Estimating Optimal Transformations for Multiple Regression and Correlation: Rejoinder

Leo Breiman; Jerome H. Friedman

Journal of the American Statistical Association, Vol. 80, No. 391 (Sep., 1985), 614-619.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198509%2980%3A391%3C614%3AEOTFMR%3E2.0.CO%3B2-I

Journal of the American Statistical Association is currently published by American Statistical Association.

Rejoinder

LEO BREIMAN and JEROME H. FRIEDMAN

The discussants have provided additional insight into theoretical and practical aspects of the ACE methodology. Much of this insight will be very valuable when one applies the method and tries to interpret its results. We greatly appreciate the considerable effort that went into these comments and the high quality of the result. An immediate consequence is that we have revised the ACE algorithm and associated software to find suboptimal eigenfunctions, as suggested by Buja and Kass.

A central point running through the comments is that the enhanced power associated with using ACE as a tool brings with it a greater potential for abuse, if not carefully used. We completely agree. An automobile is far more powerful than a bicycle and in many circumstances much more useful. In careless hands, however, it can be much more dangerous. The discussants make important contributions towards alerting us to some of the dangers in overinterpreting the output of ACE. We discuss some others here. No doubt, as experience is gained, additional dangers, as well as diagnostics for detecting and dealing with them, will emerge. We are encouraged that Buja and co-workers have embarked on such a program of research.

As the comments point out, ACE is a complex mixture of theory and application, and it can be approached on both levels. There were two main issues in the comments regarding the theoretical foundations. One concerns the nature of maximal correlation, and the other involves the eigenvalue–eigenfunction structure. We deal with the latter later in our rejoinder.

1. NATURE OF MAXIMAL CORRELATION

Both Buja and Kass and Pregibon and Vardi note that if there is any "noiseless" component in the joint distribution of X and Y, the maximal correlation will be one. In the examples they give, "noiseless" corresponds to the existence of two or more separated regimes in the relationship between X and Y. More specifically, if there are, say, two rectangles $R_1 = I_1XJ_1$, $R_2 = I_2XJ_2$ such that $I_1 \cap I_2 = 0$, $J_1 \cap J_2 = 0$, and $P((X, Y) \in R_1) + P((X, Y) \in R_2) = 1$, then maximal correlation 1 can be obtained by using the transformations $\theta(Y) = a_1(Y \in J_1)$, $O(Y) = a_2(Y \in J_2)$, $\varphi(X) = \beta_1(X \in I_1)$, and $\varphi(X) = \beta_2(X \in I_2)$. Pregibon and Vardi point out that this is true even if one of $P((X, Y) \in R_i)$ is arbitrarily small.

There is no argument about the truth of this. Whether or not it is a crushing defect in maximal correlation as a theoretical construct is more debatable.It simply means that when separate regimes exist in the X, Y relationship, maximal correlation seizes on the existence of these separate regimes as the most salient piece of information and ignores the finer structure. To us, this does not seem insensible. In our view, the most important issue, cogently raised by Fowlkes and Kettenring, and seconded by Pregibon and Vardi, is that we do not understand yet well enough what ACE does on finite data sets. We made a stab at this in Section 5 of our article, but we only scratched the surface.

There are some things, however, that we have learned heuristically and for which we can give some rough explanations. We take up these things in turn. The first is the Fowlkes–Kettenring experiment in the low-correlation case; the second is the nature of the ACE solutions, as related particularly to the existence of noiseless components; and the third is the effect of outliers.

2. ACE BEHAVIOR IN THE PRESENCE OF LOW ASSOCIATIONS

We are glad that Fowlkes and Kettenring raised the issue of ACE's performance in the presence of very low association between variables. Even before receiving their comments, we had planned to include in our rejoinder a caution to ACE users not to put too much credence on the estimated transformations when the variables involved had low associations. This was brought to our attention some time ago, when Michael Axelrod of the Lawrence Livermore Laboratory showed us the results of an ACE run on data generated from the model $Y = X_1 + \cdots + X_{10} + Z$, where $X_1, \ldots, X_{10}, Z$ are independent unit normals. The transformations on the X_i were sometimes distinctly nonlinear (especially for small sample size), and we attributed this to the low association between Y and each individual X_i.

