ESTIMATING OPTIMAL TRANSFORMATIONS FOR MULTIPLE REGRESSION AND CORRELATION

BY

LEO BREIMAN AND JEROME H. FRIEDMAN

TECHNICAL REPORT NO. 9
JULY 1982
REVISED MAY 1983

RESEARCH PARTIALLY SUPPORTED BY
OFFICE OF NAVAL RESEARCH CONTRACTS
N00014-82-K-0054 AND N00014-81-K-0340

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA

ESTIMATING OPTIMAL TRANSFORMATIONS FOR MULTIPLE REGRESSION AND CORRELATION

Leo Breiman

Department of Statistics
University of California
Berkeley, California 94720

and

Jerome H. Friedman

Stanford Linear Accelerator Center
and
Department of Statistics
Stanford University
Stanford, California 94305

Abstract

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

$$e^2 = \frac{E\{[\theta(Y) - \sum_{j=1}^{p} \phi_j(X_j)]^2\}}{Var[\theta(Y)]}$$

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.

*Work supported by Office of Naval Research under contracts N00014-82-K-0054 and N00014-81-K-0340.

KEY WORDS: Transformations, regression, correlation, smoothing

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 aid 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

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

Then define *optimal transformations* as functions $\theta^*,\phi_1^*,\ldots,\phi_p^*$ that minimize (1.1): i.e.

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

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 and/or the predictors involve arbitrary mixtures of continuous ordered variables and categorical variables (ordered or unordered). The functions $\theta,\phi_1,\ldots,\phi_p$ 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 $\{(y_k,x_{k1},\ldots,x_{kp}),\ 1\leq k\leq N\}$ 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. (We illustrate below situations where the optimal transformations are not monotone.)

It is applicable to at least three situations

random designs in regression
autoregressive schemes in stationary ergodic times series
controlled designs in regression

In the first of these, we assume the data $(y_k,\underline{x}_k)$, $k=1,\ldots,N$ 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

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

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

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

In the controlled design situation, a distribution $P(dy|\underline{x})$ for the

response variable Y is specified for every point $\underline{x} = (x_1,\ldots,x_p)$ in the design space. The N^{th} order design consists of a specification of N points $\underline{x}_1,\ldots,\underline{x}_N$ in the design space and the data consists 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 $P(dy|\underline{x}_k)$.

Denote by $\hat{P}_N(d\underline{x})$ the empirical distribution that gives mass $1/N$ to each of the points $\underline{x}_1,\ldots,\underline{x}_N$. Assume further that

$$\hat{P}_N \xrightarrow{w} P$$

where $P(d\underline{x})$ is a probability measure on the design space. Then $P(dy|\underline{x})$ and $P(d\underline{x})$ 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

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

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

(a) $0 \leq \rho^*(X,Y) \leq 1$

(b) $\rho^*(X,Y) = 0$ if and only if X and Y are independent

(c) 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 sections 5 and 6. We next illustrate the procedure in Section 3 by applying it to a simulated data set where 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 Belsey, Kuh and Welsch [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.

Sections 5 and 6 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.

Section 6 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 very restricted class of data smooths. However, in over a thousand 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.

Section 6 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 i.i.d. data there are known results concerning the consistency of various smooths. Stone's pioneering paper [1977] 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.

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

This paper is laid out in two distinct parts. Sections 1-4 give a fairly non-technical overview of the method and discuss its application to data. Sections 5 and 6 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 develope is the MORALS algorithm due to Young et. al.[1976] (see also de Leeuw et. al. [1976]). It uses a similar alternating least squares fit, but 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] gives 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 itractable with only a few known solutions. He did not consider the problem of estimating θ^*, φ^* from data.

Kolmogorov (see Sarmanov and Zaharov [1960], 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

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

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, May, and Sampson [1982], who construct 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

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

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

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

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

```
BASIC ACE ALGORITHM
   set θ(Y) = Y/‖Y‖;
   ITERATE UNTIL e²(θ,φ) fails to decrease:
      φ₁(X) = E[θ(Y)|X];
      replace φ(X) with φ₁(X);
      θ₁(Y) = E[φ(X)|Y]/‖E[φ(X)|Y]‖;
      replace θ(Y) with θ₁(Y);
   END ITERATION LOOP;
   θ and φ are the solutions θ* and φ*;
END ALGORTHM;
```

This algorithm decreases (2.1) at each step by alternatingly minimizing with respect to one function 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 ($\theta = Y/\|Y\|$ above) 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

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

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 $\theta(Y)$ yields

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

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

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

The corresponding iterative algorithm is then:

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:
    $\phi_{k,1}(X_k) = E[\theta(Y) - \sum_{i\neq k} \phi_i(X_i)|X_k]$;
    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,...,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:
        $\phi_{k,1}(X_k) = E[\theta(Y) - \sum_{i\neq k} \phi_i(X_i)|X_k]$;
        replace $\phi_k(X_k)$ with $\phi_{k,1}(X_k)$;
      END FOR LOOP;
    END INNER ITERATION LOOP;
    $\theta_1(Y) = E[\sum_{i=1}^{p} \phi_i(X_i)|Y]/\|E[\sum_{i=1}^{p} \phi_i(X_i)|Y]\|$
    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 ALGORTHM;

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 $\{(y_k,x_{k1},\ldots,x_{kp}),\ 1 \le k \le N\}$ 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 quantity 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:

$$\hat{E}[A|Z=z] = \sum_{z_j=z} A_j \Big/ \sum_{z_j=z} 1$$

where A is a real valued quantity 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 (see, for example, Gasser and Rosenblatt [1979], Cleveland [1979], 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 ("super smoother", Friedman and Stuetzle [1984]).

The algorithm evaluates $\hat{\theta}^*,\hat{\phi}_1^*,\ldots,\hat{\phi}_p^*$ at all the corresponding data values, i.e. $\hat{\theta}^*(y)$ 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 $\hat{\phi}_1^*,\ldots,\hat{\phi}_p^*$ versus the data values of $x_1,\ldots,x_p$ 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 Belsey, Kuh and Welsch [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

$$y_k = \exp[x_k^3 + e_k]$$

with the x_k^3 and the e_k drawn independently from a standard normal distribution $N(0,1)$. Figure 1a shows a scatterplot of these data. Figures 1b-1d show the results of applying the ACE algorithm to the data. The estimated optimal transformation $\hat{\theta}^*(y)$ is shown in the plot 1b of

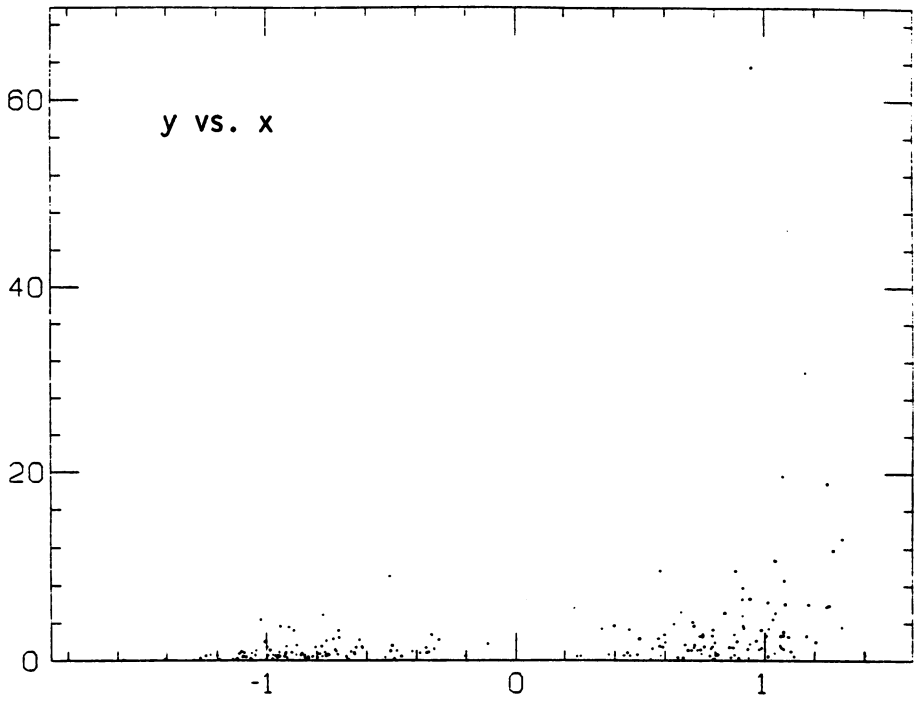


Figure 1a

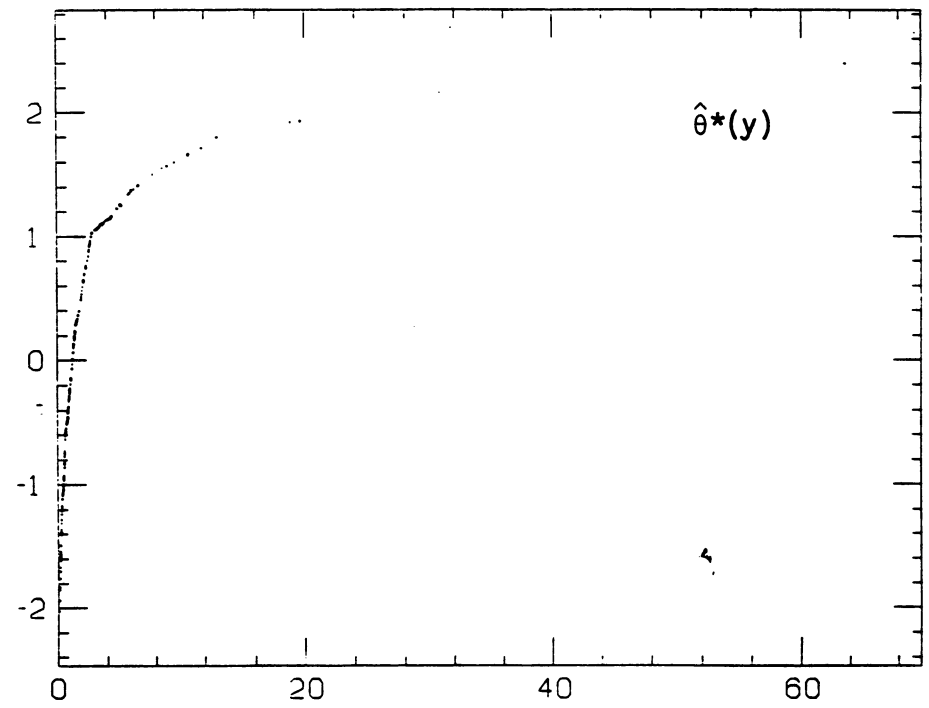


Figure 1b

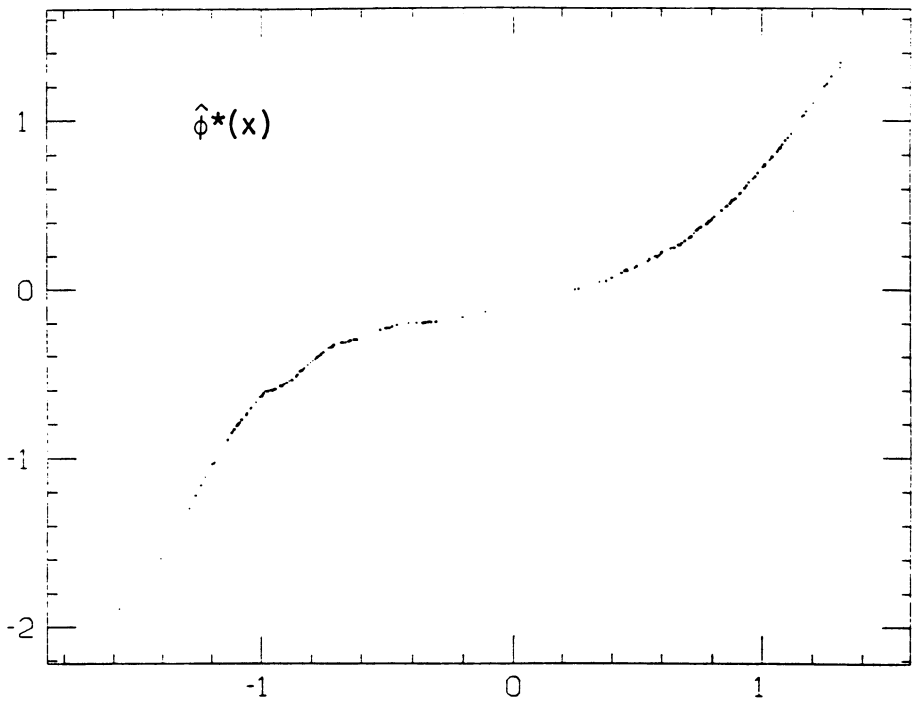


Figure 1c

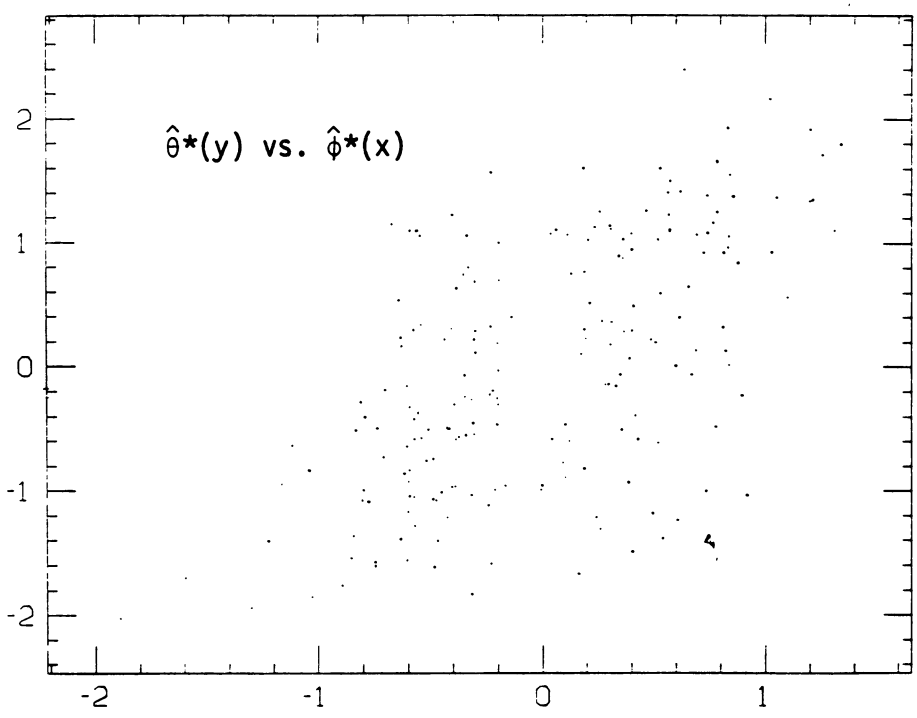


Figure 1d

$\hat{\theta}^*(y_k)$ versus y_k, $1 \leq k \leq 200$. Figure 1c 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 1d 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} using the above model. The known optimal transformations are $\theta(Y) = \log Y$, $\phi(X) = X^3$. Therefore, we define the *direct* estimate for ρ^* given any data set generated as above by

$$\hat{\hat{\rho}}^* = \frac{1}{N} \sum_{k=1}^{N} (\log y_k - \overline{\log y})(x_k^3 - \overline{x^3})$$

and $\hat{\hat{R}}^{*2} = \hat{\hat{\rho}}^{*2}$. 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 $\rho^* = 0.707$ and $R^{*2} = 0.5$.

For 100 data sets, each of size 200, generated from the above model, the means and standard deviations of the ρ^* estimates are

	mean	s.d.
ρ^* direct	.700	.034
ACE	.709	.036

The means and standard deviations of the R^{*2} estimates are

	mean	s.d.
R^{*2} direct	.492	.047
ACE	.503	.050

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

	mean	s.d.
$\hat{\rho}^* - \hat{\hat{\rho}}^*$	.001	.015
$\hat{R}^{*2} - \hat{\hat{R}}^{*2}$	.012	.022

The above experiment was duplicated for smaller sample size $N = 100$. In this case we obtain

	mean	s.d.
$\hat{\rho}^* - \hat{\hat{\rho}}^*$	.029	.034
$\hat{R}^{*2} - \hat{\hat{R}}^{*2}$	.042	.051

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

$$y_k = \exp[\sin(x_k) + e_k/2] \quad (1 \leq k \leq 200)$$

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

Figure 2a

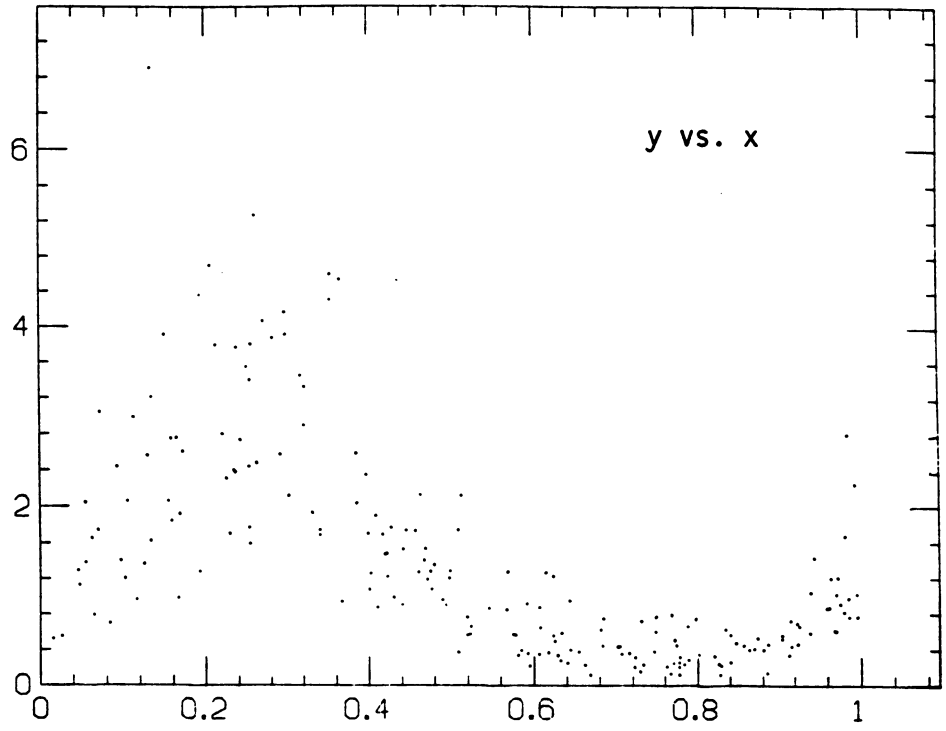


Figure 2b

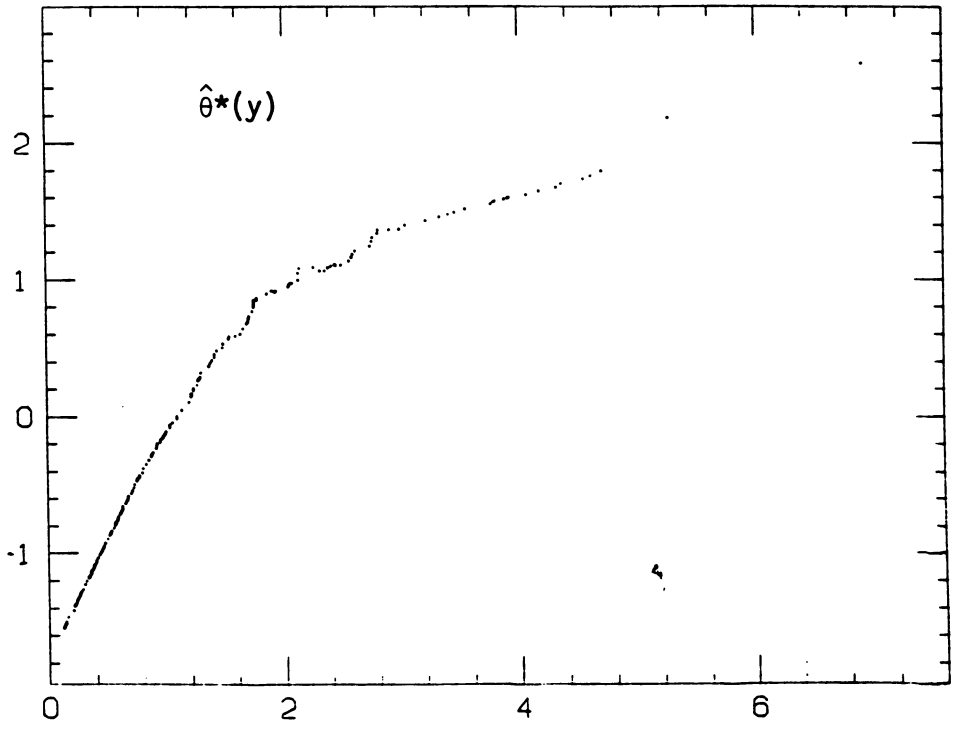
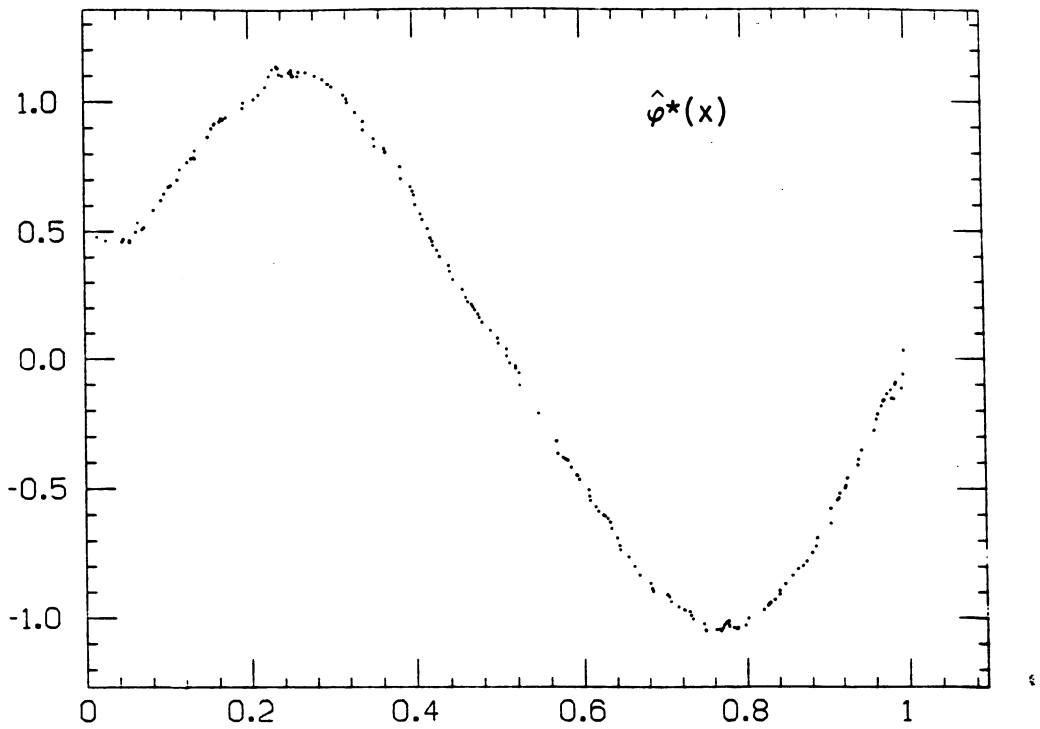


Figure 2c



Our next example consists of a sample of 200 triples $\{(y_k, x_{k1}, x_{k2}), 1 \leq k \leq 200\}$ drawn from the model $Y = X_1 X_2$ 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 3a shows a plot of $\hat{\theta}^*(y_k)$ versus y_k, while Figures 3b and 3c show corresponding plots of $\hat{\phi}_1^*(x_{k1})$ and $\hat{\phi}_2^*(x_{k2})$ $(1 \leq k \leq 200)$. 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 non-monotonic estimates for both response as well as predictor transformations.

