac{\theta_N(y)}{\|\theta_N\|_N} - \frac{\theta(y)}{\|\theta\|}\right\|_N^2 \to 0.$$

Proof. Write $\theta/\|\theta\| = \theta/\|\theta\|_N + \theta(1/\|\theta\| - 1/\|\theta\|_N)$. Then two parts are needed: first, to show that

$$E\left\|\frac{\theta_N}{\|\theta_N\|_N} - \frac{\theta}{\|\theta\|_N}\right\|_N^2 \to 0,$$

and second, to show that

$$E\left\|\theta\left(\frac{1}{\|\theta\|}-\frac{1}{\|\theta\|_N}\right)\right\|_N^2\to 0.$$

For the first part, let

$$S_N^2=\frac{1}{N}\sum_k\left(\frac{\theta_N(y_k)}{\|\theta_N\|_N}-\frac{\theta(y_k)}{\|\theta\|_N}\right)^2=2\left(1-\frac{(\theta_N,\theta)_N}{\|\theta_N\|_N\|\theta\|_N}\right).$$

Then $S_N^2\le 4$, so it is enough to show that $S_N^2\xrightarrow{P}0$ to get $ES_N^2\to 0$.

Let

$$\begin{aligned}V_N^2&=\frac{1}{N}\sum_k(\theta_N(y_k)-\theta(y_k))^2\\&=\|\theta_N\|_N^2+\|\theta\|_N^2-2(\theta_N,\theta)_N\\&=(\|\theta_N\|_N-\|\theta\|_N)^2+2(\|\theta\|_N\|\theta_N\|_N-(\theta_N,\theta)_N).\end{aligned}$$

Both terms are positive, and since $EV_N^2\to 0$, $E(\|\theta_N\|_N-\|\theta\|_N)^2\to 0$ and $E(\|\theta\|_N\|\theta_N\|_N-(\theta_N,\theta)_N)\to 0$. By assumption, $\|\theta\|_N^2\xrightarrow{P}\|\theta\|^2$, resulting in $S_N^2\xrightarrow{P}0$.

Now look at

$$\begin{aligned}W_N^2&=\frac{1}{N}\sum_k\theta^2(y_k)[1/\|\theta\|_N-1/\|\theta\|]^2\\&=\|\theta\|_N^2(1/\|\theta\|_N-1/\|\theta\|)^2\\&=(1-\|\theta\|_N/\|\theta\|)^2.\end{aligned}$$

Then $EW_N^2\to 0$ follows from the assumptions.

Using Proposition A.1, it follows that $E\|\theta_N(y;m,l)-\theta(y;m,l)\|_N^2\to 0$ and, in consequence, that $E\|\phi_{j,N}(x_j;m,l)-\phi_j(x_j;m,l)\|^2\to 0$.

In function space, define

$$\begin{aligned}P_X^{(m)}\theta&=\theta-T^m\theta\\U_m&=P_YP_X^{(m)}.\end{aligned}$$

Then

$$\theta(\cdot;m,l)=U_m^lθ_0/\|U_m^lθ_0\|.$$

The last step in the proof is showing that

$$\|U_m^lθ_0/\|U_m^lθ_0\|-θ^*\|\to 0$$

as m, l go to infinity. Begin with Proposition A.2.

Proposition A.2. As $m\to\infty$, $U_m\to U$ in the uniform operator norm.

Proof. $\|U_m\theta-U\theta\|=\|P_YT^mP_X\theta\|\le\|T^mP_X\theta\|$. Now on $H_2(Y)$, $\|T^mP_X\|\to 0$. If not, take θ_m, $\|\theta_m\|=1$ such that $\|T^mP_X\theta_m\|\ge\delta$, all m. Let $\theta_{m'}\xrightarrow{w}\theta$; then $P_X\theta_m\xrightarrow{s}P_X\theta$ and

$$\begin{aligned}\|T^{m'}P_X\theta_{m'}\|&\le\|T^{m'}P_X(\theta_{m'}-\theta)\|+\|T^{m'}P_X\theta\|\\&\le\|P_X(\theta_{m'}-\theta)\|+\|T^{m'}P_X\theta\|.\end{aligned}$$

By Proposition (5.5) the right-hand side goes to zero.