In the two-variable case, using near neighbor smooths, we can get heuristic results that illuminate some of the ACE behavior. Let N_S be the number of nearest neighbors over which the smoother averages. For linear smoothers, the transformation $\theta(Y)$ satisfies $\lambda\theta = S_YS_X\theta$, where $\lambda = \hat{\rho}^{*2}$ and S_YS_X is a matrix operator with elements determined by the data. Denote this matrix by U_{jk} so that $\lambda\theta(Y_j) = \Sigma_k U_{jk}\theta(Y_k)$. If the data are considered to be a random sample of size N from some underlying bivariate density $f_{XY}(X, Y)$, then U_{jk} can be written as $U_{jk} = E_{jk} + Z_{jk}$, where $E_{jk} = EU_{jk}$ and the Z_{jk} are the random data fluctuations in U_{jk}. If X and Y are independent, then $E_{jk} \simeq 1/N$. Since $\overline{\theta} = 0$,

$$\lambda\theta(Y_j) \simeq \frac{1}{N}\sum_k \theta(Y_k) + \sum_k Z_{jk}\theta(Y_k)$$

$$\simeq \sum_k Z_{jk}\theta(Y_k).$$

Therefore, in the case of independence, the transformations are completely determined by the noise in the data. Define $\overline{E} = \Sigma_{jk} E_{jk}/N^2$. Then a signal to noise measure for the matrix U can be defined by

$$S/N = \left[\sum_{j,k} (E_{jk} - \overline{E})^2/E \sum_{jk} Z_{jk}^2\right]^{1/2}.$$

For nearest neighbor smooths, we can show that

$$S/N \simeq \frac{N_S}{\sqrt{N}}\left[\int\int G^2(X, Y) f_X(X) f_Y(Y)\, dX\, dY - 1\right]^{1/2},$$

where $G(X, Y) = f_{XY}(X, Y)/f_X(X)f_Y(Y)$. If X and Y are unit normals with correlation ρ, then

$$S/N \simeq (|\rho|/\sqrt{1-\rho^2})(N_S/\sqrt{N}).$$

This indicates that for smaller values of $|\rho|$, the only way to obtain a decent S/N ratio is to increase the window size of the smoother.

We performed an experiment using a similar setup as in the Fowlkes and Kettenring comments. A set of 100 samples of 200 observations each was generated from an independent bivariate normal distribution. For each sample, the difference between the maximal correlation as estimated by ACE, and the absolute value of the sample correlation coefficient, was taken. When ACE was invoked using its variable window-size smoother (supersmoother, Friedman 1984), the median of these differences over the 100 trials was .258. When a constant window size equal to 50% of the data was used (largest internal supersmoother span), the median value of this difference was reduced to .095. Thus, increasing the window width (span) helps improve the estimates in this case. The estimated transformations also had much less structure.

It should be kept in mind that an estimated maximal correlation of .25 implies that the empirical model is explaining only 5% of the variance of the transformed responses. It is usually wise to avoid basing strong conclusions on any model having low explanatory power. ACE models are no exception.

A possible way to improve ACE is to adaptively increase the window sizes in the algorithm for variables whose association with $\theta(Y)$ is found to be small on the first few ACE iterations and to increase the window size for computing $\theta(Y)$ if $\widehat{e^2}$ is large. As we, and others, learn more about ACE behavior, improved implementations will be issued.