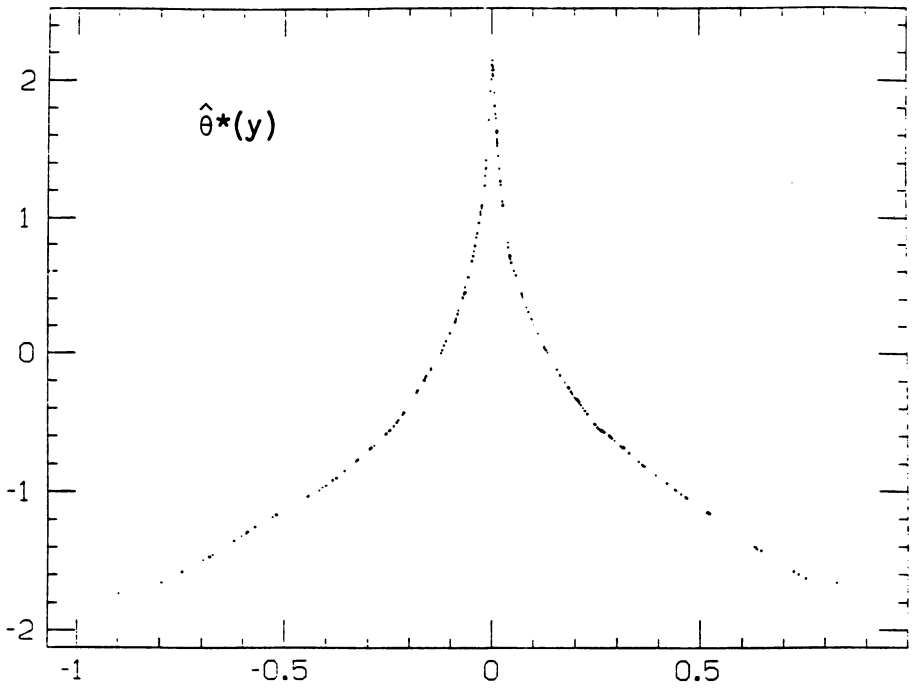


Figure 3a

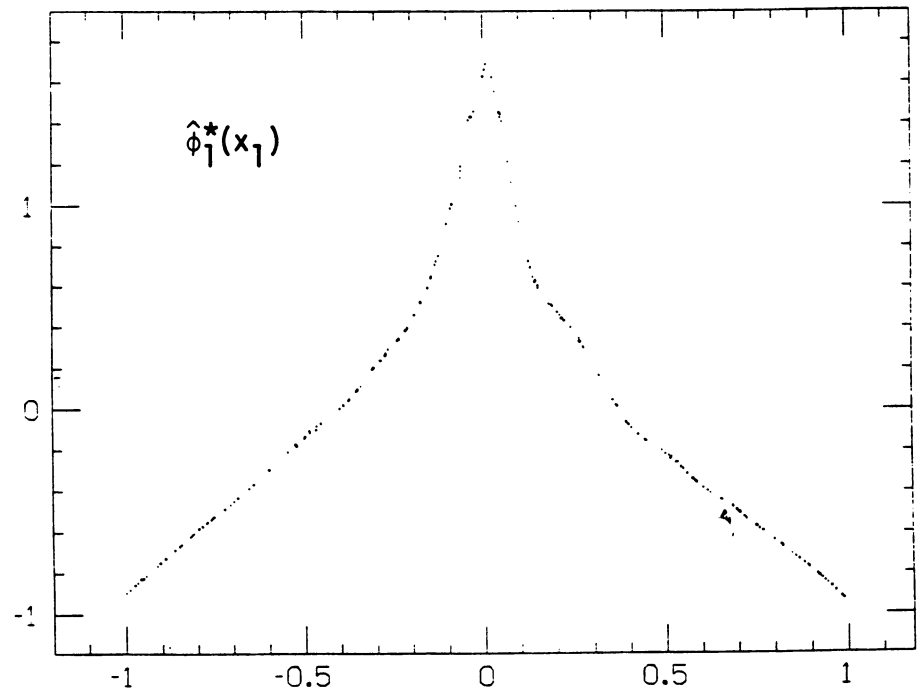


Figure 3b

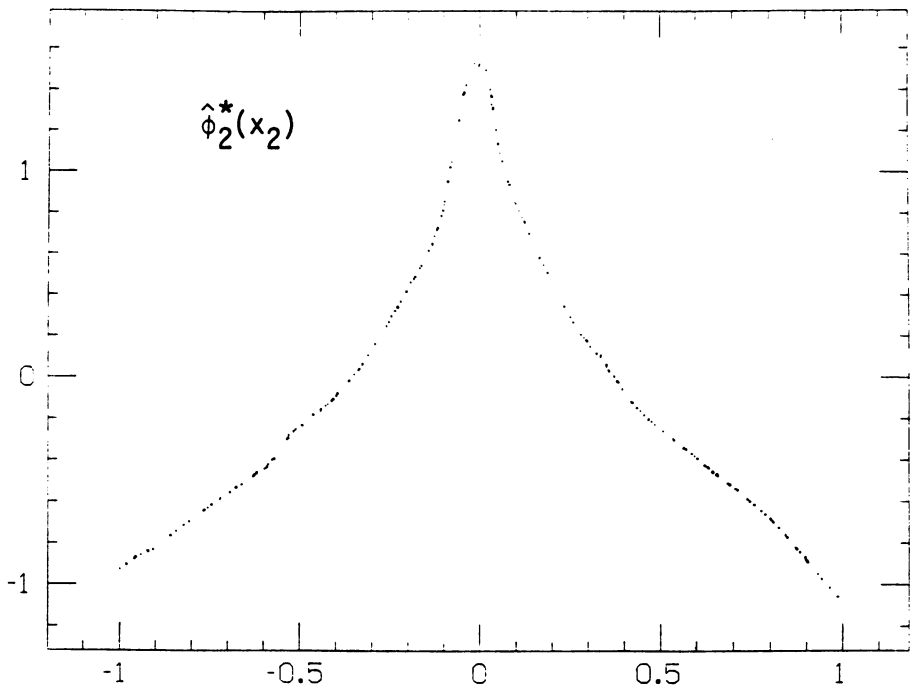


Figure 3c

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 appear in Belsey, Kuh and Welsch [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 which 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 effect 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

$$\begin{aligned}\log(MV) = \alpha_1 &+ \alpha_2(RM)^2 + \alpha_3\, AGE \\ &+ \alpha_4\, \log(DIS) + \alpha_5\, \log(RAD) + \alpha_6\, TAX \\ &+ \alpha_7\, PTRATIO + \alpha_8(B-0.63)^2 \\ &+ \alpha_9\, \log(LSTAT) + \alpha_{10}\, CRIM + \alpha_{11}\, ZN \\ &+ \alpha_{12}\, INDUS + \alpha_{13}\, CHAS + \alpha_{14}(NOX)^p + \epsilon\ .\end{aligned}$$

A brief description of each variable is given in Table 1. (For a more complete description, see Harrison and Rubinfeld [1978], Table IV.) The coefficients $\alpha_1,\ldots,\alpha_{14}$ 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.

TABLE 1

Variables Used in the Housing Value Equation
of Harrison and Rubinfeld (1978)

Variable	Definition
MV	Median value of owner-occupied homes
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 if otherwise
NOX	Nitrogen oxide concentration in pphm

Harrison and Rubinfeld note 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 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 0.01.

The resulting final model involved four predictors and had an $\hat{R}^2$ of 0.89. Applying ACE simultaneously to all 13 predictors results in an increased $\hat{R}^2$ of only 0.02.

Figure 4a shows a plot of the solution response transformation $\hat{\theta}(y')$. 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 4b 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 4c-4f 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 4c) and the logarithm of the fraction of population that is of lower status (Figure 4d). 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 less than six rooms, housing value is roughly independent of room number, while for larger values there is a strong increasing linear dependence. The remaining two variables then enter into this model are pupil-teacher ratio and property tax rate. The solution transformation for the former, Figure 4e, is seen to be approximately linear while that for the latter, Figure 4f, has considerable nonlinear structure. For tax rates up to \$320 housing price seems to fall rapidly with increasing tax, while for larger rates the association is roughly constant.

Although the variable $(\text{NOX})^2$ 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 4g. 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 4h shows a scatterplot of $\hat{\theta}^*(y_k)$ versus $\sum_{j=1}^{4} \hat{\phi}_j^*(x_{kj})$ 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 0.11 as compared to the e^2 value of 0.20 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. Table 2 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 Table 2. The resulting $\hat{R}^2$ was 0.78. Running the ACE algorithm with all eight predictor variables produces an $\hat{R}^2$ of 0.79.

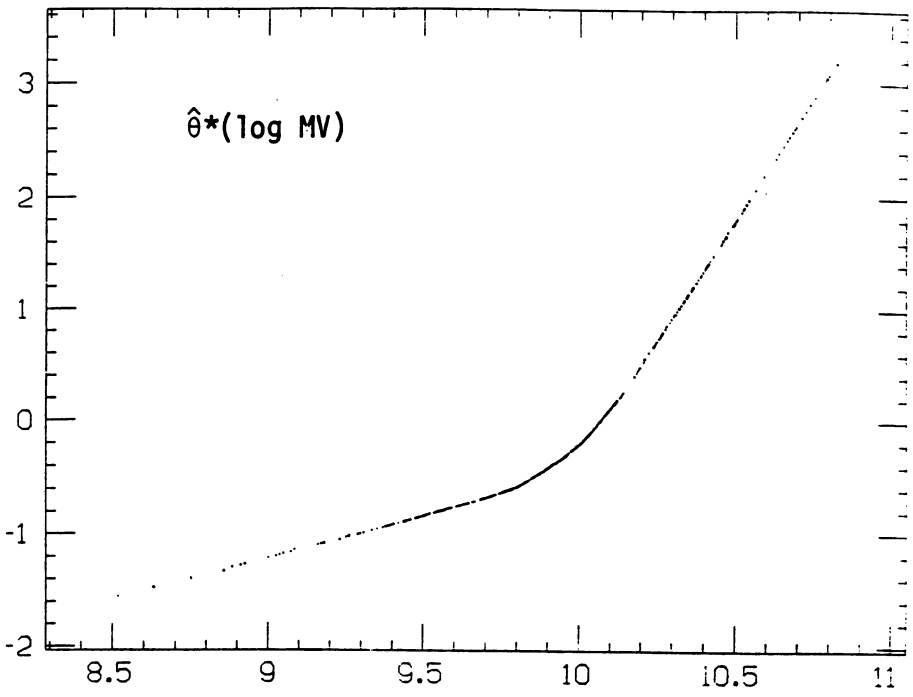


Figure 4a

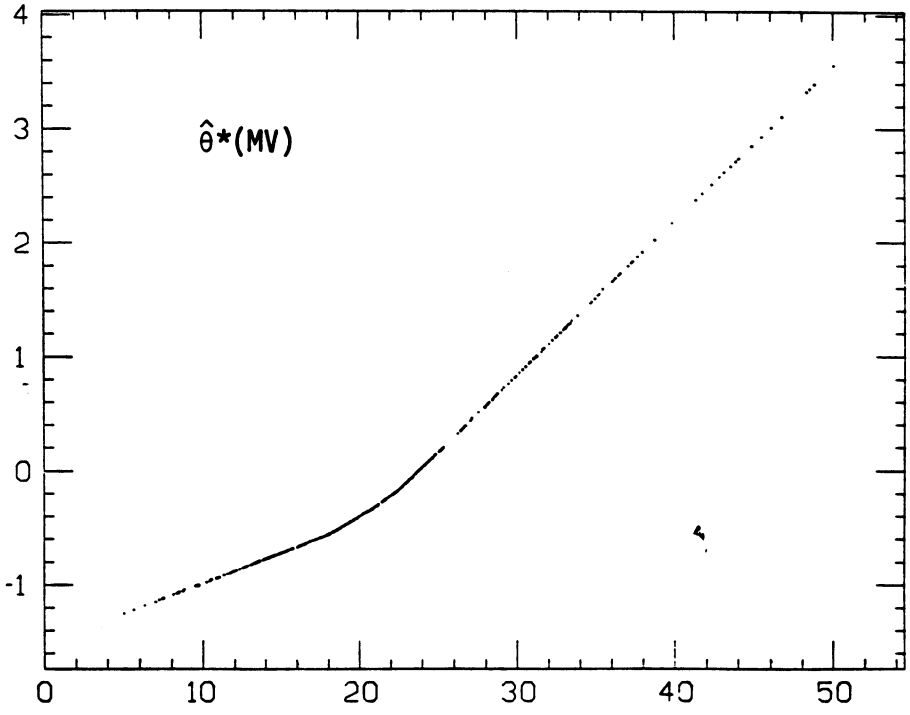
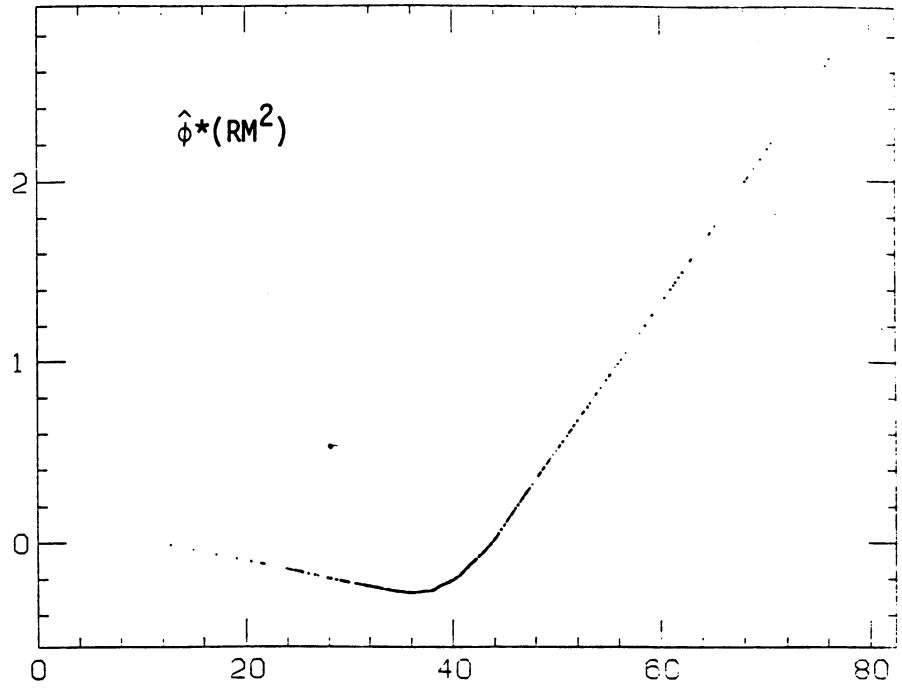
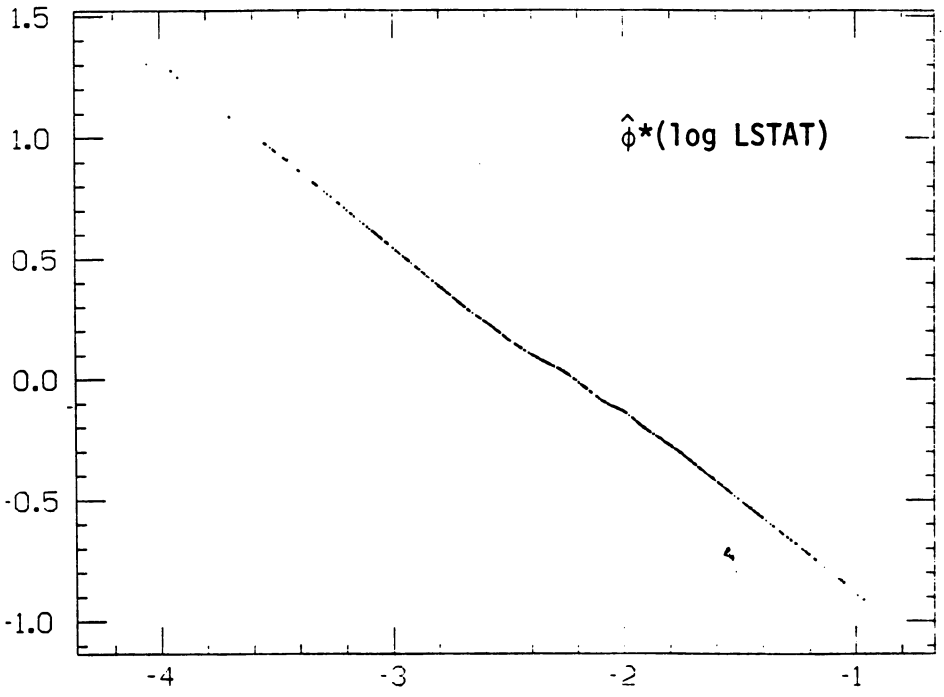