The operator U_m is not necessarily self-adjoint, but it is compact. By Proposition (A.2), if $0(\mathrm{sp}(U))$ is any open set containing $\mathrm{sp}(U)$, then for m sufficiently large, $\mathrm{sp}(U_m)\subset 0(\mathrm{sp}(U))$. Suppose, for simplicity, that the eigenspace $E_{\bar\lambda}$ corresponding to the largest eigenvalue $\bar\lambda$ of U is one-dimensional. (The proof goes through if $E_{\bar\lambda}$ is higher-dimensional, but it is more complicated.) Then for any open neighborhood 0 of $\bar\lambda$, and m sufficiently large, there is only one eigenvalue λ_m of U_m in 0, $\lambda_m\to\bar\lambda$, and the projection $P_E^{(m)}$ of U_m corresponding to λ_m converges to $P_{E_{\bar\lambda}}$ in the uniform operator topology. Moreover, λ_m can be taken as the eigenvalue of U_m having largest absolute value. If λ' is the second largest eigenvalue of U and λ_m' is the eigenvalue of U_m having the second highest absolute value, then (assuming $E_{\lambda'}$ is one-dimensional) $\lambda_m'\to\lambda'$.

Write

$$W_m=U_m-P_E^{(m)},\qquad W=U-P_{E_{\bar\lambda}};$$

so $\|W_m-W\|\to 0$ again. Now,

$$\begin{aligned}U_m^lθ_0&=\lambda_m^lP_E^{(m)}θ_0+W_m^lθ_0\\U^lθ_0&=\bar\lambda^lP_{E_{\bar\lambda}}θ_0+W^lθ_0.\end{aligned}\tag{A.16}$$

For any $\varepsilon>0$ we will show that there exists m_0, l_0 such that for $m\ge m_0$, $l\ge l_0$,

$$\|W_m^lθ_0\|/\lambda_m^l\le\varepsilon,\qquad\|W^lθ_0\|/\bar\lambda^l\le\varepsilon.\tag{A.17}$$

Take $r=(\bar\lambda+\lambda')/2$ and select m_0 such that $r>\max(\lambda',|\lambda_m'|;m\ge m_0)$. Denote by $R(\lambda,W_m)$ the resolvent of W_m. Then

$$W_m^l=\frac{1}{2\pi i}\int_{|\lambda|=r}\lambda^lR(\lambda,W_m)d\lambda$$

and

$$\|W_m^l\|\le\frac{1}{2\pi}r^l\int_{|\lambda|=r}\|R(\lambda,W_m)\|d|\lambda|,$$

where $d|\lambda|$ is arc length along $|\lambda|=r$. On $|\lambda|=r$, for $m\ge m_0$, $\|R(\lambda,W_m)\|$ is continuous and bounded. Furthermore, $\|R(\lambda,W_m)\|\to\|R(\lambda,W)\|$ uniformly. If $M(r)=\max_{|\lambda|=r}\|R(\lambda,W)\|$, then

$$\|W_m^l\|\le r^lM(r)(1+\Delta_m),$$

where $\Delta_m\to 0$ as $m\to\infty$. Certainly,

$$\|W^l\|\le r^lM(r).$$

Fix $\delta>0$ such that $(1+\delta)r<\bar\lambda$. Take m_0' such that for $m\ge\max(m_0,m_0')$, $\lambda_m\ge(1+\delta)r$. Then

$$\|W_m^l\|/\lambda_m^l\le 1/(1+\delta))^lM(r)(1+\Delta_m)$$

and

$$\|W^l\|/\bar\lambda^l\le(1/(1+\delta))^lM(r).$$

Now choose a new m_0 and l_0 such that (A.17) is satisfied.

Using (A.17),

$$\left\|\frac{U_m^lθ_0}{\|U_m^lθ_0\|}-\frac{P_E^{(m)}θ_0}{\|P_E^{(m)}θ_0\|}\right\|=\varepsilon_{m,l},$$

where $\varepsilon_{m,l}\to 0$ as $m,l\to\infty$. Thus

$$\left\|\frac{U_m^lθ_0}{\|U_m^lθ_0\|}-\theta^*\right\|=\varepsilon'_{m,l}+\left\|\frac{P_{E_m}θ_0}{\|P_{E_m}θ_0\|}-\frac{P_{E_{\bar\lambda}}θ_0}{\|P_{E_{\bar\lambda}}θ_0\|}\right\|,$$

and the right side goes to zero as $m,l\to\infty$.