The results of Fowlkes and Kettenring for the circular t distribution are difficult to interpret without knowing the population maximal correlation or optimal transformations. Buja and Kass show that for a circular uniform distribution the maximal correlation is $\frac{1}{3}$ and the optimal transformations are parabolas symmetric about the vertical axis. Running ACE on the circular t distribution, Fowlkes and Kettenring also recover transformations that are symmetric about the vertical axis and a maximal correlation estimate about $\frac{1}{3}$. Judging from their plots, the optimal transformation shapes do not appear to correspond to parabolas.

3. ACE BEHAVIOR IN THE PRESENCE OF SEPARATE REGIMES

Another interesting question is what the ACE algorithm produces on finite data sets drawn from distributions having separate regimes. On finite data sets, it is clear that ACE maximizes the sample correlation of $\theta(Y)$ and $\varphi(X)$ under some smoothness constraints on θ and φ. For linear smooths, some approximate analysis of these constraints is possible. If near neighbor smooths are used, and both the smooth on X and Y use the same number N_S of nearest neighbors, then we can show that

$$\frac{1}{N}\sum_{k=1}^{N-1}(\theta(Y_{k+1}) - \theta(Y_k))^2 \le \frac{1}{(\rho^*)^2}\frac{N}{(N_S)^3},$$

and similarly for $\varphi(X)$, where ρ^* is the maximal sample correlation produced by ACE and $\|\theta\|_N = \|\varphi\|_N = 1$.

Thus an intrinsic part of our finite sample implementation of the ACE algorithm is that only functions with a certain degree of smoothness can emerge as solutions. The degree of smoothness produced depends on the window size of the smoother and ρ^*. (The smoother used in ACE has an adaptive window size, but this size can never fall below a specified minimum percentage of the total sample size.) This also gives another indication that the lower ρ^*, the noisier the transformations $\theta(Y)$ can be.

Now consider data from the model $Y = X + U$ $(0 \le X < 10)$, $Y = 11$ $(X = 10)$, where U is uniform on $[-1, 1]$ and X is uniform on $[0, 10]$ with total probability P, and concentrated on 10 with probability Q. For $Q \ll 1$, this is a bivariate distribution with a small noiseless component. The optimal transformations are discontinuous; that is, $\theta(Y) = a_1$ $(Y < 11)$, $\theta(Y) = a_2$ $(Y \ge 11)$, $\varphi(X) = \beta_1$ $(X < 10)$, and $\varphi(X) = \beta_2$ $(X = 10)$. A suboptimal but continuous set of transformations for Y is $\theta(Y) = a + bY$ $(Y \le Y_0)$, $\theta(Y) = a + bY_0$ $(Y > Y_0)$. Our analysis shows that if we try to approximate the discontinuous transformations with transformations having a certain degree of smoothness, they quickly develop a lower correlation than that produced by one of the suboptimal solutions. Thus, as Pregibon and Vardi speculate, the smoothness constraints underlie the fact that on data ACE either transforms small noiseless components smoothly into the main structure of the data (if it can) or they show up having the same effect as outliers (see below).

If the data separate into two or more regimes, none of which are small, then it is interesting to track what ACE does. We performed two experiments. In the first, X is uniform on $[0, 10]$. For $X \in [0, 5]$, Y is $N(0, 1)$. For $X \in [5, 10]$, Y is $N(\mu, 1)$, $\mu > 0$. We did this for $\mu = 3, 5, 10$. The results were consistent. Since there was not a good suboptimal solution to fasten onto, ACE did its best to smoothly approximate the optimal discontinuous solution and got $\hat{\rho}^* = .99$. The graphs