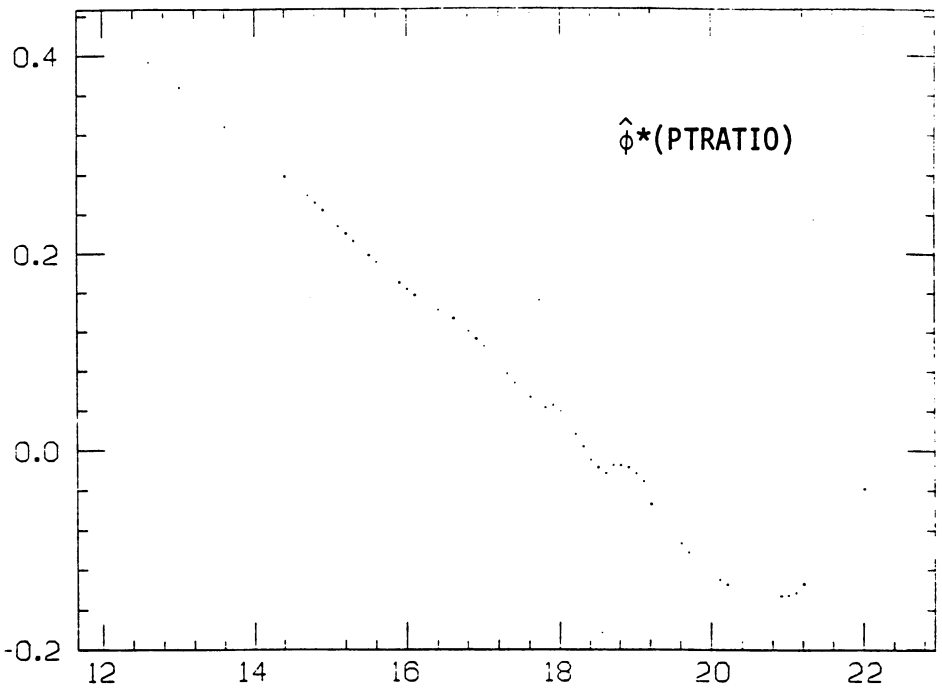
Figure 4b



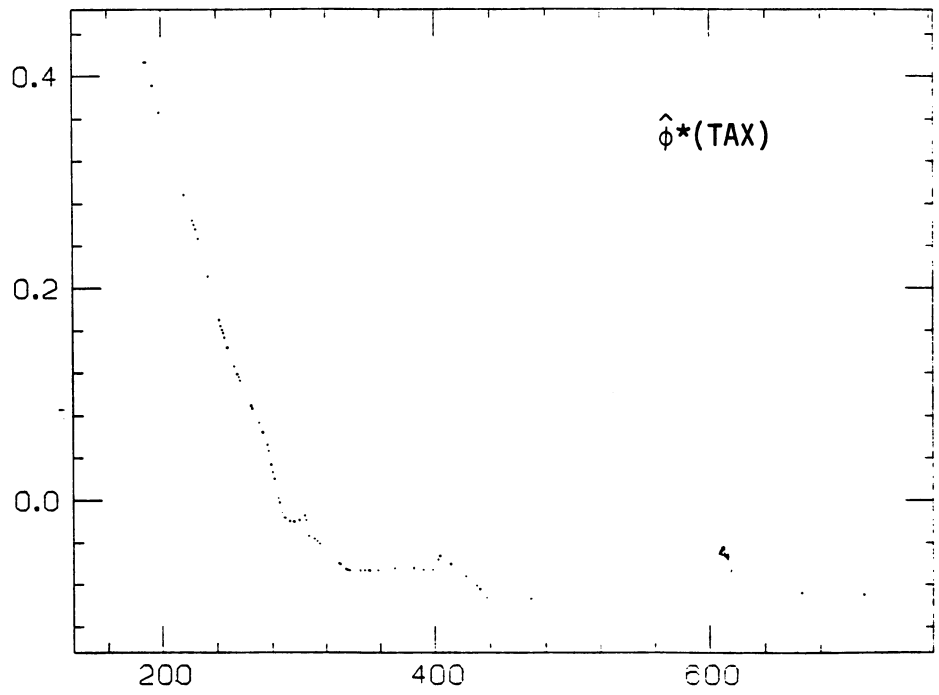
σ = 0.492 **Figure 4c**



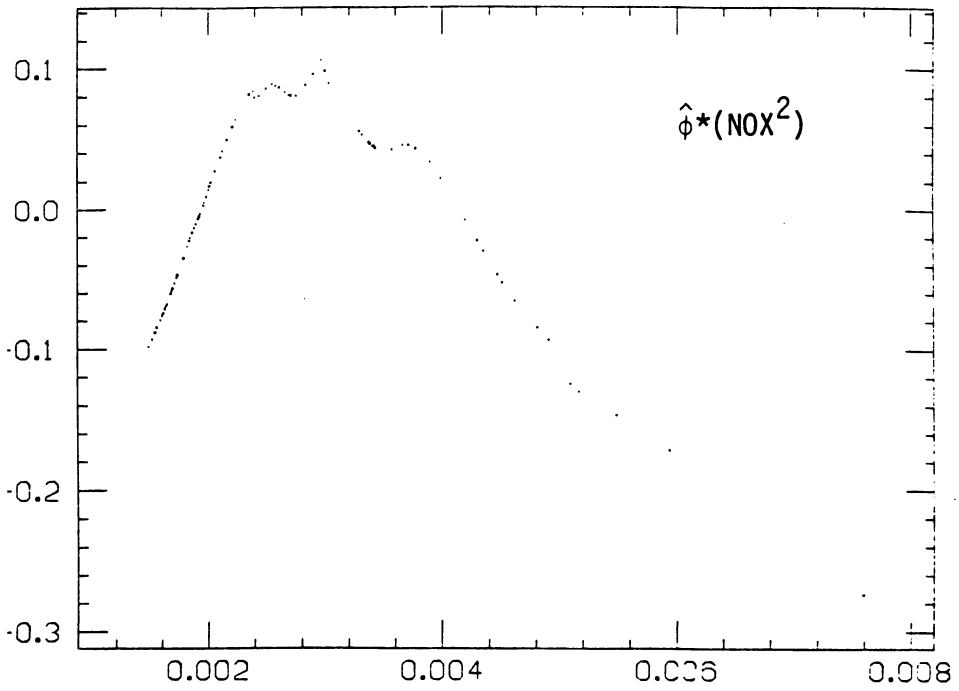
σ = 0.417 **Figure 4d**



σ = 0.147 Figure 4e



σ = 0.122 Figure 4f



σ = 0.09

Figure 4g

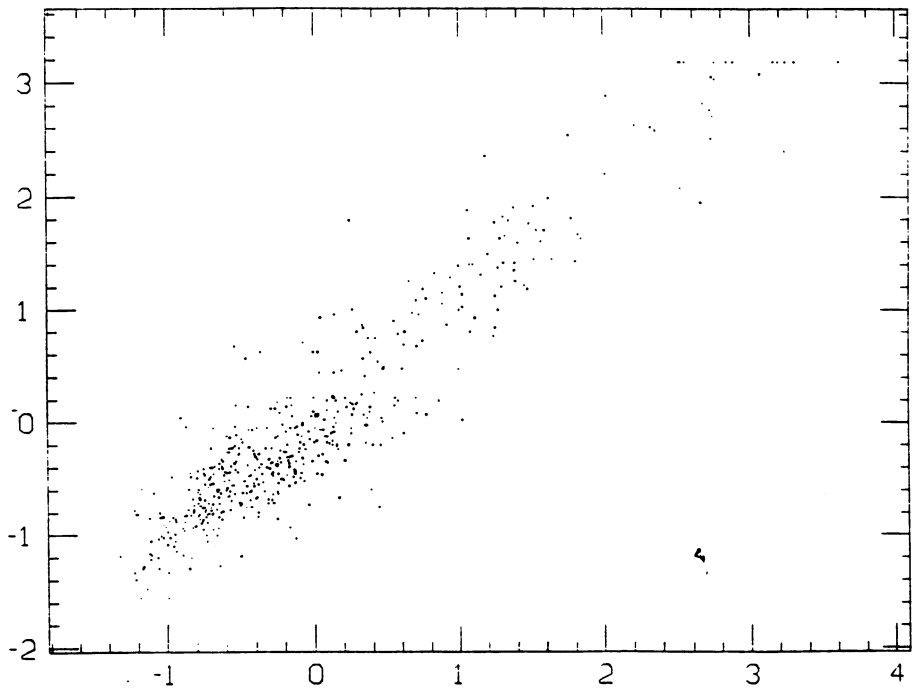


Figure 4h

TABLE 2

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)

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 0.82. Figures 5a-5f show the optimal transformation estimates.

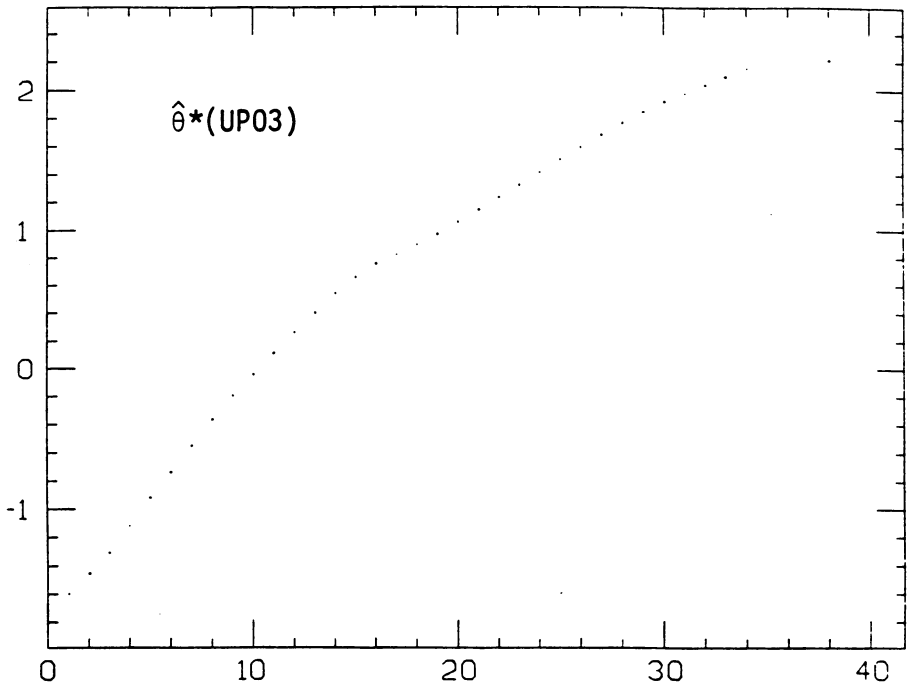
The solution for the response transformation, Figure 5a, shows that, at most, a very mild transformation with negative curvature is indicated. Similarly, Figure 5b indicates that there is no compelling necessity to consider a transformation on the most influential predictor variable, Sandburg Air Force Base Temperature. However, the solution transformation estimates for the remaining variables are all highly nonlinear (and nonmonotonic). For example, Figure 5d 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 5f, 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 5g shows a smooth of ozone concentration on day-of-the-year. This smooth has an $\hat{R}^2$ of 0.38 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 during July to September. This latter fact is confirmed by the day-of-the-year transformation with the meteorological variables

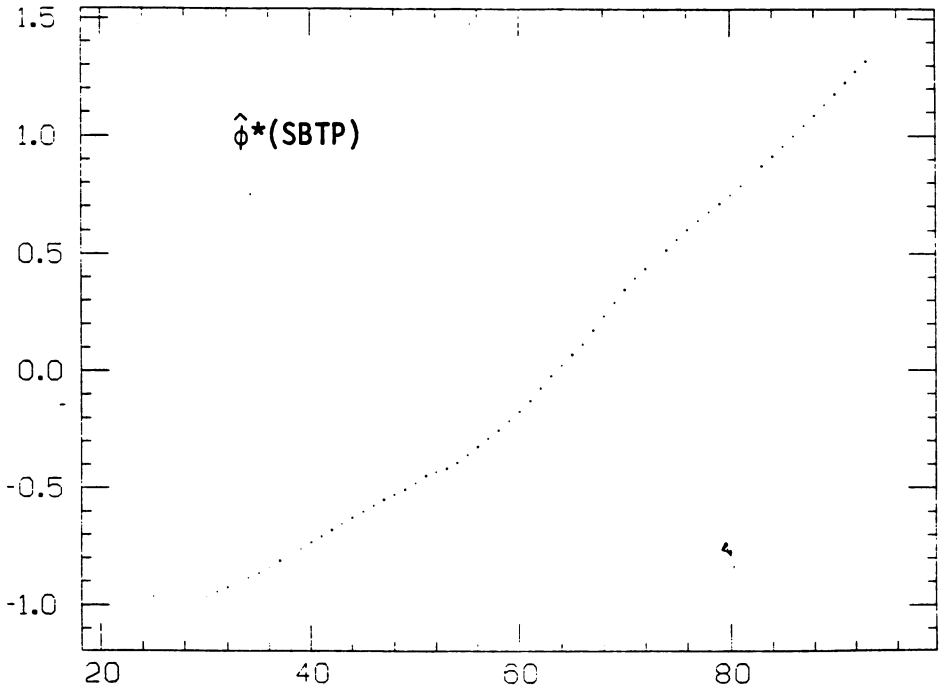
removed. Our current belief is that with the meteorological variables entered, day-of-the-year becomes a partial surrogate for hours of daylight before and during the morning commuter rush. The decline past May 1 may then be explained by the fact that daylight savings time goes into effect in Los Angeles on the last Sunday in April.

This data illustrates 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.



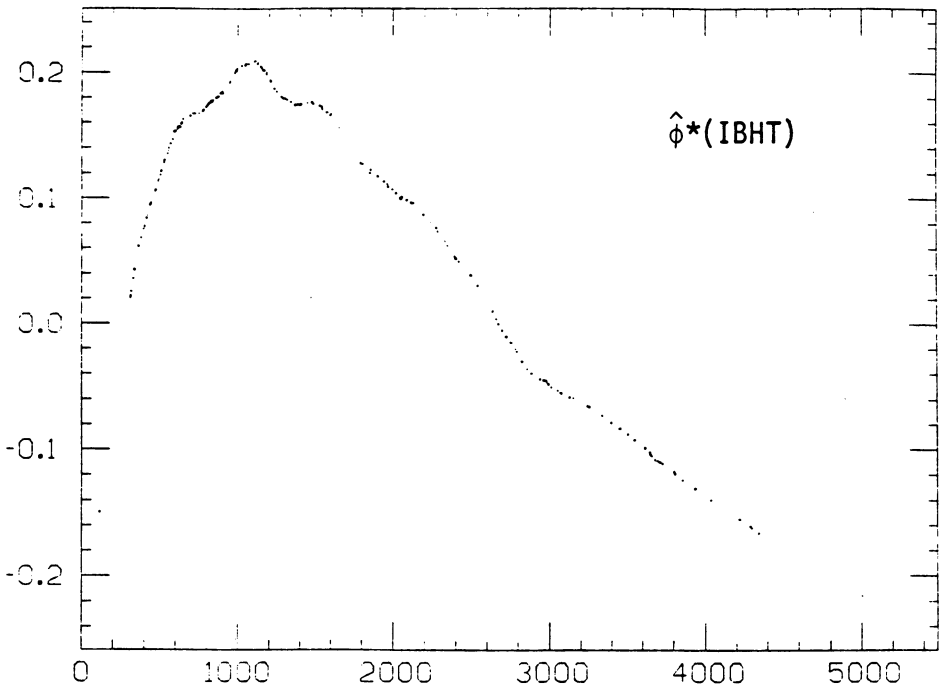
σ = 1.0

Figure 5a



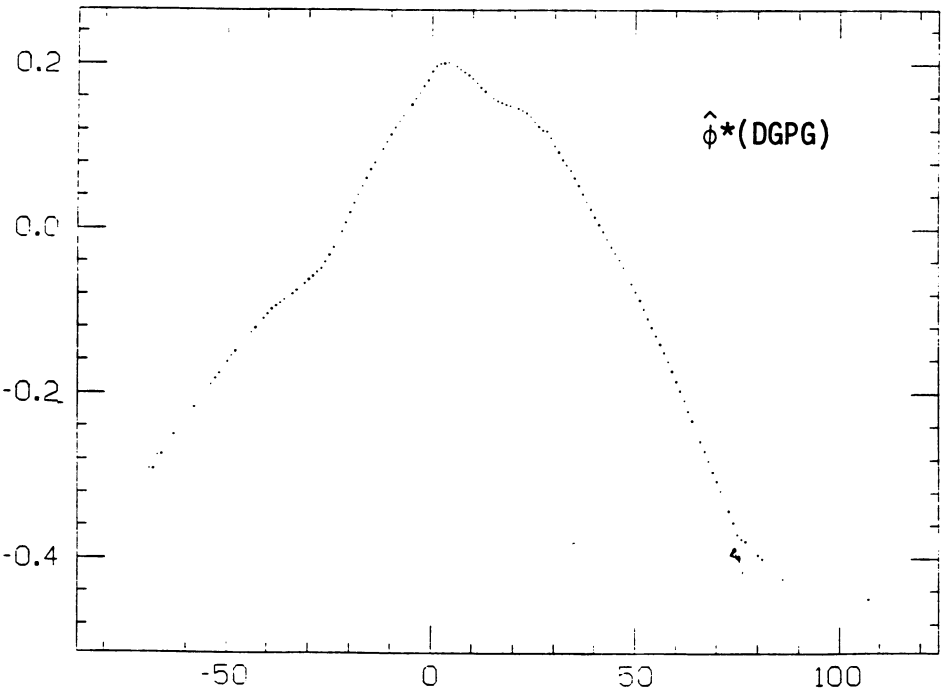
σ = 0.56

Figure 5b



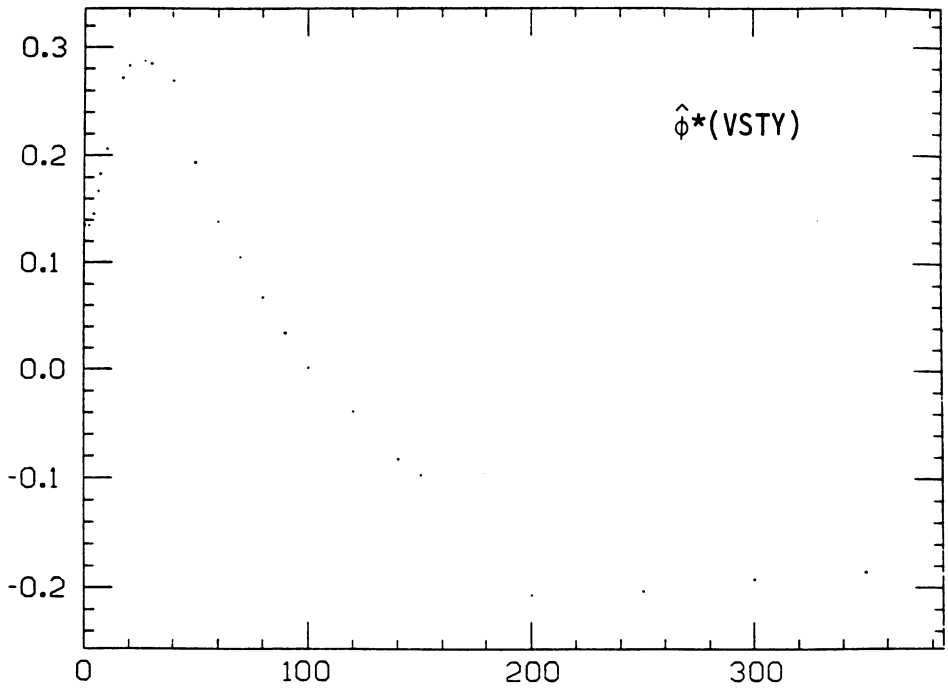
Inversion Base Height, σ = 0.16

Figure 5c



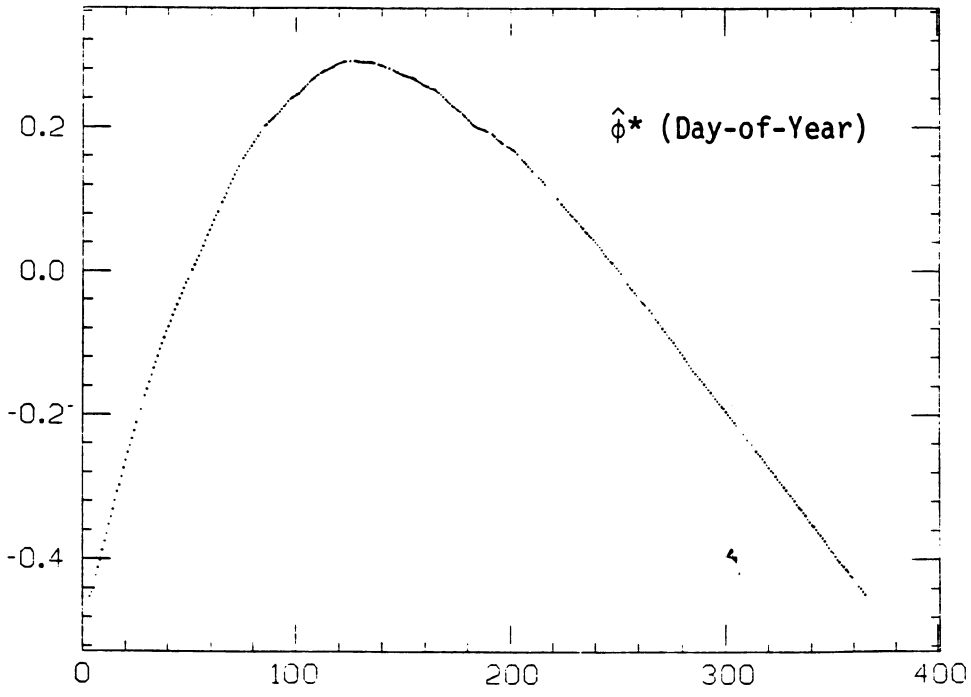
σ = 0.16

Figure 5d



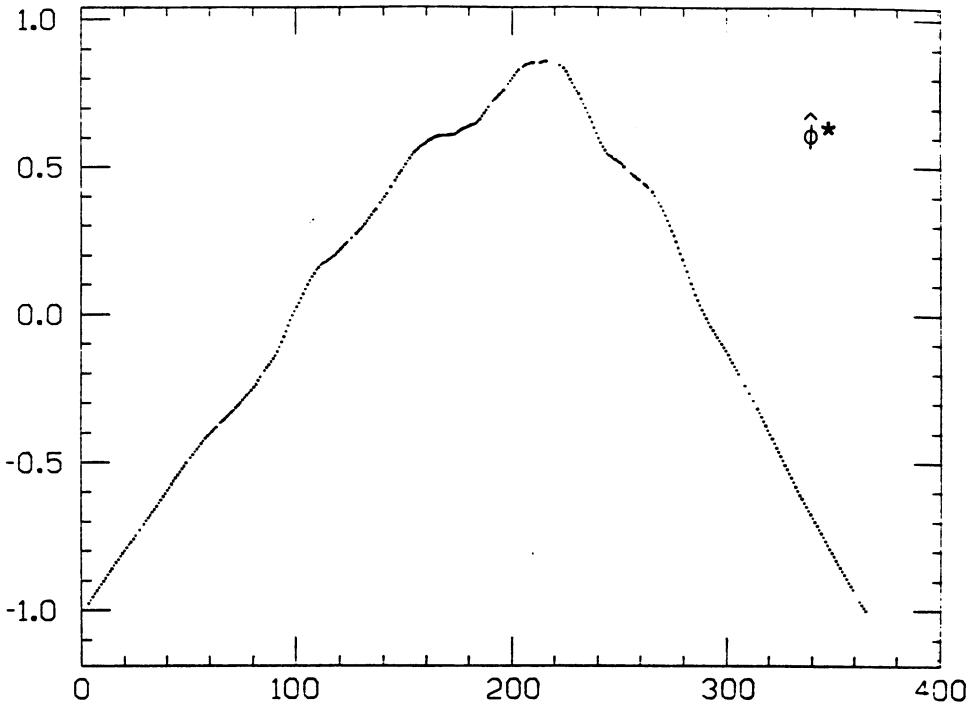
σ = 0.16

Figure 5e



σ = 0.23

Figure 5f



Day-of-the-Year as Single Independent Variable, $\hat{R}^2 = 0.38$

Figure 5g

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]; Linsey [1972]; Box and Hill [1974]; Linsey [1972], [1974]; Wood [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 CPU time on an IBM 3081. Our guess is that this translates into 2.5 minutes on a VAX 11/750 with FP. To extrapolate to other problems, use the estimate that running time is proportional to (number of variables) $\times$ (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 $\underline{x}$. One method for doing this is to attempt to compute $\hat{\theta}^{*-1}\left(\sum_{j=1}^{p} \hat{\phi}_j^*(x_j)\right)$. However, letting

$$Z = \sum_{j=1}^{p} \phi_j^*(X_j) ,$$

we know that the best least squares predictor of Y of the form $\chi(Z)$ is given by $E(Y|Z)$. This is implemented in the current ACE program by predicting y as the function of $\sum_{j=1}^{p} \hat{\phi}_j^*(x_j)$ gotten 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 $(y_k,x_{k1},\ldots,x_{kp})$, $1 \leq k \leq N$. However, since $\theta(y)$ and $\phi(x)$ 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, it can be substituted for the empirical transformation estimate and the ACE algorithm can be rerun using an option which alters only the scale and origin of that particular transformation. The resulting e^2 can be compared to the original value. 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

$$E[\textstyle\sum_{\ell=1}^{q}\theta_\ell(Y_\ell) - \sum_{j=1}^{p}\phi_j(X_j)]^2$$

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

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.