The term *weak consistency* is used above because we have in mind a desirable stronger result. We conjecture that for reasonable smooths, the set $C_N=\{(Y_1,\mathbf{X}_1),\ldots,(Y_N,\mathbf{X}_N);\text{ algorithm converges}\}$ satisfies $P(C_N)\to 1$ and that for θ_N, the limit on C_N starting from a fixed θ_0,

$$E[I_{C_N}\|\theta_N-\theta^*\|_N^2]\to 0.$$

We also conjecture that such a theorem will be difficult to prove. A weaker, but probably much easier result would be to assume the use of self-adjoint non-negative definite smooths with non-negative matrix elements. Then we know that the algorithm converges to some θ_N, and we conjecture that $E[\|\theta_N-\theta^*\|_N^2]\to 0$.

A.5 Mean Squared Consistency of Nearest Neighbor Smooths

To show that the ACE algorithm is applicable in a situation, we need to verify that the assumptions of Theorem (A.2) can be satisfied. We do this, first assuming that the data $(Y_1,X_1),\ldots,(Y_N,X_N)$ are samples from a two-dimensional stationary, ergodic process. Then the ergodic theorem implies that for any $\theta\in L_2(Y)$, $\|\theta\|_N^2\xrightarrow{p}\|\theta\|^2$ and, trivially, $E\|\theta\|_N^2\to\|\theta\|^2$.

To show that we can get a bounded, linear sequence of smooths that are mean squared consistent, we use the nearest neighbor smooths.

Theorem A.3. Let $(Y_1, X_1), \ldots, (Y_N, X_N)$ be samples from a stationary ergodic process such that the distribution of X has no atoms. Then there exists a mean squared consistent sequence of nearest-neighbor smooths of Y on X.

The proof begins with Lemma A.1.

Lemma A.1. Suppose that $P(dx)$ has no atoms, and let $P_N(dx) \xrightarrow{w} P(dx)$. Take $\delta_N > 0$, $\delta_N \to \delta > 0$; define $J(x; \varepsilon) = [x - \varepsilon, x + \varepsilon]$; and

$$\varepsilon_N(x) = \min\{\varepsilon; P_N(J(x, \varepsilon)) \geq \delta_N\}$$
$$\varepsilon(x) = \min\{\varepsilon; P(J(x, \varepsilon)) \geq \delta\}.$$

Then using Δ to denote symmetric difference,

$$P_N(J(x, \varepsilon_N(x)) \Delta J(x, \varepsilon(x))) \to 0 \text{ uniformly in } x \qquad (A.18)$$

and

$$\lim_N \sup_{\{(x,y):|x-y|\leq h\}} P_N(J(x, \varepsilon(x)) \Delta J(y, \varepsilon(y))) \leq \varepsilon_1(h), \qquad (A.19)$$

where $\varepsilon_1(h) \to 0$ as $h \to 0$.

Proof. Let $F_N(x)$, $F(x)$ be the cumulative df corresponding to P_N, P. Since $F_N \xrightarrow{w} F$ and F is continuous, then it follows that

$$\sup_x |F_N(x) - F(x)| \to 0.$$

To prove (A.18), note that

$$\begin{aligned} &P_N(J(x, \varepsilon_N) \Delta J(x, \varepsilon)) \\ &\quad \leq |P_N(J(x, \varepsilon_N)) - P_N(J(x, \varepsilon))| \\ &\quad \leq |\delta_N - P_N(J(x, \varepsilon_N))| \\ &\qquad + |\delta_N - \delta| + |F_N(x + \varepsilon(x)) - F(x + \varepsilon(x))| \\ &\qquad + |F_N(x - \varepsilon(x)) - F_N(x - \varepsilon(x))|, \end{aligned}$$

which does it. To prove (A.19), it is sufficient to show that

$$\sup_{x,y,|x-y|\leq h} P(J(x, \varepsilon(x)) \Delta J(y, \varepsilon(y))) \leq \varepsilon_1(h).$$