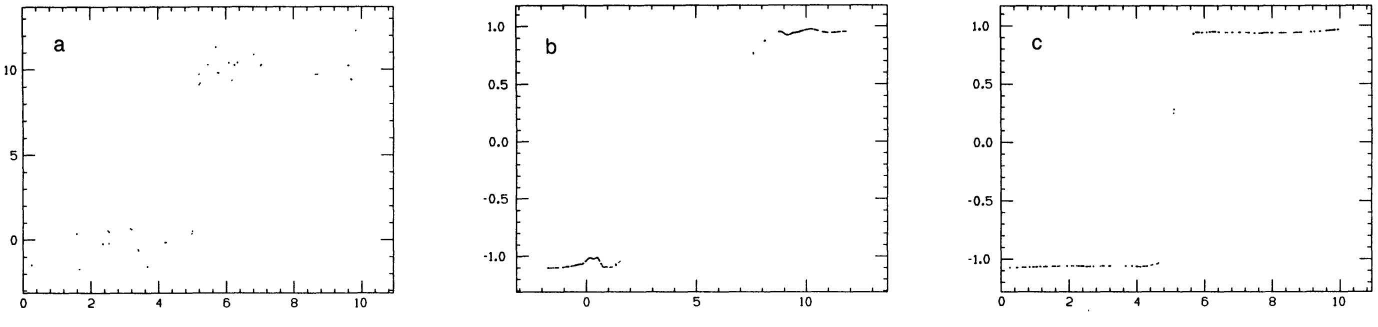

Figure 1. Piecewise Constant Data: (a) Y Versus X; (b) θ(Y) Versus Y; (c) φ(X) Versus X.

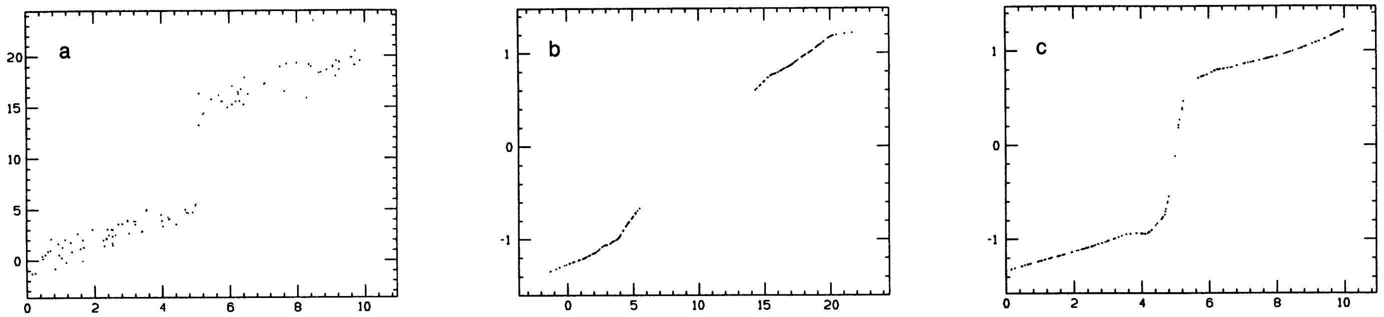

Figure 2. Piecewise Linear Data: (a) Y Versus X; (b) θ(Y) Versus Y; (c) φ(X) Versus X.

of the data and transformations for the $\mu = 10$ run are shown in Figure 1.

In the second example, there is some data structure. Again X is uniform on [0, 10]. For $X \in [0, 5]$, $Y = X + Z$, $Z \sim N(0, 1)$. For $X \in [5, 10]$, $Y = X + Z$, $Z \sim N(\mu, 1)$. Again, the results for various values of $\mu \geq 3$ were consistent. In this situation, there is a suboptimal transformation given by $\theta(Y) = Y$ and $\varphi(X) = X$ $(0 \leq X \leq 5)$, $\varphi(X) = X + \mu$ $(5 < X \leq 10)$. This transforms the data into the perfectly linear model $\theta(Y) = \varphi(X) + Z$, $Z \sim N(0, 1)$, with a θ, φ correlation of .997. Even though the $\varphi(X)$ transformation is discontinuous, ACE finds that it can get a higher correlation (.99) by smoothly approximating this suboptimal transformation than by smoothly approximating the optimal solution. The data and transformation plots for $\mu = 10$ are given in Figure 2.