5.0 Optimal Transformations in Function Space

Introduction

In this section, we first prove the existence of optimal transformations (Theorem 5.9). Then we show that the ACE algorithm converges to an optimal transformation (Theorems 5.18 and 5.19).

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 $(\theta, \phi_1, \ldots, \phi_p) = (\theta, \underline{\phi})$, each function defined on the range of the corresponding random variables, such that

$$(5.1) \qquad \begin{aligned} E\theta(Y) &= 0\,, \quad E\phi_j(X_j) = 0\,, \quad j = 1,\ldots,p \\ E\theta^2(Y) &< \infty\,, \quad E\phi_j^2(X_j) < \infty\,, \quad j = 1,\ldots,p \end{aligned}$$

Use the notation

$$(5.2) \qquad \tilde{\phi}(\underline{X}) = \sum_j^p \phi_j(X_j)$$

Denote the set of all transformations by F.

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

$$e^{*2} = E[\theta^*(Y) - \tilde{\phi}^*(\underline{X})]^2 = \inf_F \{E[\theta(Y) - \tilde{\phi}(\underline{X})]^2 ; E\theta^2 = 1\}$$

(5.4) DEFINITION. *A transformation* $(\theta^{**}, \underline{\phi}^{**})$ *is optimal for correlation if* $E(\theta^{**})^2 = 1$, $E(\tilde{\phi}^{**})^2 = 1$,

$$\rho^* = E[\theta^{**}(Y)\tilde{\phi}^{**}(\underline{X})] = \sup_F \{E[\theta(Y)\tilde{\phi}(\underline{X})]\ ;\ E(\tilde{\phi})^2 = 1, E\theta^2 = 1\}$$

(5.5) THEOREM. *If* $(\theta^{**}, \underline{\phi}^{**})$ *is optimal for correlation, then* $\theta^* = \theta^{**}$, $\underline{\phi}^* = \rho^* \underline{\phi}^{**}$ *is optimal for regression and conversely. Furthermore* $e^{*2} = 1 - \rho^{*2}$.

PROOF. Write

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

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

$$E(\theta-\tilde{\phi})^2 \geq 1 - 2\rho^*\sqrt{E\tilde{\phi}^2} + E\tilde{\phi}^2 \tag{5.6}$$

with equality only if $E\theta\hat{\phi} = \rho^*$. The minimum of the right side of (5.6) 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 $(\theta^{**},\underline{\phi}^{**})$ is optimal for correlation, then $\theta^* = \theta^{**}$, $\underline{\phi}^* = \rho^*\underline{\phi}^{**}$ is optimal for regression. The argument is reversible.

5.1 Existence of Optimal Transformations

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

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

$$\theta(Y) + \textstyle\sum_j \Phi_j(X_j) = 0 \ a.s.$$

are individually a.s. zero.

To formulate the second assumption, we use:

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

AII. *The conditional expectation operators*

$$E(\phi_j(X_j)|Y)\colon H_2(X_j) \to H_2(Y)\ ,$$
$$E(\phi_j(X_j)|X_i)\colon H_2(X_j) \to H_2(X_i)\ , \quad i \neq j$$
$$E(\theta(Y)|X_j)\colon H_2(Y) \to H_2(X_j)$$

are all compact.

Condition AII is satisfied in most cases of interest. A sufficient condition is given by: 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

$$\iint [f_{XY}^2/f_X f_Y]dxdy < \infty \tag{5.8}$$

(5.9) THEOREM. *Under AI and AII optimal transformations exist.*

Some machinery is needed.

(5.10) PROPOSITION. *The set of all functions* f *of the form*

$$f(Y,\underline{X}) = \theta(Y) + \textstyle\sum_j \phi_j(X_j)\ , \quad \theta \in H_2(Y),\ \ \phi_j \in H_2(X_j)$$

with the inner product and norm

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

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

$$\tilde{\phi}(\underline{X}) = \textstyle\sum_1^p \phi_j(X_j)\ , \quad \phi_j \in H_2(X_j)$$

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

(5.10) follows from:

(5.11) PROPOSITION. *Under AI, AII there are constants* $0 < c_1 \le c_2 < \infty$ *such that*

$$c_1(\|\theta\|^2 + \textstyle\sum_1^p \|\phi_j\|^2) \le \|\theta + \sum_1^p \phi_j\|^2 \le c_2(\|\theta\|^2 + \sum_1^p \|\phi_j\|^2) .$$

PROOF. The right hand inequality is immediate. If the left side does not hold, we can find a sequence $f_n = \theta_n + \sum \phi_{n,j}$ such that $\|\theta_n\|^2 + \sum_1^p \|\phi_{n\ j}\|^2 = 1$, 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 $H_2(Y), H_2(X_1), \ldots, H_2(X_p)$ respectively.

Write

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

to see that AII implies $E\phi_{n'j}\phi_{n'i} \to E\phi_j\phi_i$, $i \ne j$ and similarly for $E\theta_{n'}\phi_{n',j}$. Furthermore $\|\phi_i\| \le \underline{\lim}\ \|\phi_{n'i}\|$, $\|\theta\| \le \underline{\lim}\ \|\theta_{n'}\|$. Thus, defining $f = \theta + \sum_j \phi_j$

$$\|f\|^2 = \|\theta + \textstyle\sum_j \phi_j\|^2 \le \underline{\lim}\ \|f_{n'}\|^2 = 0$$

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

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

Hence, if $f = 0$, then $\underline{\lim}\ \|f_n\|^2 \ge 1$.

(5.12) COROLLARY. *If* $f_n \xrightarrow{w} f$ *in* H_2, *then* $\theta_n \xrightarrow{w} \theta$ *in* $H_2(Y)$, $\phi_{nj} \xrightarrow{w} \phi_j$ *in* $H_2(X_j)$, $j = 1, \ldots, p$, *and conversely.*

PROOF. If $f_n = \theta_n + \sum_j \phi_{nj} \xrightarrow{w} \theta + \sum_j \phi_j$, then by (5.11), $\overline{\lim}\ \|\theta_n\| < \infty$, $\overline{\lim}\ \|\phi_{nj}\| < \infty$. Take n' such that $\theta_{n'} \xrightarrow{w} \theta'$, $\phi_{n'j} \xrightarrow{w} \phi_j'$, and let

$f' = \theta' + \sum_j \phi'_j$. Then for any $g \in H_2$, $(g,f_{n'}) \to (g',f')$, so $(g,f) = (g,f')$ all g. The converse is easier.

(5.13) DEFINITION. *In* H_2, *let* P_Y, P_j *and* P_X *denote the projection operators on* $H_2(Y)$, $H_2(X_j)$ *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.

(5.14) PROPOSITION. 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 (5.12), that $\phi_{n,j} \xrightarrow{w} \phi_j$. By AII, $P_Y\phi_{n,j} \xrightarrow{s} P_Y\phi_j$ 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, $P_X\colon H_2(Y) \to H_2(X)$ is the adjoint of P_Y and hence compact.

Now to complete the proof of Theorem 5.9. 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^* which 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, $\|P_X\theta_n\| \to \|P_X\theta\| = \bar{s}$. Further, $\|\theta\| \leq 1$. 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.

5.2 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

$$U\theta = P_Y P_X \theta \ , \quad V\tilde{\phi} = P_X P_Y \tilde{\phi}$$

(5.15) PROPOSITION. *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 eigenvalue specified by*

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

PROOF. Direct verification.

Let the largest eigenvalue be denoted by $\bar{\lambda}$, $\bar{\lambda} = \|U\| = \|V\|$. Then

(5.16) THEOREM. *If θ^*, $\tilde{\phi}^*$ is an optimal transformation for regression, then*

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

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 $\theta^* = eP_Y\tilde{\phi}^*$, e constant. Therefore, $\theta^* = P_Y\tilde{\phi}^* / \|P_Y\tilde{\phi}^*\|$. This implies

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

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.

(5.17) COROLLARY. *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.*

5.3 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 $\bar{\lambda}\theta = U\theta$ 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 $P_X\theta = E(\theta|X)$, $P_Y\phi = E(\phi|Y)$. 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 $\phi_{n+1} = P_X\theta_n$, $\theta_{n+1} = P_Y\phi_{n+1}/\|P_Y\phi_{n+1}\|$. 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 above method of alternating conditionals extends to the general multivariate case. The analogue is clear; given θ_n, $\tilde{\phi}_n$, then the next iteration is

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

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 Outer Loop

1. Start with an initial guess $\theta_0(Y)$.
2. 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 $\bar{\lambda}$. Then

(5.18) THEOREM. *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, $\theta_n = \alpha_n\theta^* + g_n$, where $g_n \perp E$. Because, if it is true for n, then

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

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

$$\|g_{n+1}\|/\alpha_{n+1} = \|Ug_n\|/\bar{\lambda}\alpha_n \leq (r/\bar{\lambda})\|g_n\|/\alpha_n .$$

Thus $\|g_n\|/\alpha_n \leq c(r/\bar{\lambda})^n$. But $\|\theta_n\| = 1$, $\alpha_n^2 + \|g_n\|^2 = 1$, implying $\alpha_n^2 \to 1$. Since $\alpha_0 > 0$, then $\alpha_n > 0$, so $\alpha_n \to 1$. Now use $\|\theta_n - \theta^*\|^2 = (1-\alpha_n)^2 + \|g_n\|^2$ to reach the conclusion. Since $\|\tilde{\phi}_{n+1} - \tilde{\phi}^*\| = \|P_X\theta_n - P_X\theta^*\| \leq \|\theta_n - \theta^*\|$, the theorem follows.

The Inner Loop

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

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

(5.19) THEOREM. *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

$$\theta - \tilde{\phi}_{m+1} = T(\theta - \tilde{\phi}_m) = T^{m+1}(\theta - \tilde{\phi}_0) \tag{5.20}$$

Write $\theta - \tilde{\phi}_0 = \theta - P_X\theta + P_X\theta - \tilde{\phi}_0$. Noting that $T(\theta - P_X\theta) = \theta - P_X\theta$, (5.20) becomes

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

The theorem is then proven by

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

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

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

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

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

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

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

Put $\gamma_0 = \|W\|$, $\gamma_m = G(\gamma_{m-1})\gamma_{m-1}$, then $\|T^mW\| \leq \gamma_m$. But clearly $\gamma_m \to 0$.

The range of W is dense in $H_2(X)$. Otherwise, there is a $\tilde{\phi}' \neq 0$ 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 above leads to $\tilde{\phi}' = 0$. For any $\tilde{\phi}$ and $\varepsilon > 0$, take $W\tilde{\phi}_1$ so that $\|\tilde{\phi}-W\tilde{\phi}_1\| \leq \varepsilon$. Then $\|T^m\tilde{\phi}\| \leq \varepsilon + \|T^mW\tilde{\phi}_1\|$, 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 fresh start. Restart is a faster running algorithm, and 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

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.

6.0 The ACE Algorithm on Finite Data Sets

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 very restricted class of smooths. In practice, in over 1000 runs of ACE over 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.

6.1 Data Smooths

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

(6.1) DEFINITION. *A data smooth* S *of* $\underline{x}$ *on* x_j *is a mapping* $S: F(\underline{x}) \rightarrow F(x_j)$ *defined for every* D *in* $\mathcal{D}_N$. *If* $\phi \in F(\underline{x})$ *denote the corresponding element in* $F(x_j)$ *by* $S(\phi|x_j)$ *and its values by* $S(\phi|x_{kj})$.

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

1. Histogram: Divide the real axis up into disjoint intervals $\{I_\ell\}$. If $x_k \in I_\ell$, define

$$S(\phi|x_k) = \frac{1}{n_\ell} \sum_{\underline{x}_m;\ x_k \in I_\ell} \phi(\underline{x}_m)$$

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

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

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

$$S(\phi|x_k) = \sum_m \phi(\underline{x}_m)K(x_m - x_k) \Big/ \sum_\ell K(x_\ell - x_k)$$

4. Regression: Fix M and order x_k as in (2) above. At x_k, regress the values of $\phi(\underline{x}_{k-M}),\ldots,\phi(\underline{x}_{k+M})$ excluding $\phi(\underline{x}_k)$ on $x_{k-M},\ldots,x_{k+M}$ excluding x_k, getting a regression line $L(x)$. Put $S(\phi|x_k) = L(x_k)$. 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 below. These properties hold only if they are true for all $D \in \mathcal{D}_n$.

Linearity. A smooth is linear if

$$S(\alpha\phi_1 + \beta\phi_2) = \alpha S\phi_1 + \beta S\phi_2$$

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

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

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

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

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

Boundedness. S is bounded by M if

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

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

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

6.2 Convergence of ACE

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

$$S_j\phi = S_j\phi - \overline{S_j\phi} \tag{6.2}$$

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

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

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

   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 $H_2(\theta,\underline{\phi})$ 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 $(\theta,0,0,\ldots,0) = \underline{\theta}$ and $H_2(\underline{\phi})$ with elements $(0,\phi_1,\ldots,\phi_p) = \underline{\phi}$.

For $f = (f_0,f_1,\ldots,f_p)$ in $H_2(\theta,\underline{\phi})$ define $\hat{S}_j: H_2(\theta,\underline{\phi}) \rightarrow H_2(\theta,\underline{\phi})$ by

$$(S_j f)_i = \begin{cases} 0, & j \neq i \\ f_j + S_j(\sum_{i\neq j} f_i), & j = i \end{cases}$$

Starting with $\underline{\theta} = (\theta,0,0,\ldots,0)$, $\underline{\phi}^{(m)} = (0,\phi_1^{(m)},\ldots,\phi_p^{(m)})$ one complete cycle in the inner loop is described by

$$\underline{\theta} - \underline{\phi}^{(m+1)} = (I-S_p)(I-S_{p-1})\cdots(I-S_1)(\underline{\theta}-\underline{\phi}^{(m)}) \quad . \tag{6.3}$$

Define $\hat{T}$ on $H_2(\theta,\underline{\phi}) \rightarrow H_2(\theta,\underline{\phi})$ as the product operator in (6.3). Then

$$\underline{\phi}^{(m)} = \underline{\theta} - \hat{T}^m(\underline{\theta}-\underline{\phi}^{(0)}) \quad . \tag{6.4}$$

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

$$S_j(\underline{\theta}-\underline{\phi}) = 0 \quad , \quad j=1,\ldots,p \quad . \tag{6.5a}$$

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

Adding

$$\underline{\theta} = S_y\underline{\phi}/\|S_y\underline{\phi}\|_N \tag{6.5b}$$

to (6.5a) gives a set of equations satisfied by the estimated optimal transformations.

Assume, for the remainder of this section, that the smooths are linear. Then (6.5a) can be written as

$$S_j\underline{\phi} = S_j\underline{\theta} \quad , \quad j=1,\ldots,p \quad . \tag{6.6}$$

Let $sp(S_j)$ denote the spectrum of the matrix S_j. Assume $1 \notin sp(S_j)$. (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 $-1 \notin sp(A)$. Then (6.6) has the unique solution

$$\phi_j = A_j(I+A)^{-1}\theta \quad , \quad j=1,\ldots,p \tag{6.7}$$

The element $\underline{\phi} = (0,\phi_1,\ldots,\phi_p)$ given by (6.7) will be denoted by $\hat{P}\underline{\theta}$. Rewrite (6.3) using $(I-\hat{T})(\underline{\theta}-\hat{P}\underline{\theta}) = 0$ as

$$\underline{\phi}^{(m)} = \hat{P}\underline{\theta} - \hat{T}^m(\hat{P}\theta-\underline{\phi}^{(0)}) \tag{6.8}$$

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

(6.9) THEOREM. *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* $f \in H_2(\underline{\phi})$ *is that*

$$\det[\lambda I - \prod_1^p (I-S_j/\lambda)^{-1}(I-S_j)] \tag{6.10}$$

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

PROOF. For $\hat{T}^m f \to 0$, all $f \in H_2(\underline{\phi})$, 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

$$\lambda f_j = -S_j(\lambda \sum_{i<j} f_i + \sum_{i>j} f_i) , \quad j = 1,\ldots,p . \tag{6.11}$$

Let $s = \sum_i f_i$ and rewrite (6.11) as

$$(\lambda I-S_j)f_j = S_j((1-\lambda)\sum_{i<j} f_i - s) . \tag{6.12}$$

If $\lambda = 1$, (6.12) becomes $(I-S_j)f_j = -S_j s$ or $s = -As$. By assumption, this implies $s = 0$, and hence $f_j = 0$, 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

$$g_{j+1} = (I-S_j/\lambda)^{-1}(I-S_j)g_j \ . \tag{6.13}$$

Since $g_{p+1} = -\lambda s$, $g_1 = -s$, then (6.13) leads to

$$\lambda s = (I-S_p/\lambda)^{-1}(I-S_p)\cdots(I-S_1/\lambda)^{-1}(I-S_1)s \tag{6.14}$$

If (6.14) has no non-zero solutions, then $s = 0$, $g_j = 0$, $j = 1,\ldots,p$, implying all $f_j = 0$. Conversely, if (6.14) has a solution $s \neq 0$, it leads to a solution of (6.11).

Unfortunately, condition (6.10) 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 (6.10) 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 $\hat{P}\theta$, then the outer loop iteration is given by

$$\underline{\theta}^{(n+1)} = \frac{S_y\hat{P}\underline{\theta}^{(n)}}{\|S_y\hat{P}\underline{\theta}^{(n)}\|_N}$$

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

$$\theta^{(n+1)} = \frac{\hat{U}\theta^{(n)}}{\|\hat{U}\theta^{(n)}\|_N} \tag{6.15}$$

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 (6.5a,b). However, 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 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 above 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 above arguments.

However, ACE 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.

6.3 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 $P_j(\phi_i|x_j)$ by

$$P_j(\phi_i|x_{kj}) = E(\phi_i(X_i)|X_j = x_{kj}) \tag{6.16}$$

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 $(y_k,\underline{x}_k)$ are samples from the distribution of $(Y,X_1,\ldots,X_p)$, not necessarily independent or even random (see Section 6.4).