First, note that

$$|\varepsilon(x) - \varepsilon(y)| \leq |x - y|.$$

If $J(x, \varepsilon(x))$, $J(y, \varepsilon(y))$ overlap, then their symmetric difference consists of two intervals I_1, I_2 such that $|I_1| \leq 2|x - y|$, $|I_2| \leq 2|x - y|$. There is an $h_0 > 0$ such that if $|x - y| \leq h_0$, the two neighborhoods always overlap. Otherwise there is a sequence $\{x_n\}$, with $\varepsilon(x_n) \to 0$ and $P(J(x_n, \varepsilon(x_n))) = \delta$, which is impossible, since P has no atoms. Then for $h \leq h_0$,

$$\sup_{x,y,|x-y|\leq h} P(J(x, \varepsilon(x)) \Delta J(y, \varepsilon(y))) \leq 2 \sup_{|i|\leq 2h} P(I)$$

and the right-hand side goes to zero as $h \to 0$.

The lemma is applied as follows: Let $g(y)$ be any bounded function in $L_2(Y)$. Define $P_\delta(g \mid x)$, using $I(\cdot)$ to denote the indicator function, as

$$\begin{aligned} &1/\delta \int g(y)\, I(x' \in J(x, \varepsilon(x)))P(dy, dx') \\ &\qquad = 1/\delta \int P_X(g \mid x')\, I(x' \in J(x, \varepsilon(x)))P(dx'). \end{aligned}$$

Note that P_δ is bounded and continuous in x. Denote by $S_\delta^{(N)}$ the smooths with $M = [N\delta]$. Proposition A.3 follows.

Proposition A.3. $E\|S_\delta^{(N)} g - P_\delta g\|_N^2 \to 0$ for fixed δ.

Proof. By (A.18), with probability one,

$$S_\delta^{(N)}(g \mid x) = (1/[\delta N]) \sum_j g(y_j) I(x_j \in J(x, \varepsilon_N(x)))$$

can be replaced for all x by

$$g_N(x, \omega) = (1/[\delta N]) \sum_j g(y_j) I(x_j \in J(x, \varepsilon(x))),$$

where ω is a sample sequence.

By the ergodic theorem, for a countable $\{x_n\}$ dense on the real line, and $\omega \in W'$, $P(W') = 1$,

$$\Phi_N(x_n, \omega) = g_N(x_n, \omega) - P_\delta(g \mid x_n) \to 0.$$

Use (A.19) to establish that for any bounded interval J and any $\omega \in W'$, $\Phi_N(x, \omega) \to 0$ uniformly for $x \in J$. Then write

$$\begin{aligned} \|\Phi_N(x, \omega)\|_N^2 &= \frac{1}{N} \sum_{k=1}^N \Phi_N^2(x_k, \omega) I(x_k \in J) \\ &\quad + \frac{1}{N} \sum_{k=1}^N \Phi_N^2(x_k, \omega) I(x_k \in' J). \end{aligned}$$

The first term is bounded and goes to zero for $\omega \in W'$; hence its expectation goes to zero. The expectation of the second term is bounded by $cP(X \in' J)$. Since J can be taken arbitrarily large, this completes the proof.

Using the inequality

$$E\|S_\delta^{(N)} g - P_X g\|_N^2 \leq 2\, E\|S_\delta^{(N)} g - P_\delta g\|_N^2 + 2\|P_\delta g - P_X g\|^2$$

gives

$$\limsup E\|S_\delta^{(N)} g - P_X g\|_N^2 \leq 2\|P_\delta g - P_X g\|^2.$$

Proposition A.4. For any $\phi(x)\, \varepsilon\, L_2(X)$, $\lim_{\delta\to 0}\|P_\delta \phi - \phi\| \to 0$.

Proof. For ϕ bounded and continuous,

$$\frac{1}{\delta} \int \phi(x') I(x' \in J(x, \varepsilon(x))) P(dx') \to \phi(x)$$

as $\delta \to 0$ for every x. Since $\sup|P_\delta \phi - \phi| \leq c$ for all δ, then $\|P_\delta \phi - \phi\| \to 0$. The proposition follows if it can be shown that for every $\phi \in L_2(X)$, $\limsup_\delta \|P_\delta \phi\| < \infty$. But

$$\begin{aligned} \|P_\delta \phi\|^2 &= \int \left[\frac{1}{\delta} \int \phi(x') I(x' \in J(x, \varepsilon(x))) P(dx')\right]^2 P(dx) \\ &\leq \frac{1}{\delta} \int \phi(x')^2 P(dx') \left[\int I(x' \in J(x, \varepsilon(x))) P(dx)\right]. \end{aligned}$$