In general, if the data fall into separate regimes, and each regime has some good structure in it, the ACE solution will track these structures and paste them together if by so doing it can get a higher $\hat{R}^2$ than by a smooth approximation to the dominant discontinuous solution. There will be discontinuities across the pasted edges, and appreciable sharp changes in the ACE transformations should alert the analyst to the possible existence of disparate regimes in the data.

4. SUBOPTIMAL SOLUTIONS

The comment by Buja and Kass on the importance of suboptimal eigenfunctions is both interesting and important. Buja and Kass describe situations in which more than one set of transformations could have approximately equal, and sometimes quite high, eigenvalues. When this occurs it can be both good news and bad news. The bad news is the instability that may be introduced in the estimation. The good news is that the data analyst has at his disposal several good sets of transformations from which to choose. Each "orthogonal" set may provide different information concerning the relationship between the response and predictors.

The example given by Pregibon and Vardi provides a good illustration of the additional information that can be given by suboptimal solutions. Here $Y = X_1X_2$, with X_2 taking only positive values whereas X_1 takes on both positive and negative values. In this case the predictive relationship is exact (no noise) so that the joint distribution is degenerate. As discussed in our article, optimal transformations are often not unique for degenerate distributions, and the ambiguity is fundamental in this case. The introduction of a little noise removes the degeneracy, providing unique optimal transformations.

In this spirit consider the problem

$$Y = X_1X_2e^{a\varepsilon}, \tag{1}$$

where a is a positive constant and ε is a random variable with a standard normal distribution; X_2 has a lognormal distribution; and X_1 has a distribution that is symmetric about zero with $|X_1|$ having a lognormal distribution. The optimal transformations in this case (modulo constants) are $\theta(Y) = I[Y \geq 0]$, $\varphi_1(X_1) = I[X_1 \geq 0]$, and $\varphi(X_2) \equiv 0$. The resulting eigenvalue (maximal correlation) is 1. That is, the sign of Y is perfectly predicted by the sign of X_1. However, the transformations $\theta(Y) = \log|Y|$, $\varphi_1(X_1) = \log|X_1|$, and $\varphi_2(X_2) = \log X_2$ are also eigenfunctions, since $\log|Y| = \log|X_1| + \log X_2 + a\varepsilon$, with the resulting transformed variables all having normal distributions. This second set of transformations will not have an eigenvalue equal to 1 unless $a = 0$ (no noise). Thus, these transformations correspond to a suboptimal eigensolution. Note that both sets of transformations are informative; the first set gives the sign of Y, and the second set gives information concerning its absolute value. Taken together, both sets pin down the predictive relationship quite well (provided the value of a is not too large).

In order to get a feeling as to the nature of suboptimal eigensolutions, Table 1 shows the transformations corresponding to the four largest eigensolutions for this problem and their associated multiple correlation R, for $a = .5$.

For this problem the third and fourth eigensolutions provide little additional information.

Only a minor modification to the ACE algorithm is required to have it produce suboptimal transformations. The standardization of the current estimate of $\theta(Y)$ is replaced by a more general orthogonalization in which its projections on the previously determined (more) optimal Y-transformations are subtracted away. It should be noted that the standardization of $\theta(Y)$ in the present algorithm can be thought of as a way of keeping the solutions orthogonal to the trivial solutions $\theta(Y) \equiv \varphi_i(X_i) \equiv 0$. We have made this change in the current release of the ACE software. With this change the program is able to find all the solutions listed in Table 1 (as well as many more) from

Table 1. Eigensolutions

Eigen-solution	R	θ(Y)	φ₁(X₁)	φ₂(X₂)
1	1.0	sign(Y)	sign(X_1)	0
2	.95	log$\lvert Y\rvert$	log$\lvert X_1\rvert$	log(X_2)
3	.72	sign(Y) · log$\lvert Y\rvert$	sign(X_1) · log$\lvert X_1\rvert$	0
4	.61	sign($\lvert Y\rvert$ − 1) · log$\lvert Y\rvert$	sign($\lvert X_1\rvert$ − 1) · log$\lvert X_1\rvert$	sign(X_2 − 1) · log X_2