(6.17) DEFINITION. *Let* $S_y^{(N)}$, $S_j^{(N)}$ *be any sequence of data smooths. They are mean square consistent if*

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

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

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

Then set

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

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

(6.18) THEOREM. *For the fresh start algorithm, if the smooths* $S_y^{(N)}, S_j^{(N)}$ *are m.s. 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*

$$E\|\theta_N(y;m,\ell)-\theta(y;,m,\ell)\|_N^2 \to 0\ , \quad E\|\phi_{jN}(x_j;m,\ell)-\phi_j(x_j;m,\ell)\|_N^2 \to 0$$

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

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

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 $E\|S_i^{(N)}S_j^{(N)}\theta_0 - P_iP_j\theta_0\|_N^2 \to 0$.

(5.19) PROPOSITION. *If* θ_N *is defined in* $H_2(y)$ *for all data sets* D, *and* $\theta \in H_2(Y)$ *such that*

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

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, that $E\|\theta(\frac{1}{\|\theta\|} - \frac{1}{\|\theta\|_N})\|_N^2 \to 0$. 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 $S_N^2 \leq 4$, so it is enough to show that $S_N^2 \xrightarrow{P} 0$ to get $ES_N^2 \to 0$.

Let

$$V_N^2 = \frac{1}{N}\sum_k (\theta_N(y_k)-\theta(y_k))^2$$

$$= \|\theta_N\|_N^2 + \|\theta\|_N^2 - 2(\theta_N,\theta)_N$$

$$= (\|\theta_N\|_N - \|\theta\|_N)^2 + 2(\|\theta\|_N\|\theta_N\|_N - (\theta_N,\theta)_N)$$

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

Now look at

$$W_N^2 = \frac{1}{N}\sum_k \theta^2(y_k)\left[\frac{1}{\|\theta\|_N} - \frac{1}{\|\theta\|}\right]^2$$

$$= \|\theta\|_N^2\left(\frac{1}{\|\theta\|_N} - \frac{1}{\|\theta\|}\right)^2$$

$$= \left(1 - \frac{\|\theta\|_N}{\|\theta\|}\right)^2$$

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

Using Proposition 6.19 it follows that $E\|\theta_N(y;m,\ell)-\theta(y;m,\ell)\|_N^2 \to 0$ and in consequence, that $E\|\phi_{j,N}(x_j;m,\ell)-\phi_j(x_j;m,\ell)\|^2 \to 0$.

In function space, define

$$P_X^{(m)}\theta = \theta - T^m\theta$$

$$U_m = P_Y P_X^{(m)}$$

Then

$$\theta(\cdot;m,\ell) = \frac{U_m^\ell \theta_0}{\|U_m^\ell \theta_0\|}.$$

The last step in the proof is showing that

$$\left\| \frac{U_m^\ell \theta_0}{\|U_m^\ell \theta_0\|} - \theta^* \right\| \to 0$$

as m, ℓ go to infinity. Begin with

(6.20) PROPOSITION. *As* $m \to \infty$, $U_m \to U$ *in the uniform operator norm.*

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

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

By Proposition (5.21) the right hand side goes to zero.

The operator U_m is not necessarily self-adjoint, but it is compact. By (6.20), if $O(sp(U))$ is any open set containing $sp(U)$, then for m sufficiently large $sp(U_m) \subset O(sp(U))$. 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 is more complicated.) Then for any open neighborhood O of $\bar\lambda$, and m sufficiently large, there is only one eigenvalue λ_m of U_m in O, $\lambda_m \to \bar\lambda$ and the projection $P_E^{(m)}$ of U_m corresponding to λ_m converges to $P_{E_{\bar\lambda}}$ in the uniform operator topology. Also, 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' the eigenvalue of U_m having the second highest absolute value, then (assuming $E_{\lambda'}$ is one-dimensional) $\lambda_m' \to \lambda'$.

Write

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

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

(6.21)
$$U_m^\ell \theta_0 = \lambda_m^\ell P_E^{(m)} \theta_0 + W_m^\ell \theta_0$$
$$U^\ell \theta_0 = \lambda^\ell P_{E_{\bar\lambda}} \theta_0 + W^\ell \theta_0$$

For any $\varepsilon > 0$ we will show that there exists m_0, ℓ_0 such that for $m \geq m_0$, $\ell \geq \ell_0$

(6.22)
$$\|W_m^\ell \theta_0\| / \lambda_m^\ell \leq \varepsilon \ , \qquad \|W^\ell \theta_0\| / \bar\lambda^\ell \leq \varepsilon \ .$$

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

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

and

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

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

$$\|W_m^\ell\| \leq r^\ell M(r)(1 + \Delta_m)$$

where $\Delta_m \longrightarrow 0$ as $m \longrightarrow \infty$. Certainly

$$\|W^\ell\| \leq r^\ell M(r) \ .$$

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

$$\|W_m^\ell\| / \lambda_m^\ell \leq \left(\frac{1}{1+\delta}\right)^\ell M(r)(1+\Delta_m)$$

and

$$\|W^\ell\| / \bar\lambda^\ell \leq \left(\frac{1}{1+\delta}\right)^\ell M(r) \ .$$

Now choose a new m_0 and ℓ_0 such that (6.22)is satisfied.

Using (6.22)

$$\left\| \frac{U_m^{\ell}\theta_0}{\|U_m^{\ell}\theta_0\|} - \frac{P_E^{(m)}\theta_0}{\|P_E^{(m)}\theta_0\|} \right\| = \varepsilon_{m,\ell}$$

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

$$\left\| \frac{U_m^{\ell}\theta_0}{\|U_m^{\ell}\theta_0\|} - \theta^* \right\| = \varepsilon'_{m,\ell} + \left\| \frac{P_{E_m}\theta_0}{\|P_{E_m}\theta_0\|} - \frac{P_{E_{\bar{\lambda}}}\theta_0}{\|P_{E_{\bar{\lambda}}}\theta_0\|} \right\|$$

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

The term weak consistency is used above because we have in mind a desirable stronger result. We conjecture that for reasonable smooths, the set $C_N = \{(Y_1,\underline{X}_1),\ldots,(Y_N,\underline{X}_N);$ 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 nonnegative definite smooths with nonnegative matrix elements. Then we know that the algorithm converges to some θ_N, and we conjecture that

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

6.4 The M.S. 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 (6.18) can be satisfied. We do this first assuming that the data $(Y_1,X_1),\ldots,(Y_N,X_N)$ are samples from a 2-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, $E\|\theta\|_N^2 \longrightarrow \|\theta\|^2$.

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

(6.23) THEOREM. *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 an m.s. consistent sequence of nearest neighbor smooths of* Y *on* X.

The proof begins with;

(6.24) LEMMA. *Suppose that* $P(dx)$ *has no atoms and let* $P_N(dx) \xrightarrow{w} P(dx)$. *Take* $\delta_N > 0$, $\delta_N \longrightarrow \delta > 0$, *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*

(1) $P_N(J(x,\varepsilon_N(x)) \Delta J(x,\varepsilon(x))) \longrightarrow 0$ *uniformly in* x.

(2) $$\overline{\lim_N} \sup_{\{(x,y);|x-y|\leq h\}} P_N(J(x,\varepsilon(x)) \Delta J(y,\varepsilon(y))) \leq \varepsilon_1(h)$$

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

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

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

To prove (1) note that

$$P_N(J(x,\varepsilon_N) \Delta J(x,\varepsilon)) \leq |P_N(J(x,\varepsilon_N)) - P_N(J(x,\varepsilon))| \leq$$

$$|\delta_N - P_N(J(x,\varepsilon_N))| + |\delta_N - \delta| + |F_N(x+\varepsilon(x)) - F(x+\varepsilon(x))| + |F_N(x-\varepsilon(x)) - F_N(x-\varepsilon(x))|$$

which does it. To prove (2) 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 $\varepsilon(x_n) \longrightarrow 0$ and $P(J(x_n,\varepsilon(x_n)) = \delta$, which is impossible since P has no atoms. Then for $h \leq$ ho

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

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

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

$$\frac{1}{\delta} \int g(y)\ I(x' \in J(x,\varepsilon(x))\ P(dy,dx') =$$

$$\frac{1}{\delta} \int P_X(g|x')\ I(x' \in J(x,\varepsilon(x))\ P(dx')$$

Note that P_δ is bounded and continuous in x . Denote by $S_\delta^{(N)}$ the smooths

with $M = [N\delta]$. Then

(6.25) PROPOSITION. $E\|S_\delta^{(N)} g - P_\delta g\|_N^2 \longrightarrow 0$ *for fixed* δ.

proof. By (1) of the lemma, with probability one

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

can be replaced for all x by

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

where ω is a sample sequence.

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

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

Use (2) of the lemma to establish that for any bounded interval J and any $\omega \in W'$, $\Phi_N(x,\omega) \longrightarrow 0$ uniformly for $x \in J$. Then write

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

The 1st term is bounded and goes to zero for $\omega \in W'$, hence its expectation goes to zero. The expectation of the 2nd 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

$$\overline{\lim}\; E\|S_\delta^{(N)} g - P_X g\|_N^2 \leq 2\|P_\delta g - P_X g\|^2$$

(6.26) PROPOSITION. *For any* $\phi(x) \in L_2(X)$, $\lim_{\delta \to 0} \|P_\delta \phi - \phi\| \longrightarrow 0$.

proof. For ϕ bounded and continuous

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

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

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

Suppose that x' is such that there are numbers $\varepsilon^+, \varepsilon^-$ 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

$$(6.27)\quad \frac{1}{\delta}\int I(x' \in J(x,\varepsilon(x)))P(dx) \leq 2 .$$

If, say, $P([x',\infty)) < \delta$, then $x \geq x' - \varepsilon^-$ and (6.27) 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 (6.25) and (6.26), for any $\varepsilon > 0$, we can select $\delta(\varepsilon,n), N(\delta,n)$ so that for all n

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

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

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

Put $M(N) = \max\{M;\ N \geq \max(M,N(M))\}$. Then $M(N) \longrightarrow \infty$ as $N \longrightarrow \infty$ and the sequence of smooths $S_{\delta_{M(N)}}^{(N)}$ is m.s. 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 $S_\delta^{(N)}g = S_\delta^{(N)}g - \overline{S_\delta^{(N)}g}$

and functions g such that $Eg = 0$ causes no problems, since

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

and

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

using the notation of Proposition 6.25 .

Assume g is bounded and write

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

By the ergodic theorem, the 2nd term goes a.s. to $EP_\delta(g|X)$, and an argument mimicking the proof of 6.25 shows that the first term goes to zero a.s.

Finally, write

$$|EP_\delta(g|X)| = |EP_\delta(g|X) - EP_X g| \leq \|P_\delta\phi - \phi\|$$

where $\phi = P_X g$. Thus, Theorem 6.23 can be easily changed to account for modified smooths.

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

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

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

A necessary result is;

(6.28) PROPOSITION. *For* $\theta(y)$ *bounded in* $L_2(Y)$ *and* $\phi(\underline{x})$ *bounded and continuous*

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

Let $T_N = \sum_{j=1}^{N} \theta(Y_j)\phi(\underline{x}_j)$. Then $ET_N = \sum_{j=1}^{N} g(\underline{x}_j)\phi(\underline{x}_j)$, $g(\underline{x}) = E[\theta(Y)|\underline{X}=x]$.

By hypothesis, $ET_N/N \longrightarrow E\theta(Y)\phi(X)$. Further

$$\delta_N^2 = Var(T_N) = \sum_{j=1}^{N} E[\theta(y_j)-g(\underline{x}_j)]^2 \phi(\underline{x}_j)$$

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

where $h(\underline{x}) = E[(\theta(Y) - g(\underline{x}))^2 | \underline{X} = \underline{x}]$. Since $h\phi$ is continuous and bounded, then $\delta_N^2/N \longrightarrow Eh(\underline{X})\phi(\underline{X})$. Now the application of Kolmogorov's exponential bound (see Loeve, 1960, pp.254-255) gives

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

proving the proposition.

In the consistency theorem (6.18) we add the restriction that θ_0 be a bounded function in $L_2(Y)$. Then the condition on θ may be relaxed to : *For* θ *any bounded function in* $L_2(Y)$, $\|\theta\|_N^2 \xrightarrow{P} \|\theta\|^2$, $E\|\theta\|_N^2 \longrightarrow \|\theta\|^2$. These follow from (6.28) and its proof. Further, because of the H1,H2, m.s. consistency of the smooths can be relaxed to the requirements that;

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

$$\| S_j\phi - P_j\phi \|_N^2 \longrightarrow 0$$

(H.3,ii) For every bounded function $\theta(y) \in L_2(Y)$

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

(H.3,iii) For every bounded continuous function $\phi(x_i)$

$$E\| S_y\phi - P_y\phi \|_N^2 \longrightarrow 0$$

The existence of sequences of nearest neighbor smooths satisfying (A.3) can be proven in a fashion very similar to the proof of Theorem 6.23. (H.3,i) is proven using Lemma 6.24 and Proposition 6.26. (H.3,ii) and (H.3,iii) require Proposition 6.28 in addition.

If the data is i.i.d. stronger results can be gotten. For instance m.s. consistency can be proven for a modified regression smooth similar to supersmoother. For x any point, let $J(x)$ be the indices 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

$$S(\phi|x) = \bar{\phi}_x + \frac{\Gamma_x(\phi,x)}{\sigma_x^2}(x-\bar{x}_x) \tag{6.29}$$

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

Write the second term in (6.30) as

$$\frac{\Gamma_x(\phi,x)}{\sigma_x}\,\frac{(x-\bar{x}_x)}{\sigma_x}\,.$$

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

$$\left|\frac{x-\bar{x}_x}{\sigma_x}\right| \leq 1.$$

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

$$[x]_t = \begin{cases} x, & |x| \leq 1 \\ \text{sgn}(x), & |x| > 1, \end{cases}$$

and define the *modified regression smooth* by

$$S(\phi|x) = \bar{\phi}_x + \frac{\Gamma_x(\phi,x)}{\sigma_x}\left[\frac{x-\bar{x}_x}{\sigma_x}\right]_t \tag{6.31}$$

This modified smooth is bounded by 2.

(6.32) THEOREM. *If, as* $N \longrightarrow \infty$, $M \longrightarrow \infty$, $M/N \longrightarrow 0$ *and* $P(dx)$ *has no atoms, then the modified regression smooths are m.s. consistent.*

The proof is in Breiman and Friedman (1983). We are almost certain that the modified regression smooths are also m.s. consistent for stationary ergodic time series and in the weaker sense for controlled experiments, but under less definitive conditions on rate at which $M \longrightarrow \infty$.

References

Anscombe, F.J. and Tukey, J.W. (1963). The examination and analysis of residuals. *Technometrics* 5, 141-160.

Belsey, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics*, John Wiley and Sons.

Box, G.E.P. and Tidwell, P.W. (1962). Transformations of the independent variables. *Technometrics* 4, 531-550.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *J.R. Statist. Soc. 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.

Breiman, L. and Friedman, J. (July 1982). Estimating optimal transformations for multiple regression and correlation. Dept. of Statistics, University of California, Berkeley, Tech. Report. #9.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* 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.

deBoor, C. (1978). *A Practical Guide to Splines*, Springer-Verlag.

de Leeuw, J., Young, F.W., Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 1976, 471-503.

Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.* 9, 1310-1319.

Devroye, L. and Wagner, T.J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* 8, 231-239.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7, 1-26.

Fraser, D.A.S. (1967). Data transformations and the linear model. *Ann. Math. Statist.* 38, 1456-1465.

Friedman, J.H. and Stuetzle, W. (1982). Smoothing of scatterplots. Dept. of Statistics, Stanford University, Tech. Report ORION006.

Gasser, T. and Rosenblatt, M. (eds.) (1979). *Smoothing Techniques for Curve Estimation*, in Lecture Notes in Mathematics 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. *Z. Angew. Math. Mech.* 21, 364-379.

Harrison, D. and Rubinfeld, D.L. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Mngmnt* 5, 81-102.

Kendall, M.A. and Stuart, A. (1967). *The Advanced Theory of Statistics*, Volume 2, Hafner.

Kimeldorf, G., May, J.H., and Sampson, A.R. (1982). Concordant and discordant monotone correlations and their evaluations by nonlinear optimization. In *Studies in the Management Sciences* 19, Optimization in Statistics, S.H. Zanakis and J.S. Rustagi, eds., North-Holland, pp. 117-130.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29, 115-129.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *J.R. Statist. Soc. B* 27, 251-263.

Lancaster, H.O. (1958). The structure of bivariate distributions. *Ann. Math. Statist.* 29, 719-736.

Lancaster, H.O. (1969). *The Chi-Squared Distribution*, John Wiley and Sons.

Linsey, J.K. (1972). Fitting response surfaces with power transformations. *J. R. Statist. Soc. C* 21, 234-237.

Linsey, J.K. (1974). Construction and comparison of statistical models. *J.R. Statist. Soc. B* 36, 418-425.

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

Renyi, A. (1959). On measures of dependence. *Acta. Math. Acad. Sci. Hungar.* 10, 441-451.

Sarmanov, O.V. (1958a). The maximal correlation coefficient (symmetric case). *Dokl. Acad. Nauk. SSSR* 120, 715-718.

Sarmanov, O.V. (1958b). The maximal correlation coefficient (nonsymmetric case). *Dokl. Acad. Nauk. SSSR* 121, 52-55.

Sarmanov, O.V. and Zaharov, V.K. (1960). Maximum coefficients of multiple correlation. *Dokl. Akad. Nauk SSSR* 130, 269-271.

Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression. *Ann. Statist.* 8, 240-246.

Stone, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.* 7, 139-149.

Tukey, J.W. (1982). The use of smelting in guiding re-expression, in *Modern Data Analysis*, Laurner and Siegel (eds.), Academic Press.

Wood, J.T. (1974). An extension of the analysis of transformations of Box and Cox. *Appl. Statist. (J.R. Statist. Soc. C)* 23.

Young, F.W., de Leeuw, J., Takane, Y. (1976). Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. *Psychometrika*, 1976, 505-529.

APPENDIX 1

Proof of Theorem 6.32

If the window size goes to zero at the right rate as $N \longrightarrow \infty$, then for i.i.d data most "reasonable" smooths which utilize local smoothing are m.s. consistent. There is a substantial literature on consistency, usually in higher dimensional spaces.

The common definition of consistency is: given a set of N-1 independent copies $(X_1,Y_1),\ldots,(X_{N-1},Y_{N-1})$, of (X,Y) drawn from the same bivariate distribution, and $\phi \in L_2(Y)$, call $S^{(N)}$ L_2-consistent if $E[S^{(N)}(\phi|X)-E(\phi(Y)|X)]^2 \longrightarrow 0$. To see that our definition is equivalent, put down the fixed point (x,y) and then the other N-1 random points $(x_1,y_1),\ldots,(x_{N-1},y_{N-1})$. Now compute $E[S^{(N)}(\phi|x)-E(\phi(Y)|x)]^2 = g_N(x)$. Our definition of m.s. consistency is then

$$E[\frac{1}{n}\sum_k g_N(X_k)] \longrightarrow 0$$

or $Eg_N(X) \longrightarrow 0$.

Uniform boundedness is a critical condition for consistency proofs. A key element in Stone's [1977] proof is (put in different form)

(A.1) PROPOSITION. *Take data sets drawn from a bivariate distribution* (X,Y), $S^{(N)}$ *a uniformly bounded sequence of smooths on* x, P_X *the conditional expectation operator. If, for a set of functions* $\{\phi\}$ *dense in* $H_2(Y)$,

$$E\|S^{(N)}(\phi|x)-P_X(\phi|x)\|_N^2 \to 0 ,$$

then the $S^{(N)}$ *as m.s. consistent. If, for a set of functions* $\{h\}$ *dense in* $H_2(X)$,

(A.2) $$E\|S^{(N)}(h|x)-h(x)\|_N^2 \to 0$$

then (A.2) holds for all $h \in H_2(X)$.

The proof is simple and is omitted.

Assume $S^{(N)}$ is linear. Then

(A.3) $$E\|S^{(N)}\phi-P_X\phi\|_N^2 \leq 2E\|S^{(N)}(\phi-P_X\phi)\|_N^2 + 2E\|S^{(N)}P_X\phi-P_X\phi\|_N^2 .$$

If it can be shown that $E\|S^{(N)}h-h\|_N^2 \to 0$ for all continuous $h \in H_2(X)$ vanishing off at finite intervals, and if the first term on the right in (A.3) goes to zero for all ϕ such that $\|\phi\|_\infty < \infty$, then (A.1) implies that $S^{(N)}$ is m.s. consistent. This strategy works for a wide variety of smooths.

To illustrate, because Stone's results [1977] do not seem immediately applicable to bivariate regression smooths, m.s. consistency is proven for the modified regression smooths (Theorem 6.32).

Assume $\|\phi\|_\infty < \infty$ and use the inequality (A.3) with $g(x) = P_X(\phi|x)$. Then

$$S(\phi-g|x) = \frac{1}{2M} \sum_{j\in I(x)} (\phi(y_j)-g(x_j)) \{1+ \frac{(x_j-\bar{x}_X)}{\sigma_X}[\frac{x-\bar{x}_X}{\sigma_X}]t\} .$$

$\Gamma_x(\phi,x)$, σ_x^2 the covariance between $\phi(y_k)$, x_k and the variance of x_k over the indices in $J(x)$.