Suppose that x' is such that there are numbers ε^+, ε^- with $P([x', x' + \varepsilon^+]) = \delta$, $P([x', x' - \varepsilon^-]) = \delta$. Then $x' \in J(x, \varepsilon(x))$ implies $x' - \varepsilon^- \leq x \leq x' + \varepsilon^+$, and

$$1/\delta \int I(x' \in J(x, \varepsilon(x))) P(dx) \leq 2. \qquad (A.20)$$

If, say, $P([x', \infty)) < \delta$, then $x \geq x' - \varepsilon^-$ and (A.20) still holds, and similarly if $P((-\infty, x']) < \delta$.

Take $\{\theta_n\}$ to be a countable set of functions dense in $L_2(Y)$. By Propositions A.3 and A.4, for any $\varepsilon > 0$, we can select $\delta(\varepsilon, n)$, $N(\delta, n)$ so that for all n,

$$E\|S_\delta^{(N)} \theta_n - P_X \theta_n\|_N^2 \leq \varepsilon \quad \text{for } \delta \leq \delta(\varepsilon, n),\ N \geq N(\delta, n).$$

Let $\varepsilon_M \downarrow 0$ as $M \to \infty$; define $\delta_M = \min_{n\leq M} \delta(\varepsilon, n)$ and $N(M) = \max_{n\leq M} N(\delta_M, n)$. Then

$$E\|S_{\delta_M}^{(N)} \theta_n - P_X \theta_n\|_N^2 \leq \varepsilon_M \quad \text{for } n \leq M,\ N \geq N(M).$$

Put $M(N) = \max\{M; N \geq \max(M, N(M))\}$. Then $M(N) \to \infty$ as $N \to \infty$, and the sequence of smooths $S_{\delta_{M(N)}}^{(N)}$ is mean squared consistent for all θ_n. Noting that for $\theta \in L_2(Y)$,

$$E\|S_\delta^{(N)} \theta - P_x \theta\|^2 \leq 3E\|S_\delta^{(N)} \theta_n - P_X \theta_n\|_N^2 + 9\|\theta - \theta_n\|^2$$

completes the proof of the theorem.

The fact that ACE uses modified smooths $S_\delta^{(N)} g = S_\delta^{(N)} g - Av(S_\delta^{(N)} g)$ and functions g such that $Eg = 0$ causes no problems, since

$$\|Av(S_\delta^{(N)} g)\|_N^2 = (Av(S_\delta^{(N)} g))^2$$

and

$$Av(S_\delta^{(N)} g) = \frac{1}{N} \sum_{k=1}^N g_N(x_k, \omega),$$

using the notation of Proposition A.3.

Assume g is bounded, and write

$$Av(S_\delta^{(N)}g) = \frac{1}{N}\sum_{k=1}^{N}\Phi_N(x_k, \omega) + \frac{1}{N}\sum_{k=1}^{N}P_\delta(g \mid x_k).$$

By the ergodic theorem, the second term goes a.s. to $EP_\delta(g \mid X)$, and an argument mimicking the proof of Proposition A.3 shows that the first term goes to zero a.s.

Finally, write

$$|EP_\delta(g \mid X)| = |EP_\delta(g \mid X) - EP_Xg| \le \|P_\delta\phi - \phi\|,$$

where $\phi = P_Xg$. Thus, Theorem A.3 can be easily changed to account for modified smooths.

In the controlled experiment situation, the $\{\mathbf{x}_k\}$ are not random, but the condition $\hat{P}_N(d\mathbf{x}) \xrightarrow{w} P(d\mathbf{x})$ is imposed. Additional assumptions are necessary.

Assumption A.1. For $\theta(Y)$ any bounded function in $L_2(Y)$, $E(\theta(Y) \mid \mathbf{X} = \mathbf{x})$ is continuous in $\mathbf{x}$.

Assumption A.2. For $i \neq j$ and $\phi(x)$ any bounded continuous function, $E(\phi(X_i) \mid X_j = x)$ is continuous in x.

A necessary result is Proposition A.5.