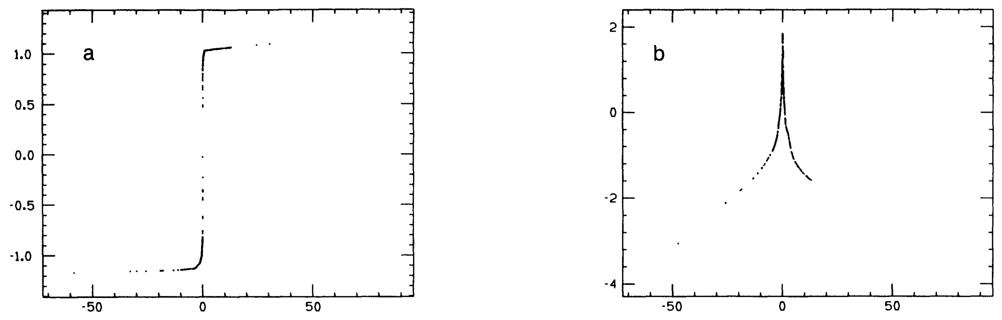

Figure 3. $Y = X_1X_2C^{.5\varepsilon}$: (a) $\hat{\theta}(Y)$ for the First Eigensolution; (b) $\hat{\theta}(Y)$ for the Second Eigensolution.

data simulated from (1). Figure 3 shows the response transformations $\hat{\theta}(Y)$ corresponding to the first two eigensolutions obtained by running ACE on a sample of 500 observations generated from (1).

For the Boston housing data example, the first three eigensolutions have $\hat{e}^2$ (unexplained variance) values of .11, .28, and .45, respectively. For the air pollution example these values are .22, .63, and .84. The solutions corresponding to the largest eigenvalues (smallest $\hat{e}^2$) are the ones shown and discussed in our article. Besides having the most explanatory power, these solutions are generally preferred, since they are the only ones with a monotone response transform $\theta(Y)$. Note that if the optimal eigensolution results in a monotone $\theta(Y)$, then the suboptimal ones cannot, due to the orthogonality (and zero mean) requirement. A real benefit from finding suboptimal solutions would arise if the optimal solution gave a nonmonotone $\theta(Y)$, whereas a close suboptimal one had a monotone response transform.

We realized, early in the development of ACE, that the suboptimal eigenfunctions should be of some use. In fact, Breiman and Ihaka (1985) use them in a generalization of discriminant analysis. One particular point that the analysis of Buja and Kass forcibly makes is that the analyst should know if there are other additive models around with error only slightly greater than the model selected by ACE. A question that we have considered but not satisfactorily resolved is how to incorporate all solution models (i.e., all sets of eigenfunctions) into a single overall model that makes sense in a theoretical context. A sound resolution of this issue would be a significant step forward in the ACE methodology.