Write the second term in (A.3) as

$$\frac{\Gamma_x(\phi,x)}{\sigma_x}\,\frac{(x-\bar{x}_x)}{\sigma_x}\ .$$

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

$$\left|\frac{x-\bar{x}_x}{\sigma_x}\right| \leq 1.$$

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

$$[x]_t = \begin{cases} x\ , & |x| \leq 1 \\ \operatorname{sgn}(s)\ , & |x| > 1\ , \end{cases}$$

and define the *modified regression smooth* by

$$S(\phi|x) = \bar{\phi}_x + \frac{\Gamma_x(\phi,x)}{\sigma_x}\left[\frac{x-\bar{x}_x}{\sigma_x}\right]_t \tag{A.5}$$

This modified smooth is bounded by 2.

(A.6) THEOREM. *If, as* $N\to\infty$, $M\to\infty$, $M\log N/N\to 0$ *and* $P(dx)$ *has no atoms, then the modified regression smooths are m.s. consistent.*

PROOF. Assume $\|\phi\|_\infty < \infty$ and use the inequality (A.3) with $g(x) = P_x(\phi|x)$. Then

$$S(\phi-g|x) = \frac{1}{2M}\sum_{j\in I(x)}(\phi(y_j)-g(x_j))\{1+\frac{(x_j-\bar{x}_x)}{\sigma_x}\left[\frac{x-\bar{x}_x}{\sigma_x}\right]_t\}\ .$$

The conditional expectation of $[S(\phi-g|x)]^2$ given $\{x_k\}$ is

$$\frac{1}{4M^2}\sum_{j\in I(x)}E[(\phi(y_j)-g(x_j))^2|x_j]\{1+\frac{(x_j-\bar{x}_x)}{\sigma_x}[\frac{x-\bar{x}_x}{\sigma_x}]\}^2$$
$$\leq \frac{1}{M}\|\phi\|_\infty .$$

Thus, the first term in (A.3) is asymptotically zero. Now look at $S(h|x)-h(x)$, $h\in H_2(X)$, h continuous and zero outside a finite interval;

$$S(h|x)-h(x)=\frac{1}{2M}\sum_{j\in J(x)}(h(x_j)-h(x))\{1+\frac{(x_j-\bar{x}_x)}{\sigma_x}[\frac{x-\bar{x}_x}{\sigma_x}]_t\}$$

Then, for $H(\delta)=\max\{|h(x')-h(x")|;\ |x'-x"|\leq\delta\}$,

$$[S(h|x)-h(x)]^2\leq[\frac{1}{2M}\sum_{j\in J(x)}(h(x_j)-h(x))^2]\cdot 2$$
$$\leq 2H(\max_{j\in J(x)}|x_j-x|) .$$

Then to get $E[S(h|x)-h(x)]^2\to 0$, it is enough to show that Δ $\Delta_N=\max\{|x_j-x|;\ x_j\in J(x)\}$ converges in probability to zero. Take x to be a point such that $P[(x,x+\varepsilon)]>0$, $P[(x-\varepsilon,x)]>0$ for all $\varepsilon>0$. The set S of all such points has $P(S)=1$. Then

$$\{\Delta_N>\varepsilon\}\subset\{\text{at most 2M-1 of }\{x_k\}\text{ in }(x-\varepsilon,x)\}$$
$$\cup\{\text{at most 2M-1 of }\{x_k\}\text{ in }(x,x+\varepsilon)\} .$$

But

$$\{\text{ at most 2M-1 of }\{x_k\}\text{ in }(x-\varepsilon,x)\}$$
$$=\{\frac{1}{N}\sum_{k=1}^{N}I(x_k\in(x-\varepsilon,x))\leq\frac{2M-1}{N}\}$$