Proposition A.5. For $\theta(y)$ bounded in $L_2(Y)$ and $\phi(\mathbf{x})$ bounded and continuous,

$$\frac{1}{N}\sum_{j=1}^{N}\theta(y_j)\phi(\mathbf{x}_j) \xrightarrow{a.s.} E\theta(Y)\phi(X).$$

Let $T_N = \sum_{j=1}^{N}\theta(Y_j)\phi(\mathbf{x}_j)$. Then $ET_N = \sum_{j=1}^{N}g(\mathbf{x}_j)\phi(\mathbf{x}_j)$, $g(\mathbf{x}) = E[\theta(Y) \mid \mathbf{X} = \mathbf{x}]$. By hypothesis, $ET_N/N \to E\theta(Y)\phi(X)$. Furthermore,

$$s_N^2 = \operatorname{var}(T_N) = \sum_{j=1}^{N}E[\theta(y_j) - g(\mathbf{x}_j)]^2\phi(\mathbf{x}_j)$$

$$= \sum_{j=1}^{N}h(\mathbf{x}_j)\phi(\mathbf{x}_j),$$

where $h(\mathbf{x}) = E[(\theta(Y) - g(\mathbf{x}))^2 \mid \mathbf{X} = \mathbf{x}]$. Since $h\phi$ is continuous and bounded, then $s_N^2/N \to Eh(\mathbf{X})\phi(\mathbf{X})$. Now the application of Kolmogorov's exponential bound gives

$$T_N/N - ET_N/N \xrightarrow{a.s.} 0,$$

proving the proposition.

In Theorem A.2 we add the restriction that θ_0 be a bounded function in $L_2(Y)$. Then the condition on θ may be relaxed to the following: For θ, any bounded function in $L_2(Y)$, $\|\theta\|_N^2 \xrightarrow{P} \|\theta\|^2$, $E\|\theta\|_N^2 \to \|\theta\|^2$. These follow from Proposition A.5 and its proof. Furthermore, because of Assumptions A.1 and A.2, mean squared consistency of the smooths can be relaxed to the following requirements.

Assumption A.3. For $i \neq j$ and every bounded continuous function $\phi(x_i)$,

$$\|S_j\phi - P_j\phi\|_N^2 \to 0.$$

Assumption A.4. For every bounded function $\theta(y) \in L_2(Y)$,

$$E\|S_j\theta - P_j\theta\|_N^2 \to 0.$$

Assumption A.5. For every bounded continuous function $\phi(x_i)$,

$$E\|S_v\phi - P_v\phi\|_N^2 \to 0.$$

The existence of sequences of nearest-neighbor smooths satisfying Assumptions A.3, A.4, and A.5 can be proven in a fashion similar to the proof of Theorem A.3. Assumption A.3 is proven using Lemma A.1 and Proposition A.4. Assumptions A.4 and A.5 require Proposition A.5 in addition.

If the data are iid, stronger results can be obtained. For instance, mean squared consistency can be proven for a modified regression smooth similar to the supersmoother. For x of any point, let $J(x)$ be the indexes of the M points in $\{x_k\}$ directly above x plus the M below. If there are only $M' < M$ above (below), then include the $M+$ ($M - M'$) directly below (above). For a regression smooth,

$$S(\phi \mid x) = \bar{\phi}_x + [\Gamma_x(\phi, x)/\sigma_x^2](x - \bar{x}_x), \qquad (A.21)$$

where $\bar{\phi}_x$, $\bar{x}_x$ are the averages of $\phi(y_k)$, x_k over the indexes in $J(x)$, and $\Gamma_x(\phi, x)$, σ_x^2 are the covariance between $\phi(y_k)$, x_k and the variance of x_k over the indexes in $J(x)$.

Write the second term in (A.21) as

$$[\Gamma_x(\phi, x)/\sigma_x][(x - \bar{x}_x)/\sigma_x].$$

If there are M points above and below in $J(x)$, it is not hard to show that

$$|(x - \bar{x}_x)/\sigma_x| \le 1.$$

This is not true near endpoints where $(x - \bar{x}_x)/\sigma_x$ can become arbitrarily large as M gets large. This endpoint behavior keeps regression from being uniformly bounded. To remedy this, define a function

$$[x]_t = x, \qquad |x| \le 1$$
$$= \operatorname{sign}(x), \qquad |x| > 1,$$

and define the *modified regression smooth* by

$$S(\phi \mid x) = \bar{\phi}_x + \Gamma_x(\phi, x)/\sigma_x[(x - \bar{x}_x)/\sigma_x]_t. \qquad (A.22)$$

This modified smooth is bounded by 2.