5. ROBUSTNESS

Robustness is an important issue raised by Buja and Kass. ACE (in its current formulation) is definitely not robust to extreme outliers. Peter Bickel has shown us some applications of ACE to data for which outliers significantly distorted the estimated transformations. This is also illustrated in the data example provided by Pregibon and Vardi. They use it as an illustration of the success of heuristic reasoning over automated methods. Although we agree with their conclusion that automated tools and heuristic reasoning must work together, this example is actually illustrative of lack of robustness. Inspection of the marginal distribution of Y reveals three large positive outlying values. Removing those observations results in the joint XY distribution (scatterplot) shown in Figure 4(a). Application of ACE to this reduced ($N = 59$) data set results in the transformations plotted in Figures 4(b) and 4(c). The resulting $\hat{\rho}^* = .95$. The plots suggest monotonic transformations with decreasing curvature, looking much like logarithmic transformations. Of course other transformations, such as square roots, also come to mind. However, taking the square roots of both X and Y to define new variables and running ACE results in appreciably concave graphs, indicating the need for a stronger transformation. When logarithms are used to define new variables, however, and ACE is run on these new variables, the resulting transformations were very nearly linear with a resulting correlation of .96. This again indicates that ACE finds the log transformation nearly optimal for this problem. Thus ACE, in this case, gives results far more precise than a simple heuristic statement. (It should be noted that it is easy to imagine situations in which the variables are highly skewed and extend over many powers of 10, but for which logarithms are far from the best transformations.) Until a good way is found to make ACE robust, it should be used with a great deal of caution in the suspected presence of extreme outliers.

6. THE FOWLKES–KETTENRING DATA

Fowlkes and Kettenring point out that on their three-variable data set $Y(CCS)$, X_1 (residence local), and X_2 (residence metro), ACE gave some ambiguous results. That is, starting with log

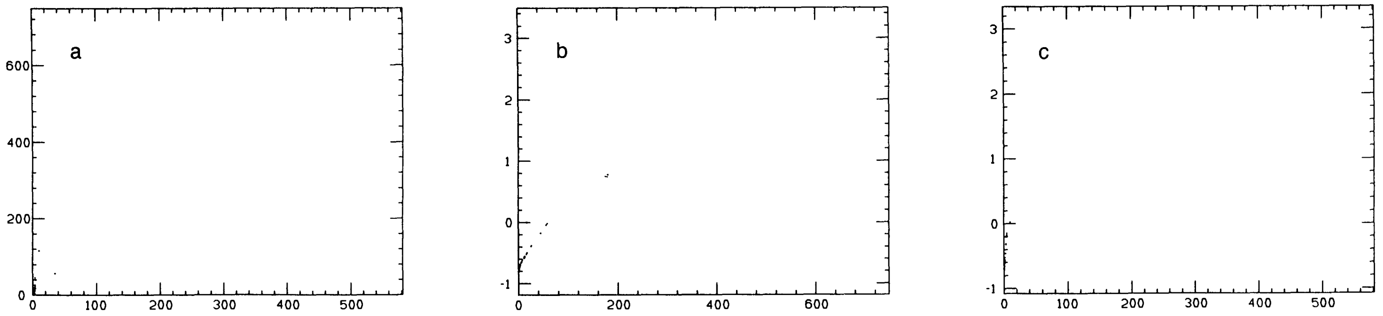

Figure 4. Pregibon and Vardi Data Example, With Outliers Removed: (a) Y Versus X; (b) $\theta(Y)$ Versus Y; (c) $\varphi(X)$ Versus X.

Table 2. Transformations

$\theta(Y)$	$\varphi_1(X_1)$	$\hat{R}$
Y	X_1	.9829
$Y^{.9}$	$X_1^{.9}$	.9839
$Y^{.8}$	$X_1^{.8}$	.9840
$Y^{.7}$	$X_1^{.7}$	.9841
$Y^{.6}$	$X_1^{.6}$	.9840
$Y^{.5}$	$X_1^{.5}$	.9838
$\log(Y)$	$\log(X_1)$	.9810

Y, $\log X_1$ transformations, ACE stayed with these transformations, but starting with linear transformations, ACE stayed with the linear.

There is one large outlier in the data, and since outliers do not sit well with ACE, we deleted it. We noticed that X_2 is only weakly associated with Y, but there is a strong linear association between Y and X_1. Thus, within a wide range, any model that applies the same transformation to both Y and X_1 should have a high $\hat{R}$. To check this, we fixed various transformations for Y and X_1 and held them constant, letting ACE find only the transformation on X_2. The results are in Table 2.