and similarly for the number of $\{x_k\}$ in $(x,x+\varepsilon)$. Using the Law of Large Numbers and $M/N\to 0$ results in $P(\Delta_N>\varepsilon)\to 0$, proving the theorem.

```
      SUBROUTINE ACE (P,N,X,Y,W,L,DELRSQ,TX,TY,RSQ,IERR,M,Z)
C
C-----------------------------------------------------------------
C
C ESTIMATE OPTIMAL TRANSFORMATIONS FOR MULTIPLE REGRESSION AND
C CORRELATION BY ALTERNATING CONDITIONAL EXPECTATION ESTIMATES.
C
C (7/11/84)
C
C (BREIMAN AND FRIEDMAN, 1984, REV.)
C
C CODED BY: J. H. FRIEDMAN
C           DEPARTMENT OF STATISTICS AND
C           STANFORD LINEAR ACCELERATOR CENTER
C           STANFORD UNIVERSITY
C           STANFORD, CA. 94305
C
C INPUT:
C
C    N : NUMBER OF OBSERVATIONS.
C    P : NUMBER OF PREDICTOR VARIABLES FOR EACH OBSERVATION.
C    X(P,N) : PREDICTOR DATA MATRIX.
C    Y(N) : RESPONSE VALUE FOR EACH OBSERVATION.
C       MISSING VALUES ARE SIGNIFIED BY A VALUE (RESPONSE OR
C       PREDICTOR) GREATER THAN OR EQUAL TO BIG.
C       (SEE BELOW - DEFAULT, BIG = 1.0E20)
C    W(N) : WEIGHT FOR EACH OBSERVATION.
C    L(P+1) : FLAG FOR EACH VARIABLE.
C       L(1) THROUGH L(P) : PREDICTOR VARIABLES.
C       L(P+1) : RESPONSE VARIABLE.
C       L(I)=0 => ITH VARIABLE NOT TO BE USED.
C       L(I)=1 => ITH VARIABLE ASSUMES ORDERABLE VALUES.
C       L(I)=2 => ITH VARIABLE ASSUMES CIRCULAR (PERIODIC) VALUES
C                 IN THE RANGE (0.0,1.0) WITH PERIOD 1.0.
C       L(I)=3 => ITH VARIABLE TRANSFORMATION IS TO BE MONOTONE.
C       L(I)=4 => ITH VARIABLE TRANSFORMATION IS TO BE LINEAR.
C       L(I)=5 => ITH VARIABLE ASSUMES CATEGORICAL (UNORDERABLE) VALUES.
C   DELRSQ : TERMINATION THRESHOLD. ITERATION STOPS WHEN
C       RSQ CHANGES LESS THAN DELRSQ IN NTERM
C       CONSECUTIVE ITERATIONS (SEE BELOW - DEFAULT, NTERM=3).
C
C OUTPUT:
C
C   TX(N,P) : PREDICTOR TRANSFORMATIONS.
C      TX(J,I) = TRANSFORMED VALUE OF ITH PREDICTOR FOR JTH OBS.
C   TY(N) = RESPONSE TRANSFORMATION.
C      TY(J) = TRANSFORMED RESPONSE VALUE FOR JTH OBSERVATION.
C   RSQ = FRACTION OF VARIANCE(TY<Y>)
C                          P
C         EXPLAINED BY SUM TX(I)<X(I)> .
C                         I=1
C   IERR : ERROR FLAG.
C      IERR = 0 : NO ERRORS DETECTED.
C      IERR > 0 : ERROR DETECTED - SEE FORMAT STATEMENTS BELOW.
C
```

```
C SCRATCH:
C
C    M(N,P+1), Z(N,12) : INTERNAL WORKING STORAGE.
C    (Z(J,1), J=1,N) CONTAIN (TRANSFORMED) RESIDUALS AS OUTPUT.
C
C NOTE: ACE USES AN ITERATIVE PROCEDURE FOR SOLVING THE OPTIMIZATION
C    PROBLEM. DEFAULT STARTING TRANSFORMATIONS ARE TY(J)=Y(J),
C    TX(J,I)=X(I,J) : J=1,N, I=1,P. OTHER STARTING TRANSFORMATIONS CAN
C    BE SPECIFIED (IF DESIRED) FOR EITHER THE RESPONSE AND/OR ANY OF
C    THE PREDICTOR VARIABLES. THIS IS SIGNALED BY NEGATING THE
C    CORRESPONDING L(I) VALUE AND STORING THE STARTING TRANSFORMED
C    VALUES IN THE CORRESPONDING ARRAY (TY(J), TX(J,I)) BEFORE
C    CALLING ACE.
C
C-----------------------------------------------------------------
C
      INTEGER P,PP1,M(N,1),L(1)
      REAL Y(N),X(P,N),W(N),TY(N),TX(N,P),Z(N,12),CT(10)
      COMMON /PARMS/ ITAPE,MAXIT,NTERM,SPAN,ALPHA,BIG
      DOUBLE PRECISION SM,SV,SW,SW1
      IERR=0
      PP1=P+1
      SM=0.0
      SV=SM
      SW=SV
      SW1=SW
      DO 10 I=1,PP1
      IF (L(I).GE.-5.AND.L(I).LE.5) GO TO 10
      IERR=6
      IF (ITAPE.GT.0) WRITE (ITAPE,550) I,L(I)
 10   CONTINUE
      IF (IERR.NE.0) RETURN
      IF (L(PP1).NE.0) GO TO 20
      IERR=4
      IF (ITAPE.GT.0) WRITE (ITAPE,530) PP1
      RETURN
 20   NP=0
      DO 30 I=1,P
      IF (L(I).NE.0) NP=NP+1
 30   CONTINUE
      IF (NP.GT.0) GO TO 40
      IERR=5
      IF (ITAPE.GT.0) WRITE (ITAPE,540) P
      RETURN
 40   DO 50 J=1,N
      SW=SW+W(J)
      IF (L(PP1).GT.0) TY(J)=Y(J)
 50   CONTINUE
      IF (SW.GT.0.0) GO TO 60
      IERR=1
      IF (ITAPE.GT.0) WRITE (ITAPE,500)
      RETURN
 60   DO 160 I=1,P
      IF (L(I).NE.0) GO TO 80
      DO 70 J=1,N
```

```
      TX(J,I)=0.0
 70   CONTINUE
      GO TO 160
 80   IF (L(I).LE.0) GO TO 100
      DO 90 J=1,N
      TX(J,I)=X(I,J)
 90   CONTINUE
 100  DO 110 J=1,N
      IF (TX(J,I).GE.BIG) GO TO 110
      SM=SM+W(J)*TX(J,I)
      SW1=SW1+W(J)
 110  CONTINUE
      IF (SW1.GT.0.0) GO TO 130
      DO 120 J=1,N
      TX(J,I)=0.0
 120  CONTINUE
      SM=0.0
      SW1=SM
      GO TO 160
 130  SM=SM/SW1
      DO 150 J=1,N
      IF (TX(J,I).GE.BIG) GO TO 140
      TX(J,I)=TX(J,I)-SM
      GO TO 150
 140  TX(J,I)=0.0
 150  CONTINUE
      SM=0.0
      SW1=SM
 160  CONTINUE
      DO 170 J=1,N
      M(J,PP1)=J
      Z(J,2)=Y(J)
      IF (TY(J).GE.BIG) GO TO 170
      SM=SM+W(J)*TY(J)
      SW1=SW1+W(J)
 170  CONTINUE
      IF (SW1.GT.0.0) GO TO 180
      IERR=1
      IF (ITAPE.GT.0) WRITE (ITAPE,500)
      RETURN
 180  SM=SM/SW1
      DO 200 J=1,N
      IF (TY(J).GE.BIG) GO TO 190
      TY(J)=TY(J)-SM
      GO TO 200
 190  TY(J)=0.0
 200  CONTINUE
      DO 210 J=1,N
      SV=SV+W(J)*TY(J)**2
 210  CONTINUE
      SV=SV/SW
      IF (SV.LE.0.0) GO TO 220
      SV=1.0/DSQRT(SV)
      GO TO 250
 220  IF (L(PP1).LE.0) GO TO 230
```

```
      IERR=2
      IF (ITAPE.GT.0) WRITE (ITAPE,510)
      GO TO 240
  230 IERR=3
      IF (ITAPE.GT.0) WRITE (ITAPE,520)
  240 RETURN
  250 DO 260 J=1,N
      TY(J)=TY(J)*SV
  260 CONTINUE
      CALL SORT (Z(1,2),M(1,PP1),1,N)
      DO 280 I=1,P
      IF (L(I).EQ.0) GO TO 280
      DO 270 J=1,N
      M(J,I)=J
      Z(J,2)=X(I,J)
  270 CONTINUE
      CALL SORT (Z(1,2),M(1,I),1,N)
  280 CONTINUE
      CALL SCALE (P,N,W,SW,TY,TX,DELRSQ,P,Z(1,5),Z(1,6))
      RSQ=0.0
      ITER=0
      NTERM=MIN0(NTERM,10)
      NT=0
      DO 290 I=1,NTERM
      CT(I)=100.0
  290 CONTINUE
  300 ITER=ITER+1
      NIT=0
  310 RSQI=RSQ
      NIT=NIT+1
      DO 330 J=1,N
      Z(J,5)=TY(J)
      DO 320 I=1,P
      IF (L(I).NE.0) Z(J,5)=Z(J,5)-TX(J,I)
  320 CONTINUE
  330 CONTINUE
      DO 390 I=1,P
      IF (L(I).EQ.0) GO TO 390
      DO 340 J=1,N
      K=M(J,I)
      Z(J,1)=Z(K,5)+TX(K,I)
      Z(J,2)=X(I,K)
      Z(J,4)=W(K)
  340 CONTINUE
      CALL SMOTHR (IABS(L(I)),N,Z(1,2),Z,Z(1,4),Z(1,3),Z(1,6))
      SM=0.0
      DO 350 J=1,N
      SM=SM+Z(J,4)*Z(J,3)
  350 CONTINUE
      SM=SM/SW
      DO 360 J=1,N
      Z(J,3)=Z(J,3)-SM
  360 CONTINUE
      SV=0.0
      DO 370 J=1,N
```

```
      SV=SV+Z(J,4)*(Z(J,1)-Z(J,3))**2
  370 CONTINUE
      SV=1.0-SV/SW
      IF (SV.LE.RSQ) GO TO 390
      RSQ=SV
      DO 380 J=1,N
      K=M(J,I)
      TX(K,I)=Z(J,3)
      Z(K,5)=Z(J,1)-Z(J,3)
  380 CONTINUE
  390 CONTINUE
      IF ((NP.NE.1).AND.((RSQ-RSQI.GT.DELRSQ).AND.(NIT.LT.MAXIT))) GO TO
     1 310
      DO 410 J=1,N
      K=M(J,PP1)
      Z(J,2)=Y(K)
      Z(J,4)=W(K)
      Z(J,1)=0.0
      DO 400 I=1,P
      IF (L(I).NE.0) Z(J,1)=Z(J,1)+TX(K,I)
  400 CONTINUE
  410 CONTINUE
      CALL SMOTHR (IABS(L(PP1)),N,Z(1,2),Z,Z(1,4),Z(1,3),Z(1,6))
      SM=0.0
      SV=SM
      DO 420 J=1,N
      K=M(J,PP1)
      SM=SM+W(K)*Z(J,3)
      Z(K,2)=Z(J,1)
  420 CONTINUE
      SM=SM/SW
      DO 430 J=1,N
      Z(J,3)=Z(J,3)-SM
      SV=SV+Z(J,4)*Z(J,3)**2
  430 CONTINUE
      SV=SV/SW
      IF (SV.LE.0.0) GO TO 440
      SV=1.0/DSQRT(SV)
      GO TO 450
  440 IERR=3
      IF (ITAPE.GT.0) WRITE (ITAPE,520)
      RETURN
  450 DO 460 J=1,N
      K=M(J,PP1)
      TY(K)=Z(J,3)*SV
  460 CONTINUE
      SV=0.0
      DO 470 J=1,N
      SV=SV+W(J)*(TY(J)-Z(J,2))**2
  470 CONTINUE
      RSQ=1.0-SV/SW
      IF (ITAPE.GT.0) WRITE (ITAPE,490) ITER,RSQ
      NT=MOD(NT,NTERM)+1
      CT(NT)=RSQ
      CMN=100.0
```

```
      CMX=-100.0
      DO 480 I=1,NTERM
      CMN=AMIN1(CMN,CT(I))
      CMX=AMAX1(CMX,CT(I))
 480  CONTINUE
      IF ((CMX-CMN.GT.DELRSQ).AND.(ITER.LT.MAXIT)) GO TO 300
      RETURN
 490  FORMAT(   11H ITERATION I2,   23H   R**2  =  1 - E**2  =G12.4)
 500  FORMAT(   41H IERR=1: SUM OF WEIGHTS (W) NOT POSITIVE.)
 510  FORMAT(   29H IERR=2: Y HAS ZERO VARIANCE.)
 520  FORMAT(   30H IERR=3: TY HAS ZERO VARIANCE.)
 530  FORMAT(   11H IERR=4: L(I2,   18H) MUST BE NONZERO.)
 540  FORMAT(   29H IERR=5: AT LEAST ONE L(1)-L(I2,   18H) MUST BE NONZE
     1RO.)
 550  FORMAT(   11H IERR=6: L(I2,    3H) =G12.4,   30H MUST BE IN THE RA
     1NGE (-5, 5).)
      END
      SUBROUTINE MODEL (P,N,Y,W,L,TX,TY,F,T,M,Z)
C
C-----------------------------------------------------------------------
C
C COMPUTES RESPONSE PREDICTIVE  FUNCTION F FOR THE MODEL YHAT = F(T),
C WHERE
C                                        P
C            F(T) = E(Y : T),     T =   SUM  TX<I> ( X<I> )
C                                       I=1
C USING THE X TRANSFORMATIONS TX CONSTRUCTED BY SUBROUTINE ACE.
C IF Y IS A CATEGORICAL VARIABLE (CLASSIFICATION) THEN
C                                  -1
C                        F(T) = TY  (T).
C INPUT:
C
C    P,N,Y,W,L : SAME INPUT AS FOR SUBROUTINE ACE.
C    TX,TY,M,Z : OUTPUT FROM SUBROUTINE ACE.
C
C OUTPUT:
C
C    F(N),T(N) : INPUT FOR SUBROUTINE ACEMOD.
C
C NOTE: THIS SUBROUTINE MUST BE CALLED BEFORE SUBROUTINE ACEMOD.
C
C-----------------------------------------------------------------------
C
      INTEGER P,PP1,M(N,1),L(1)
      REAL Y(N),W(N),TX(N,P),TY(N),F(N),T(N),Z(N,12)
      COMMON /PARMS/ ITAPE,MAXIT,NTERM,SPAN,ALPHA,BIG
      PP1=P+1
      IF (IABS(L(PP1)).NE.5) GO TO 20
      DO 10 J=1,N
      T(J)=TY(J)
      M(J,PP1)=J
 10   CONTINUE
      GO TO 50
 20   DO 40 J=1,N
      S=0.0
```

```
      DO 30 I=1,P
      S=S+TX(J,I)
 30   CONTINUE
      T(J)=S
      M(J,PP1)=J
 40   CONTINUE
 50   CALL SORT (T,M(1,PP1),1,N)
      DO 140 J=1,N
      K=M(J,PP1)
      Z(J,2)=W(K)
      IF (Y(K).GE.BIG) GO TO 60
      Z(J,1)=Y(K)
      GO TO 140
 60   J1=J
      J2=J1
 70   IF (Y(M(J1,PP1)).LT.BIG) GO TO 80
      J1=J1-1
      IF (J1.GE.1) GO TO 70
 80   IF (Y(M(J2,PP1)).LT.BIG) GO TO 90
      J2=J2+1
      IF (J2.LE.N) GO TO 80
 90   IF (J1.GE.1) GO TO 100
      K=J2
      GO TO 130
 100  IF (J2.LE.N) GO TO 110
      K=J1
      GO TO 130
 110  IF (T(J)-T(J1).GE.T(J2)-T(J)) GO TO 120
      K=J1
      GO TO 130
 120  K=J2
 130  Z(J,1)=Y(M(K,PP1))
      T(J)=T(K)
 140  CONTINUE
      IF (IABS(L(PP1)).NE.5) GO TO 160
      DO 150 J=1,N
      F(J)=Z(J,1)
 150  CONTINUE
      GO TO 170
 160  CALL SMOTHR (1,N,T,Z,Z(1,2),F,Z(1,6))
 170  RETURN
      END
      SUBROUTINE ACEMOD (V,P,N,X,L,TX,F,T,M,YHAT)
C
C-----------------------------------------------------------------
C
C COMPUTES RESPONSE Y ESTIMATES FROM THE MODEL
C
C                YHAT =  F ( T( V ) )
C
C USING THE X TRANSFORMATIONS TX CONSTRUCTED BY SUBROUTINE ACE AND
C THE PREDICTOR FUNCTION (F,T) CONSTRUCTED BY SUBROUTINE MODEL.
C
C INPUT:
C
```

```
C       V(P) : VECTOR OF PREDICTOR VALUES.
C    P,N,X,L : SAME INPUT AS FOR SUBROUTINE ACE.
C       TX,M : OUTPUT FROM SUBROUTINE ACE.
C        F,T : OUTPUT FROM SUBROUTINE MODEL.
C
C OUTPUT:
C
C    YHAT : ESTIMATED RESPONSE VALUE FOR V.
C
C NOTE: THIS SUBROUTINE MUST NOT BE CALLED BEFORE SUBROUTINE MODEL.
C
C-----------------------------------------------------------------
C
      INTEGER P,M(N,1),L(1),LOW,HIGH,PLACE
      REAL V(P),X(P,N),F(N),T(N),TX(N,P)
      COMMON /PARMS/ ITAPE,MAXIT,NTERM,SPAN,ALPHA,BIG
      TH=0.0
      DO 90 I=1,P
      IF (L(I).EQ.0) GO TO 90
      VI=V(I)
      IF (VI.LT.BIG) GO TO 10
      IF (X(I,M(N,I)).GE.BIG) TH=TH+TX(M(N,I),I)
      GO TO 90
 10   IF (VI.GT.X(I,M(1,I))) GO TO 20
      PLACE=1
      GO TO 80
 20   IF (VI.LT.X(I,M(N,I))) GO TO 30
      PLACE=N
      GO TO 80
 30   LOW=0
      HIGH=N+1
 40   IF (LOW+1.GE.HIGH) GO TO 60
      PLACE=(LOW+HIGH)/2
      XT=X(I,M(PLACE,I))
      IF (VI.EQ.XT) GO TO 80
      IF (VI.GE.XT) GO TO 50
      HIGH=PLACE
      GO TO 40
 50   LOW=PLACE
      GO TO 40
 60   IF (IABS(L(I)).EQ.5) GO TO 90
      JL=M(LOW,I)
      JH=M(HIGH,I)
      IF (X(I,JH).LT.BIG) GO TO 70
      TH=TH+TX(JL,I)
      GO TO 90
 70   TH=TH+TX(JL,I)+(TX(JH,I)-TX(JL,I))*(VI-X(I,JL))/(X(I,JH)-X(I,JL))
      GO TO 90
 80   TH=TH+TX(M(PLACE,I),I)
 90   CONTINUE
      IF (TH.GT.T(1)) GO TO 100
      YHAT=F(1)
      RETURN
 100  IF (TH.LT.T(N)) GO TO 110
      YHAT=F(N)
```

```
      RETURN
  110 LOW=0
      HIGH=N+1
  120 IF (LOW+1.GE.HIGH) GO TO 150
      PLACE=(LOW+HIGH)/2
      XT=T(PLACE)
      IF (TH.NE.XT) GO TO 130
      YHAT=F(PLACE)
      RETURN
  130 IF (TH.GE.XT) GO TO 140
      HIGH=PLACE
      GO TO 120
  140 LOW=PLACE
      GO TO 120
  150 IF (IABS(L(P+1)).NE.5) GO TO 170
      IF (TH-T(LOW).GT.T(HIGH)-TH) GO TO 160
      YHAT=F(LOW)
      GO TO 180
  160 YHAT=F(HIGH)
      GO TO 180
  170 YHAT=F(LOW)+(F(HIGH)-F(LOW))*(TH-T(LOW))/(T(HIGH)-T(LOW))
  180 RETURN
      END
      BLOCK DATA
C
C-----------------------------------------------------------------
C
C THESE PROCEDURE PARAMETERS CAN BE CHANGED IN THE CALLING ROUTINE
C BY DEFINING THE ABOVE LABELED COMMON AND RESETTING THE VALUES WITH
C EXECUTABLE STATEMENTS.
C
C ITAPE : FORTRAN FILE NUMBER FOR PRINTER OUTPUT.
C         (ITAPE.LE.0 => NO PRINTER OUTPUT.)
C MAXIT : MAXIMUM NUMBER OF ITERATIONS.
C NTERM : NUMBER OF CONSECUTIVE ITERATIONS FOR WHICH
C         RSQ MUST CHANGE LESS THAN DELCOR FOR CONVERGENCE.
C SPAN, ALPHA : SUPER SMOOTHER PARAMETERS (SEE BELOW).
C BIG : A LARGE REPRESENTABLE FLOATING POINT NUMBER.
C
C-----------------------------------------------------------------
C
      COMMON /PARMS/ ITAPE,MAXIT,NTERM,SPAN,ALPHA,BIG
      DATA ITAPE,MAXIT,NTERM,SPAN,ALPHA,BIG /6,20,3,0.0,5.0,1.0E20/
      END
      SUBROUTINE SMOTHR (L,NN,X,Y,W,SMO,SCR)
      REAL X(NN),Y(NN),W(NN),SMO(NN),SCR(NN,7)
      COMMON /PARMS/ ITAPE,MAXIT,NTERM,SPAN,ALPHA,BIG
      DOUBLE PRECISION SM,SW,A,B,D
      N=NN
      SM=0.0
      SW=SM
  10  IF (X(N).LT.BIG) GO TO 20
      SM=SM+W(N)*Y(N)
      SW=SW+W(N)
      N=N-1
```

```
      IF (N.GE.1) GO TO 10
 20   IF (N.GE.NN) GO TO 40
      NP1=N+1
      SM=SM/SW
      DO 30 J=NP1,NN
      SMO(J)=SM
 30   CONTINUE
 40   IF (N.LT.1) RETURN
      IF (L.LT.5) GO TO 90
      J=1
 50   J0=J
      SM=W(J)*Y(J)
      SW=W(J)
      IF (J.GE.N) GO TO 70
 60   IF (X(J+1).GT.X(J)) GO TO 70
      J=J+1
      SM=SM+W(J)*Y(J)
      SW=SW+W(J)
      IF (J.LT.N) GO TO 60
 70   SM=SM/SW
      DO 80 I=J0,J
      SMO(I)=SM
 80   CONTINUE
      J=J+1
      IF (J.LE.N) GO TO 50
      GO TO 200
 90   IF (L.NE.4) GO TO 140
      SM=0.0
      SW=SM
      B=SW
      D=B
      DO 100 J=1,N
      SM=SM+W(J)*X(J)*Y(J)
      SW=SW+W(J)*X(J)**2
      B=B+W(J)*X(J)
      D=D+W(J)
 100  CONTINUE
      A=SW-(B**2)/D
      IF (A.GT.0.0) GO TO 110
      A=0.0
      GO TO 120
 110  A=SM/A
 120  B=B/D
      DO 130 J=1,N
      SMO(J)=A*(X(J)-B)
 130  CONTINUE
      GO TO 200
 140  CALL SUPSMU (N,X,Y,W,L,SPAN,ALPHA,SMO,SCR)
      IF (L.NE.3) GO TO 200
      DO 150 J=1,N
      SCR(J,1)=SMO(J)
      SCR(N-J+1,2)=SCR(J,1)
 150  CONTINUE
      CALL MONTNE (SCR,N)
      CALL MONTNE (SCR(1,2),N)
```

```
      SM=0.0
      SW=SM
      DO 160 J=1,N
      SM=SM+(SMO(J)-SCR(J,1))**2
      SW=SW+(SMO(J)-SCR(N-J+1,2))**2
160   CONTINUE
      IF (SM.GE.SW) GO TO 180
      DO 170 J=1,N
      SMO(J)=SCR(J,1)
170   CONTINUE
      GO TO 200
180   DO 190 J=1,N
      SMO(J)=SCR(N-J+1,2)
190   CONTINUE
200   RETURN
      END
      SUBROUTINE MONTNE (X,N)
      REAL X(N)
      INTEGER BB,EB,BR,ER,BL,EL
      BB=0
      EB=BB
10    IF (EB.GE.N) GO TO 110
      BB=EB+1
      EB=BB
20    IF (EB.GE.N) GO TO 30
      IF (X(BB).NE.X(EB+1)) GO TO 30
      EB=EB+1
      GO TO 20
30    IF (EB.GE.N) GO TO 70
      IF (X(EB).LE.X(EB+1)) GO TO 70
      BR=EB+1
      ER=BR
40    IF (ER.GE.N) GO TO 50
      IF (X(ER+1).NE.X(BR)) GO TO 50
      ER=ER+1
      GO TO 40
50    PMN=(X(BB)*(EB-BB+1)+X(BR)*(ER-BR+1))/(ER-BB+1)
      EB=ER
      DO 60 I=BB,EB
      X(I)=PMN
60    CONTINUE
70    IF (BB.LE.1) GO TO 10
      IF (X(BB-1).LE.X(BB)) GO TO 10
      BL=BB-1
      EL=BL
80    IF (BL.LE.1) GO TO 90
      IF (X(BL-1).NE.X(EL)) GO TO 90
      BL=BL-1
      GO TO 80
90    PMN=(X(BB)*(EB-BB+1)+X(BL)*(EL-BL+1))/(EB-BL+1)
      BB=BL
      DO 100 I=BB,EB
      X(I)=PMN
100   CONTINUE
      GO TO 30
```

```
110   RETURN
      END
      SUBROUTINE SCALE (P,N,W,SW,TY,TX,EPS,MAXIT,R,SC)
      INTEGER P
      REAL W(N),TY(N),TX(N,P),R(N),SC(P,5)
      DOUBLE PRECISION S,H,T,U,GAMA,DELTA,SW
      DO 10 I=1,P
      SC(I,1)=0.0
10    CONTINUE
      NIT=0
20    NIT=NIT+1
      DO 30 I=1,P
      SC(I,5)=SC(I,1)
30    CONTINUE
      DO 160 ITER=1,P
      DO 50 J=1,N
      S=0.0
      DO 40 I=1,P
      S=S+SC(I,1)*TX(J,I)
40    CONTINUE
      R(J)=(TY(J)-S)*W(J)
50    CONTINUE
      DO 70 I=1,P
      S=0.0
      DO 60 J=1,N
      S=S+R(J)*TX(J,I)
60    CONTINUE
      SC(I,2)=-2.0*S/SW
70    CONTINUE
      S=0.0
      DO 80 I=1,P
      S=S+SC(I,2)**2
80    CONTINUE
      IF (S.LE.0.0) GO TO 170
      IF (ITER.NE.1) GO TO 100
      DO 90 I=1,P
      SC(I,3)=-SC(I,2)
90    CONTINUE
      H=S
      GO TO 120
100   GAMA=S/H
      H=S
      DO 110 I=1,P
      SC(I,3)=-SC(I,2)+GAMA*SC(I,4)
110   CONTINUE
120   S=0.0
      T=S
      DO 140 J=1,N
      U=0.0
      DO 130 I=1,P
      U=U+SC(I,3)*TX(J,I)
130   CONTINUE
      S=S+U*R(J)
      T=T+W(J)*U**2
140   CONTINUE
```

```
      DELTA=S/T
      DO 150 I=1,P
      SC(I,1)=SC(I,1)+DELTA*SC(I,3)
      SC(I,4)=SC(I,3)
  150 CONTINUE
  160 CONTINUE
  170 V=0.0
      DO 180 I=1,P
      V=AMAX1(V,ABS(SC(I,1)-SC(I,5)))
  180 CONTINUE
      IF ((V.GE.EPS).AND.(NIT.LT.MAXIT)) GO TO 20
      DO 200 I=1,P
      DO 190 J=1,N
      TX(J,I)=SC(I,1)*TX(J,I)
  190 CONTINUE
  200 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE SUPSMU (N,X,Y,W,IPER,SPAN,ALPHA,SMO,SC)
C
C------------------------------------------------------------
C
C SUPER SMOOTHER (FRIEDMAN AND STUETZLE, 1984).
C
C VERSION 3/10/84
C
C CODED BY: J. H. FRIEDMAN
C           DEPARTMENT OF STATISTICS AND
C           STANFORD LINEAR ACCELERATOR CENTER
C           STANFORD UNIVERSITY
C           STANFORD CA. 94305
C
C INPUT:
C    N : NUMBER OF OBSERVATIONS (X,Y - PAIRS).
C    X(N) : ORDERED ABSCISSA VALUES.
C    Y(N) : CORRESPONDING ORDINATE (RESPONSE) VALUES.
C    W(N) : WEIGHT FOR EACH (X,Y) OBSERVATION.
C    IPER : PERIODIC VARIABLE FLAG.
C       IPER=1 => X IS ORDERED INTERVAL VARIABLE.
C       IPER=2 => X IS A PERIODIC VARIABLE WITH VALUES
C                 IN THE RANGE (0.0,1.0) AND PEROID 1.0.
C    SPAN : SMOOTHER SPAN (FRACTION OF OBSERVATIONS IN WINDOW).
C           SPAN=0.0 => AUTOMATIC (VARIABLE) SPAN SELECTION.
C    ALPHA : CONTROLES HIGH FREQUENCY (SMALL SPAN) PENALITY
C            USED WITH AUTOMATIC SPAN SELECTION (BASE TONE CONTROL).
C            (ALPHA.LE.0.0 OR ALPHA.GT.10.0 => NO EFFECT.)
C OUTPUT:
C   SMO(N) : SMOOTHED ORDINATE (RESPONSE) VALUES.
C SCRATCH:
C   SC(N,7) : INTERNAL WORKING STORAGE.
C
C NOTE:
C    FOR SMALL SAMPLES (N < 40) OR IF THERE ARE SUBSTANTIAL SERIAL
C    CORRELATIONS BETWEEN OBSERATIONS CLOSE IN X - VALUE, THEN
C    A PRESPECIFIED FIXED SPAN SMOOTHER (SPAN > 0) SHOULD BE
C    USED. REASONABLE SPAN VALUES ARE 0.3 TO 0.5.
C
C------------------------------------------------------------
C
      DIMENSION X(N),Y(N),W(N),SMO(N),SC(N,7)
      COMMON /SPANS/ SPANS(3) /CONSTS/ BIG,SML,EPS
      IF (X(N).GT.X(1)) GO TO 30
      SY=0.0
      SW=SY
      DO 10 J=1,N
      SY=SY+W(J)*Y(J)
      SW=SW+W(J)
 10   CONTINUE
      A=SY/SW
      DO 20 J=1,N
      SMO(J)=A
 20   CONTINUE
      RETURN
```

```
 30   I=N/4
      J=3*I
      SCALE=X(J)-X(I)
 40   IF (SCALE.GT.0.0) GO TO 50
      IF (J.LT.N) J=J+1
      IF (I.GT.1) I=I-1
      SCALE=X(J)-X(I)
      GO TO 40
 50   VSMLSQ=(EPS*SCALE)**2
      JPER=IPER
      IF (IPER.EQ.2.AND.(X(1).LT.0.0.OR.X(N).GT.1.0)) JPER=1
      IF (JPER.LT.1.OR.JPER.GT.2) JPER=1
      IF (SPAN.LE.0.0) GO TO 60
      CALL SMOOTH (N,X,Y,W,SPAN,JPER,VSMLSQ,SMO,SC)
      RETURN
 60   DO 70 I=1,3
      CALL SMOOTH (N,X,Y,W,SPANS(I),JPER,VSMLSQ,SC(1,2*I-1),SC(1,7))
      CALL SMOOTH (N,X,SC(1,7),W,SPANS(2),-JPER,VSMLSQ,SC(1,2*I),H)
 70   CONTINUE
      DO 90 J=1,N
      RESMIN=BIG
      DO 80 I=1,3
      IF (SC(J,2*I).GE.RESMIN) GO TO 80
      RESMIN=SC(J,2*I)
      SC(J,7)=SPANS(I)
 80   CONTINUE
      IF (ALPHA.GT.0.0.AND.ALPHA.LE.10.0.AND.RESMIN.LT.SC(J,6)) SC(J,7)=
     1SC(J,7)+(SPANS(3)-SC(J,7))*AMAX1(SML,RESMIN/SC(J,6))**(10.0-ALPHA)
 90   CONTINUE
      CALL SMOOTH (N,X,SC(1,7),W,SPANS(2),-JPER,VSMLSQ,SC(1,2),H)
      DO 110 J=1,N
      IF (SC(J,2).LE.SPANS(1)) SC(J,2)=SPANS(1)
      IF (SC(J,2).GE.SPANS(3)) SC(J,2)=SPANS(3)
      F=SC(J,2)-SPANS(2)
      IF (F.GE.0.0) GO TO 100
      F=-F/(SPANS(2)-SPANS(1))
      SC(J,4)=(1.0-F)*SC(J,3)+F*SC(J,1)
      GO TO 110
 100  F=F/(SPANS(3)-SPANS(2))
      SC(J,4)=(1.0-F)*SC(J,3)+F*SC(J,5)
 110  CONTINUE
      CALL SMOOTH (N,X,SC(1,4),W,SPANS(1),-JPER,VSMLSQ,SMO,H)
      RETURN
      END
      BLOCK DATA
C
C-------------------------------------------------------------
C
C THIS SETS THE COMPILE TIME (DEFAULT) VALUES FOR VARIOUS
C INTERNAL PARAMETERS :
C
C SPANS : SPAN VALUES FOR THE THREE RUNNING LINEAR SMOOTHERS.
C SPANS(1) : TWEETER SPAN.
C SPANS(2) : MIDRANGE SPAN.
C SPANS(3) : WOOFER SPAN.
```

```
C (THESE SPAN VALUES SHOULD BE CHANGED ONLY WITH CARE.)
C BIG : A LARGE REPRESENTABLE FLOATING POINT NUMBER.
C SML : A SMALL NUMBER. SHOULD BE SET SO THAT (SML)**(10.0) DOES
C       NOT CAUSE FLOATING POINT UNDERFLOW.
C EPS : USED TO NUMERICALLY STABILIZE SLOPE CALCULATIONS FOR
C       RUNNING LINEAR FITS.
C
C THESE PARAMETER VALUES CAN BE CHANGED BY DECLARING THE
C RELEVANT LABELED COMMON IN THE MAIN PROGRAM AND RESETTING
C THEM WITH EXECUTABLE STATEMENTS.
C
C-------------------------------------------------------------
C
      COMMON /SPANS/ SPANS(3) /CONSTS/ BIG,SML,EPS
      DATA SPANS,BIG,SML,EPS /0.05,0.2,0.5,1.0E20,1.0E-7,1.0E-3/
      END
      SUBROUTINE SMOOTH (N,X,Y,W,SPAN,IPER,VSMLSQ,SMO,ACVR)
      DIMENSION X(N),Y(N),W(N),SMO(N),ACVR(N)
      INTEGER IN,OUT
      DOUBLE PRECISION WT,FBO,FBW,XM,YM,TMP,VAR,CVAR,A,H,SY
      XM=0.0
      YM=XM
      VAR=YM
      CVAR=VAR
      FBW=CVAR
      JPER=IABS(IPER)
      IBW=0.5*SPAN*N+0.5
      IF (IBW.LT.2) IBW=2
      IT=2*IBW+1
      DO 20 I=1,IT
      J=I
      IF (JPER.EQ.2) J=I-IBW-1
      XTI=X(J)
      IF (J.GE.1) GO TO 10
      J=N+J
      XTI=X(J)-1.0
 10   WT=W(J)
      FBO=FBW
      FBW=FBW+WT
      XM=(FBO*XM+WT*XTI)/FBW
      YM=(FBO*YM+WT*Y(J))/FBW
      TMP=0.0
      IF (FBO.GT.0.0) TMP=FBW*WT*(XTI-XM)/FBO
      VAR=VAR+TMP*(XTI-XM)
      CVAR=CVAR+TMP*(Y(J)-YM)
 20   CONTINUE
      DO 70 J=1,N
      OUT=J-IBW-1
      IN=J+IBW
      IF ((JPER.NE.2).AND.(OUT.LT.1.OR.IN.GT.N)) GO TO 60
      IF (OUT.GE.1) GO TO 30
      OUT=N+OUT
      XTO=X(OUT)-1.0
      XTI=X(IN)
      GO TO 50
```

```
 30   IF (IN.LE.N) GO TO 40
      IN=IN-N
      XTI=X(IN)+1.0
      XTO=X(OUT)
      GO TO 50
 40   XTO=X(OUT)
      XTI=X(IN)
 50   WT=W(OUT)
      FBO=FBW
      FBW=FBW-WT
      TMP=0.0
      IF (FBW.GT.0.0) TMP=FBO*WT*(XTO-XM)/FBW
      VAR=VAR-TMP*(XTO-XM)
      CVAR=CVAR-TMP*(Y(OUT)-YM)
      XM=(FBO*XM-WT*XTO)/FBW
      YM=(FBO*YM-WT*Y(OUT))/FBW
      WT=W(IN)
      FBO=FBW
      FBW=FBW+WT
      XM=(FBO*XM+WT*XTI)/FBW
      YM=(FBO*YM+WT*Y(IN))/FBW
      TMP=0.0
      IF (FBO.GT.0.0) TMP=FBW*WT*(XTI-XM)/FBO
      VAR=VAR+TMP*(XTI-XM)
      CVAR=CVAR+TMP*(Y(IN)-YM)
 60   A=0.0
      IF (VAR.GT.VSMLSQ) A=CVAR/VAR
      SMO(J)=A*(X(J)-XM)+YM
      IF (IPER.LE.0) GO TO 70
      H=1.0/FBW
      IF (VAR.GT.VSMLSQ) H=H+(X(J)-XM)**2/VAR
      ACVR(J)=ABS(Y(J)-SMO(J))/(1.0-W(J)*H)
 70   CONTINUE
      J=1
 80   J0=J
      SY=SMO(J)*W(J)
      FBW=W(J)
      IF (J.GE.N) GO TO 100
 90   IF (X(J+1).GT.X(J)) GO TO 100
      J=J+1
      SY=SY+W(J)*SMO(J)
      FBW=FBW+W(J)
      IF (J.LT.N) GO TO 90
100   IF (J.LE.J0) GO TO 120
      SY=SY/FBW
      DO 110 I=J0,J
      SMO(I)=SY
110   CONTINUE
120   J=J+1
      IF (J.LE.N) GO TO 80
      RETURN
      END
```

```
      SUBROUTINE SORT (V,A,II,JJ)
C
C     PUTS INTO A THE PERMUTATION VECTOR WHICH SORTS V INTO
C     INCREASING ORDER.  ONLY ELEMENTS FROM II TO JJ ARE CONSIDERED.
C     ARRAYS IU(K) AND IL(K) PERMIT SORTING UP TO 2**(K+1)-1 ELEMENTS
C
C     THIS IS A MODIFICATION OF CACM ALGORITHM #347 BY R. C. SINGLETON,
C     WHICH IS A MODIFIED HOARE QUICKSORT.
C
      DIMENSION A(JJ),V(1),IU(20),IL(20)
      INTEGER T,TT
      INTEGER A
      REAL V
      M=1
      I=II
      J=JJ
 10   IF (I.GE.J) GO TO 80
 20   K=I
      IJ=(J+I)/2
      T=A(IJ)
      VT=V(IJ)
      IF (V(I).LE.VT) GO TO 30
      A(IJ)=A(I)
      A(I)=T
      T=A(IJ)
      V(IJ)=V(I)
      V(I)=VT
      VT=V(IJ)
 30   L=J
      IF (V(J).GE.VT) GO TO 50
      A(IJ)=A(J)
      A(J)=T
      T=A(IJ)
      V(IJ)=V(J)
      V(J)=VT
      VT=V(IJ)
      IF (V(I).LE.VT) GO TO 50
      A(IJ)=A(I)
      A(I)=T
      T=A(IJ)
      V(IJ)=V(I)
      V(I)=VT
      VT=V(IJ)
      GO TO 50
 40   A(L)=A(K)
      A(K)=TT
      V(L)=V(K)
      V(K)=VTT
 50   L=L-1
      IF (V(L).GT.VT) GO TO 50
      TT=A(L)
      VTT=V(L)
 60   K=K+1
      IF (V(K).LT.VT) GO TO 60
      IF (K.LE.L) GO TO 40
```

```
      IF (L-I.LE.J-K) GO TO 70
      IL(M)=I
      IU(M)=L
      I=K
      M=M+1
      GO TO 90
 70   IL(M)=K
      IU(M)=J
      J=L
      M=M+1
      GO TO 90
 80   M=M-1
      IF (M.EQ.0) RETURN
      I=IL(M)
      J=IU(M)
 90   IF (J-I.GT.10) GO TO 20
      IF (I.EQ.II) GO TO 10
      I=I-1
 100  I=I+1
      IF (I.EQ.J) GO TO 80
      T=A(I+1)
      VT=V(I+1)
      IF (V(I).LE.VT) GO TO 100
      K=I
 110  A(K+1)=A(K)
      V(K+1)=V(K)
      K=K-1
      IF (VT.LT.V(K)) GO TO 110
      A(K+1)=T
      V(K+1)=VT
      GO TO 100
      END
```