Theorem A.4. If, as $N \to \infty$, $M \to \infty$, $M/N \to 0$, and $P(dx)$ has no atoms, then the modified regression smooths are mean squared consistent.

The proof is in Breiman and Friedman (1982). We are almost certain that the modified regression smooths are also mean squared consistent for stationary ergodic time series and in the weaker sense for controlled experiments, but under less definitive conditions on rates at which $M \to \infty$.

APPENDIX B: VARIABLES USED IN THE HOUSING VALUE EQUATION OF HARRISON AND RUBINFELD (1978)

MV—median value of owner-occupied home
RM—average number of rooms in owner units
AGE—proportion of owner units built prior to 1940
DIS—weighted distances to five employment centers in the Boston region
RAD—index of accessibility to radial highways
TAX—full property tax rate ($/$10,000)
PTRATIO—pupil-teacher ratio by town school district
B—black proportion of population
LSTAT—proportion of population that is lower status
CRIM—crime rate by town
ZN—proportion of town's residential land zoned for lots greater than 25,000 square feet
INDUS—proportion of nonretail business acres per town
CHAS—Charles River dummy = 1 if tract bounds the Charles River, 0 otherwise
NOX—nitrogen oxide concentration in pphm

APPENDIX C: VARIABLES USED IN THE OZONE-POLLUTION EXAMPLE

SBTP—Sandburg Air Force Base temperature (C°)
IBHT—inversion base height (ft.)
DGPG—Daggett pressure gradient (mmhg)
VSTY—visibility (miles)
VDHT—Vandenburg 500 millibar height (m)
HMDT—humidity (percent)

IBTP—inversion base temperature (F°)
WDSP—wind speed (mph)

Dependent Variable:

UPO3—Upland ozone concentration (ppm)

[*Received August 1982. Revised July 1984.*]

REFERENCES

Anscombe, F. J., and Tukey, J. W. (1963), "The Examination and Analysis of Residuals," *Technometrics*, 5, 141–160.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics*, New York: John Wiley.

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

Box, G. E. P., and Hill, W. J. (1974), "Correcting Inhomogeneity of Variance With Power Transformation Weighting," *Technometrics*, 16, 385–389.

Box, G. E. P., and Tidwell, P. W. (1962), "Transformations of the Independent Variables," *Technometrics*, 4, 531–550.

Breiman, L., and Friedman, J. (1982), "Estimating Optimal Transformations for Multiple Regression and Correlation," Technical Report 9, University of California, Berkeley, Dept. of Statistics.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 828–836.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 317–403.

Csáki, P., and Fisher, J. (1963), "On the General Notion of Maximal Correlation," *Magyar Tudomanyos Akademia, Budapest, Matematikai Kotato Intezet, Kozlemenyei*, 8, 27–51.

DeBoor, C. (1978), *A Practical Guide to Splines*, New York: Springer-Verlag.

De Leeuw, J., Young, F. W., and Takane, Y. (1976), "Additive Structure in Qualitative Data: An Alternating Least Squares Method With Optimal Scaling Features," *Psychometrika*, 41, 471–503.

Devroye, L. (1981), "On the Almost Everywhere Convergence of Nonparametric Regression Function Estimates," *The Annals of Statistics*, 9, 1310–1319.

Devroye, L., and Wagner, T. J. (1980), "Distribution-Free Consistency Results in Nonparametric Discrimination and Regression Function Estimation," *The Annals of Statistics*, 8, 231–239.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

Fraser, D. A. S. (1967), "Data Transformations and the Linear Model," *Annals of Mathematical Statistics*, 38, 1456–1465.

Friedman, J. H., and Stuetzle, W. (1982), "Smoothing of Scatterplots," Technical Report ORION006, Stanford University, Dept. of Statistics.

Gasser, T., and Rosenblatt, M. (eds.) (1979), "Smoothing Techniques for Curve Estimation," in *Lecture Notes in Mathematics*, No. 757, New York: Springer-Verlag.