TECHNICAL REPORTS

Statistics Department

University of California, Berkeley

1. BREIMAN, L. and FREEDMAN, D. (Nov. 1981, revised Feb. 1982). How many variables should be entered in a regression equation? Jour. Amer. Statist. Assoc., March 1983, 78, No. 381, 131-136.

2. BRILLINGER, D. R. (Jan. 1982). Some contrasting examples of the time and frequency domain approaches to time series analysis. Time Series Methods in Hydrosciences. (A. H. El-Shaarawi and S. R. Esterby, eds.) Elsevier Scientific Publishing Co., Amsterdam, 1982, pp. 1-15.

3. DOKSUM, K. A. (Jan. 1982). On the performance of estimates in proportional hazard and log-linear models. Survival Analysis. (John Crowley and Richard A. Johnson, eds.) IMS Lecture Notes - Monograph Series, (Shanti S. Gupta, series ed.) 1982, 74-84.

4. BICKEL, P. J. and BREIMAN, L. (Feb. 1982). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. Ann. Prob., Feb. 1982, 11. No. 1, 185-214.

5. BRILLINGER, D. R. and TUKEY, J. W. (March 1982). Spectrum estimation and system identification relying on a Fourier transform. The Collected Works of J. W. Tukey, vol. 2, Wadsworth, 1985, 1001-1141.

6. BERAN, R. (May 1982). Jackknife approximation to bootstrap estimates. Ann. Statist., March 1984, 12 No. 1, 101-118.

7. BICKEL, P. J. and FREEDMAN, D. A. (June 1982). Bootstrapping regression models with many parameters. Lehmann Festschrift. (P. J. Bickel, K. Doksum and J. L. Hodges, Jr., eds.) Wadsworth Press, Belmont, 1983, 28-48.

8. BICKEL, P. J. and COLLINS, J. (March 1982). Minimizing Fisher information over mixtures of distributions. Sankhyā, 1983, 45, Series A, Pt. 1, 1-19.

9. BREIMAN, L. and FRIEDMAN, J. (July 1982). Estimating optimal transformations for multiple regression and correlation.

10. FREEDMAN, D. A. and PETERS, S. (July 1982, revised Aug. 1983). Bootstrapping a regression equation: some empirical results. JASA, 1984, 79, 97-106.

11. EATON, M. L. and FREEDMAN, D. A. (Sept. 1982). A remark on adjusting for covariates in multiple regression.

12. BICKEL, P. J. (April 1982). Minimax estimation of the mean of a mean of a normal distribution subject to doing well at a point. Recent Advances in Statistics, Academic Press, 1983.

14. FREEDMAN, D. A., ROTHENBERG, T. and SUTCH, R. (Oct. 1982). A review of a residential energy end use model.

15. BRILLINGER, D. and PREISLER, H. (Nov. 1982). Maximum likelihood estimation in a latent variable problem. Studies in Econometrics, Time Series, and Multivariate Statistics, (eds. S. Karlin, T. Amemiya, L. A. Goodman). Academic Press, New York, 1983, pp. 31-65.

16. BICKEL, P. J. (Nov. 1982). Robust regression based on infinitesimal neighborhoods. Ann. Statist., Dec. 1984, 12, 1349-1368.

17. DRAPER, D. C. (Feb. 1983). Rank-based robust analysis of linear models. I. Exposition and review.

18. DRAPER, D. C. (Feb 1983). Rank-based robust inference in regression models with several observations per cell.

19. FREEDMAN, D. A. and FIENBERG, S. (Feb. 1983, revised April 1983). Statistics and the scientific method, Comments on and reactions to Freedman, A rejoinder to Fienberg's comments. Springer New York 1985 Cohort Analysis in Social Research. (W. M. Mason and S. E. Fienberg, eds.).

20. FREEDMAN, D. A. and PETERS, S. C. (March 1983, revised Jan. 1984). Using the bootstrap to evaluate forecasting equations. J. of Forecasting, 1985, Vol. 4, 251-262.

21. FREEDMAN, D. A. and PETERS, S. C. (March 1983, revised Aug. 1983). Bootstrapping an econometric model: some empirical results. JBES, 1985, 2, 150-158.

22. FREEDMAN, D. A. (March 1983). Structural-equation models: a case study.

23. DAGGETT, R. S. and FREEDMAN, D. (April 1983, revised Sept. 1983). Econometrics and the law: a case study in the proof of antitrust damages. Proc. of the Berkeley Conference, in honor of Jerzy Neyman and Jack Kiefer. Vol I pp. 123-172. (L. Le Cam, R. Olshen eds.) Wadsworth, 1985.

24. DOKSUM, K. and YANDELL, B. (April 1983). Tests for exponentiality. Handbook of Statistics, (P. R. Krishnaiah and P. K. Sen, eds.) 4, 1984.

25. FREEDMAN, D. A. (May 1983). Comments on a paper by Markus.

26. FREEDMAN, D. (Oct. 1983, revised March 1984). On bootstrapping two-stage least-squares estimates in stationary linear models. Ann. Statist., 1984, 12, 827-842.

27. DOKSUM, K. A. (Dec. 1983). An extension of partial likelihood methods for proportional hazard models to general transformation models. Ann. Statist., 1987, 15, 325-345.

28. BICKEL, P. J., GOETZE, F. and VAN ZWET, W. R. (Jan. 1984). A simple analysis of third order efficiency of estimates. Proc. of the Neyman-Kiefer Conference, (L. Le Cam, ed.) Wadsworth, 1985.

29. BICKEL, P. J. and FREEDMAN, D. A. Asymptotic normality and the bootstrap in stratified sampling. Ann. Statist. 12 470-482.

30. FREEDMAN, D. A. (Jan. 1984). The mean vs. the median: a case study in 4-R Act litigation. JBES, 1985 Vol 3 pp. 1-13.

31. STONE, C. J. (Feb. 1984). An asymptotically optimal window selection rule for kernel density estimates. Ann. Statist., Dec. 1984, 12, 1285-1297.

32. BREIMAN, L. (May 1984). Nail finders, edifices, and Oz.

33. STONE, C. J. (Oct. 1984). Additive regression and other nonparametric models. Ann. Statist., 1985, 13, 689-705.

34. STONE, C. J. (June 1984). An asymptotically optimal histogram selection rule. Proc. of the Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer (L. Le Cam and R. A. Olshen, eds.), II, 513-520.

35. FREEDMAN, D. A. and NAVIDI, W. C. (Sept. 1984, revised Jan. 1985). Regression models for adjusting the 1980 Census. Statistical Science, Feb 1986, Vol. 1, No. 1, 3-39.

36. FREEDMAN, D. A. (Sept. 1984, revised Nov. 1984). De Finetti's theorem in continuous time.

37. DIACONIS, P. and FREEDMAN, D. (Oct. 1984). An elementary proof of Stirling's formula. Amer. Math Monthly, Feb 1986, Vol. 93, No. 2, 123-125.

38. LE CAM, L. (Nov. 1984). Sur l'approximation de familles de mesures par des familles Gaussiennes. Ann. Inst. Henri Poincaré, 1985, 21, 225-287.

39. DIACONIS, P. and FREEDMAN, D. A. (Nov. 1984). A note on weak star uniformities.

40. BREIMAN, L. and IHAKA, R. (Dec. 1984). Nonlinear discriminant analysis via SCALING and ACE.

41. STONE, C. J. (Jan. 1985). The dimensionality reduction principle for generalized additive models.

42. LE CAM, L. (Jan. 1985). On the normal approximation for sums of independent variables.

43. BICKEL, P. J. and YAHAV, J. A. (1985). On estimating the number of unseen species: how many executions were there?

44. BRILLINGER, D. R. (1985). The natural variability of vital rates and associated statistics. Biometrics, to appear.

45. BRILLINGER, D. R. (1985). Fourier inference: some methods for the analysis of array and nonGaussian series data. Water Resources Bulletin, 1985, 21, 743-756.

46. BREIMAN, L. and STONE, C. J. (1985). Broad spectrum estimates and confidence intervals for tail quantiles.

47. DABROWSKA, D. M. and DOKSUM, K. A. (1985, revised March 1987). Partial likelihood in transformation models with censored data.

48. HAYCOCK, K. A. and BRILLINGER, D. R. (November 1985). LIBDRB: A subroutine library for elementary time series analysis.

49. BRILLINGER, D. R. (October 1985). Fitting cosines: some procedures and some physical examples. Joshi Festschrift. 1986. D. Reidel.

50. BRILLINGER, D. R. (November 1985). What do seismology and neurophysiology have in common? - Statistics! Comptes Rendus Math. Rep. Acad. Sci. Canada. January, 1986.

51. COX, D. D. and O'SULLIVAN, F. (October 1985). Analysis of penalized likelihood-type estimators with application to generalized smoothing in Sobolev Spaces.

52. O'SULLIVAN, F. (November 1985). A practical perspective on ill-posed inverse problems: A review with some new developments. To appear in Journal of Statistical Science.

53. LE CAM, L. and YANG, G. L. (November 1985, revised March 1987). On the preservation of local asymptotic normality under information loss.

54. BLACKWELL, D. (November 1985). Approximate normality of large products.

55. FREEDMAN, D. A. (December 1985, revised Dec. 1986). As others see us: A case study in path analysis. Prepared for the Journal of Educational Statistics.

56. LE CAM, L. and YANG, G. L. (January 1986). Distinguished Statistics, Loss of information and a theorem of Robert B. Davies.

57. LE CAM, L. (February 1986). On the Bernstein - von Mises theorem.

58. O'SULLIVAN, F. (January 1986). Estimation of Densities and Hazards by the Method of Penalized likelihood.

59. ALDOUS, D. and DIACONIS, P. (February 1986). Strong Uniform Times and Finite Random Walks.

60. ALDOUS, D. (March 1986). On the Markov Chain simulation Method for Uniform Combinatorial Distributions and Simulated Annealing.

61. CHENG, C-S. (April 1986). An Optimization Problem with Applications to Optimal Design Theory.

62. CHENG, C-S., MAJUMDAR, D., STUFKEN, J. & TURE, T. E. (May 1986, revised Jan 1987). Optimal step type design for comparing test treatments with a control.

63. CHENG, C-S. (May 1986, revised Jan. 1987). An Application of the Kiefer-Wolfowitz Equivalence Theorem.

64. O'SULLIVAN, F. (May 1986). Nonparametric Estimation in the Cox Proportional Hazards Model.

65. ALDOUS, D. (JUNE 1986). Finite-Time Implications of Relaxation Times for Stochastically Monotone Processes.

66. PITMAN, J. (JULY 1986, revised November 1986). Stationary Excursions.

67. DABROWSKA, D. and DOKSUM, K. (July 1986, revised November 1986). Estimates and confidence intervals for median and mean life in the proportional hazard model with censored data.

68. LE CAM, L. and YANG, G.L. (July 1986). Distinguished Statistics, Loss of information and a theorem of Robert B. Davies (Fourth edition).

69. STONE, C.J. (July 1986). Asymptotic properties of logspline density estimation.

71. BICKEL, P.J. and YAHAV, J.A. (July 1986). Richardson Extrapolation and the Bootstrap.

72. LEHMANN, E.L. (July 1986). Statistics - an overview.

73. STONE, C.J. (August 1986). A nonparametric framework for statistical modelling.

74. BIANE, PH. and YOR, M. (August 1986). A relation between Lévy's stochastic area formula, Legendre polynomials, and some continued fractions of Gauss.

75. LEHMANN, E.L. (August 1986, revised July 1987). Comparing Location Experiments.

76. O'SULLIVAN, F. (September 1986). Relative risk estimation.

77. O'SULLIVAN, F. (September 1986). Deconvolution of episodic hormone data.

78. PITMAN, J. & YOR, M. (September 1987). Further asymptotic laws of planar Brownian motion.

79. FREEDMAN, D.A. & ZEISEL, H. (November 1986). From mouse to man: The quantitative assessment of cancer risks.

80. BRILLINGER, D.R. (October 1986). Maximum likelihood analysis of spike trains of interacting nerve cells.

81. DABROWSKA, D.M. (November 1986). Nonparametric regression with censored survival time data.

82. DOKSUM, K.J. and LO, A.Y. (November 1986). Consistent and robust Bayes Procedures for Location based on Partial Information.

83. DABROWSKA, D.M., DOKSUM, K.A. and MIURA, R. (November 1986). Rank estimates in a class of semiparametric two-sample models.

84. BRILLINGER, D. (December 1986). Some statistical methods for random process data from seismology and neurophysiology.

85. DIACONIS, P. and FREEDMAN, D. (December 1986). A dozen de Finetti-style results in search of a theory.

86. DABROWSKA, D.M. (January 1987). Uniform consistency of nearest neighbour and kernel conditional Kaplan - Meier estimates.

87. FREEDMAN, D.A., NAVIDI, W. and PETERS, S.C. (February 1987). On the impact of variable selection in fitting regression equations.

88. ALDOUS, D. (February 1987, revised April 1987). Hashing with linear probing, under non-uniform probabilities.

89. DABROWSKA, D.M. and DOKSUM, K.A. (March 1987, revised November 1987). Estimating and testing in a two sample generalized odds rate model.

90. DABROWSKA, D.M. (March 1987). Rank tests for matched pair experiments with censored data.

91. DIACONIS, P and FREEDMAN, D.A. (March 1987). A finite version of de Finetti's theorem for exponential families, with uniform asymptotic estimates.

92. DABROWSKA, D.M. (April 1987, revised September 1987). Kaplan-Meier estimate on the plane.

92a. ALDOUS, D. (April 1987). The Harmonic mean formula for probabilities of Unions: Applications to sparse random graphs.

93. DABROWSKA, D.M. (June 1987). Nonparametric quantile regression with censored data.

94. DONOHO, D.L. & STARK, P.B. (June 1987). Uncertainty principles and signal recovery.

95. RIZZARDI, F. (Aug 1987). Two-Sample t-tests where one population SD is known.

96. BRILLINGER, D.R. (June 1987). Some examples of the statistical analysis of seismological data. To appear in *Proceedings, Centennial Anniversary Symposium, Seismographic Stations, University of California, Berkeley.*

97. FREEDMAN, D.A. and NAVIDI, W. (June 1987). On the multi-stage model for cancer.

98. O'SULLIVAN, F. and WONG, T. (June 1987). Determining a function diffusion coefficient in the heat equation.

99. O'SULLIVAN, F. (June 1987). Constrained non-linear regularization with application to some system identification problems.

100. LE CAM, L. (July 1987, revised Nov 1987). On the standard asymptotic confidence ellipsoids of Wald.

101. DONOHO, D.L. and LIU, R.C. (July 1987). Pathologies of some minimum distance estimators.

102. BRILLINGER, D.R., DOWNING, K.H. and GLAESER, R.M. (July 1987). Some statistical aspects of low-dose electron imaging of crystals.

103. LE CAM, L. (August 1987). Harald Cramér and sums of independent random variables.

104. DONOHO, A.W., DONOHO, D.L. and GASKO, M. (August 1987). Macspin: Dynamic graphics on a desktop computer.

105. DONOHO, D.L. and LIU, R.C. (August 1987). On minimax estimation of linear functionals.

106. DABROWSKA, D.M. (August 1987). Kaplan-Meier estimate on the plane: weak convergence, LIL and the bootstrap.

107. CHENG, C-S. (August 1987). Some orthogonal main-effect plans for asymmetrical factorials.

108. CHENG, C-S. and JACROUX, M. (August 1987). On the construction of trend-free run orders of two-level factorial designs.

109. KLASS, M.J. (August 1987). Maximizing $E \max_{1 \le k \le n} S_k^+ / ES_n^+$: A prophet inequality for sums of I.I.D. mean zero variates.

110. DONOHO, D.L. and LIU, R.C. (August 1987). The "automatic" robustness of minimum distance functionals.

111. BICKEL, P.J. and GHOSH, J.K. (August 1987). A decomposition for the likelihood ratio statistic and the Bartlett correction — a Bayesian argument.

112. BURDZY, K., PITMAN, J.W. and YOR, M. (September 1987). Some asymptotic laws for crossings and excursions.

113. ADHIKARI, A. and PITMAN, J. (September 1987). The shortest planar arc of width 1.

114. RITOV, Y. (September 1987). Estimation in a linear regression model with censored data.

115. BICKEL, P.J. and RITOV, Y. (September 1987). Large sample theory of estimation in biased sampling regression models I.

116. RITOV, Y. and BICKEL, P.J. (September 1987). Unachievable information bounds in non and semiparametric models.

117. RITOV, Y. (October 1987). On the convergence of a maximal correlation algorithm using alternating projections.

118. ALDOUS, D.J. (October 1987). Meeting times for independent Markov chains.

119. HESSE, C.H. (October 1987). An asymptotic expansion for the mean of the passage-time distribution of integrated Brownian Motion.

120. DONOHO, D. (October 1987). Geometrizing rates of convergence, II.

121. BRILLINGER, D.R. (October 1987). Estimating the chances of large earthquakes by radiocarbon dating and statistical modelling. To appear in *Statistics a Guide to the Unknown.*

122. ALDOUS, D., FLANNERY, B. and PALACIOS, J.L. (November 1987). Two applications of urn processes: The fringe analysis of search trees and the simulation of quasi-stationary distributions of Markov chains.

123. DONOHO, D.L. and TAYLOR, B.M. (November 1987). Minimax risk for hyperrectangles.

124. ALDOUS, D. (November 1987). Stopping times and tightness II.

125. HESSE, C.H. (November 1987). The present state of a stochastic model for sedimentation.

126. DALANG, R.C. (December 1987). Optimal stopping of two-parameter processes on hyperfinite probability spaces.

127. DONOHO, D. and GASKO, M. (December 1987). Multivariate generalizations of the median and trimmed mean I.

128. DONOHO, D. and GASKO, M. (December 1987). Multivariate generalizations of the median and trimmed mean II.

Copies of these Reports plus the most recent additions to the Technical Report series are available from the Statistics Department technical typist in room 379 Evans Hall or may be requested by mail from:

Department of Statistics
University of California
Berkeley, California 94720

Cost: $1 per copy.