Gebelein, H. (1947), "Das Statitistiche Problem der Korrelation als Variations und Eigenwert Problem und Sein Zusammenhang mit der Ausgleichung-Srechnung," *Zeitschrift Fuer Angewandte Mathematik und Mechanik*, 21, 364–379.

Harrison, D., and Rubinfeld, D. L. (1978), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics Management*, 5, 81–102.

Kendall, M. A., and Stuart, A. (1967), *The Advanced Theory of Statistics* (Vol. 2), New York: Hafner Publishing.

Kimeldorf, G., May, J. H., and Sampson, A. R. (1982), "Concordant and Discordant Monotone Correlations and Their Evaluations by Nonlinear Optimization," *Studies in the Management Sciences* (19): *Optimization in Statistics*, eds. S. H. Zanakis and J. S. Rustagi, Amsterdam: North-Holland, pp. 117–130.

Kruskal, J. B. (1964), "Nonmetric Multidimensional Scaling: A Numerical Method," *Psychometrika*, 29, 115–129.

——— (1965), "Analysis of Factorial Experiments by Estimating Monotone Transformations of the Data," *Journal of the Royal Statistical Society*, Ser. B, 27, 251–263.

Lancaster, H. O. (1958), "The Structure of Bivariate Distributions," *Annals of Mathematical Statistics*, 29, 719–736.

——— (1969), *The Chi-Squared Distribution*, New York: John Wiley.

Linsey, J. K. (1972), "Fitting Response Surfaces With Power Transformations," *Journal of the Royal Statistical Society*, Ser. C, 21, 234–237.

——— (1974), "Construction and Comparison of Statistical Models," *Journal of the Royal Statistical Society*, Ser. B, 36, 418–425.

Mosteller, F., and Tukey, J. W. (1977); *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Renyi, A. (1959), "On Measures of Dependence," *Acta Mathematica Academiae Scientiarum Hungaricae*, 10, 441–451.

Sarmanov, O. V. (1958a), "The Maximal Correlation Coefficient (Symmetric Case)," *Doklady Akademii Nauk UzSSR*, 120, 715–718.

——— (1958b), "The Maximal Correlation Coefficient (Nonsymmetric Case)," *Doklady Akademii Nauk UzSSR*, 121, 52–55.

Sarmanov, O. V., and Zaharov, V. K. (1960), "Maximum Coefficients of Multiple Correlation," *Doklady Akademii Nauk UzSSR*, 130, 269–271.

Spiegelman, C., and Sacks, J. (1980), "Consistent Window Estimation in Nonparametric Regression," *The Annals of Statistics*, 8, 240–246.

Stone, C. J. (1977), "Consistent Nonparametric Regression," *The Annals of Statistics*, 7, 139–149.

Tukey, J. W. (1982), "The Use of Smelting in Guiding Re-Expression," in *Modern Data Analysis*, eds. J. Laurner and A. Siegel, New York: Academic Press.

Wood, J. T. (1974), "An Extension of the Analysis of Transformations of Box and Cox," *Journal of the Royal Statistical Society*, Ser. C, 23, 278–283.

Young, F. W., de Leeuw, J., and Takane, Y. (1976), "Regression With Qualitative and Quantitative Variables: An Alternating Least Squares Method With Optimal Scaling Features," *Psychometrika*, 41, 505–529.

Comment

DARYL PREGIBON and YEHUDA VARDI*

In data analysis, the choice of transformations is often done subjectively. ACE is a major attempt to bring objectivity to this area. As Breiman and Friedman have demonstrated with their examples, and as we have experienced with our own, ACE is a powerful tool indeed. Our comments are sometimes critical in nature and reflect our view that there is much more to be done on the subject. We consider the methodology a significant contribution to statistics, however, and would like to compliment the authors for attacking an important problem, for narrowing the gap between mathematical statistics and data analysis, and for providing the data analyst with a useful tool.

1. ACE IN THEORY: HOW MEANINGFUL IS MAXIMAL CORRELATION?

To keep our discussion simple we limit it here to the bivariate case, though the issues that we raise are equally relevant to the general case. The basis of ACE lies in the properties of maximal

* Daryl Pregibon and Yehuda Vardi are Members of Technical Staff, AT & T Bell Laboratories, Murray Hill, NJ 07974.

Journal of the American Statistical Association
September 1985, Vol. 80, No. 391, Theory and Methods