Now the cause of ambiguous behavior pointed out by Fowlkes and Kettenring becomes clear. The termination criterion in ACE stops the iteration process when the increase in $\hat{R}$ falls below a threshold value. Given that the $\hat{R}$ for the range of models given in Table 2 varies so slightly, ACE would find very little $\hat{R}$ improvement starting from any of these models and terminate quickly with Y, X_1 transformations close to the initial ones.

The problem disappeared when we reduced the threshold value in the termination rule. ACE converged to the same set of Y, X_1, X_2 transformations, starting from any of the initial transformations given in Table 2. This set is given in Figure 5. The transformations for Y, X_1 are slightly concave, indicating transformations somewhere near the Y, X_1 transformations that give the highest $\hat{R}$ in Table 2. The $\hat{R}$ achieved by these transformations is .9849.

With this relative insensitivity of $\hat{R}$ to the choice of the Y, X_1 transformation, however, it seems reasonable that the analyst select the transformation to be used on other grounds, and we agree that the choice made by Fowlkes and Kettenring is based on some sensible heuristics. (We thank Fowlkes and Kettenring for permitting us to use their data and Paul Tukey for sending it to us.)

7. PARTIAL RESIDUAL PLOTS

Fowlkes and Kettenring raise the issue of comparing ACE to familiar partial residual methods. Smoothing a partial residual plot as illustrated by Fowlkes and Kettenring can be viewed as corresponding to the following two steps. First a linear (least squares) regression is performed minimizing $E[Y - a_0 - \sum_{i=1}^{p} a_i X_i]^2$ with respect to the parameters $a_0 \cdots a_p$ yielding estimates $\hat{a}_0 \cdots \hat{a}_p$. (Here, $Y, X_1, \ldots, X_p$ represent either the variables themselves or predetermined transformations on them.) Then $E[Y - \hat{a}_0 - \sum_{i \neq k} \hat{a}_i X_i - \varphi_k(X_k)]^2$ is minimized with respect to φ_k (smoothing of the partial residual plot), given the coefficient values $\hat{a}_i$ obtained in the linear regression. This estimate $\hat{\varphi}_k(X_k)$ is used to judge the need for (and suggest) a transformation for X_k. This can be done for each X_k $(1 \leq k \leq p)$ in turn.

A problem with this procedure is that in searching for $\varphi_k(X_k)$, the coefficients of the other explanatory variables are not allowed to adjust to the presence of the transformation on X_k. This limitation of partial residual methods was pointed out by Colin Mallows at the 1983 Princeton regression meeting. The problem can be overcome by minimizing $E[Y - a_0 - \sum_{i \neq k} a_i X_i - \varphi_k(X_k)]^2$ simultaneously with respect to the coefficient values a_i $(i \neq k)$ and the function φ_k. If transformations are being considered on more than one variable, this procedure can be done for several of the X_k individually. However, this still does not allow the transformation being considered for one variable to take into account the possibility of transformations on other variables. This is accomplished by minimizing $E[Y - \sum_{i=1}^{p} \varphi_i(X_i)]^2$ simultaneously with respect to all $\varphi_k(X_k)$ for which transformations are being considered. Thus, in fixing up partial residual plots so as to validly locate nonlinearities, one is led to a restricted ACE procedure. Full (unrestricted) ACE goes one step further by simultaneously minimizing with respect to the response transformation $\theta(Y)$ as well, allowing it to account for possible transformations on all the predictors X_k as well as vice versa.

In the preceding sense ACE generalizes partial residual methods and should result in a procedure that is more sensitive to combined nonlinear effects in the explanatory variables. The two methods give similar results in the example presented by Fowlkes and Kettenring, because there are only two explanatory variables and the transformations on Y, X_1 are already close to optimal.

8. OTHER REMARKS

Finally, we would like to relate some additional cautionary notes concerning the use of ACE from our experiences. As noted before, the "shapes" of the transformations for variables weakly associated with the response can be highly variable. They can depend somewhat on the order in which the variables

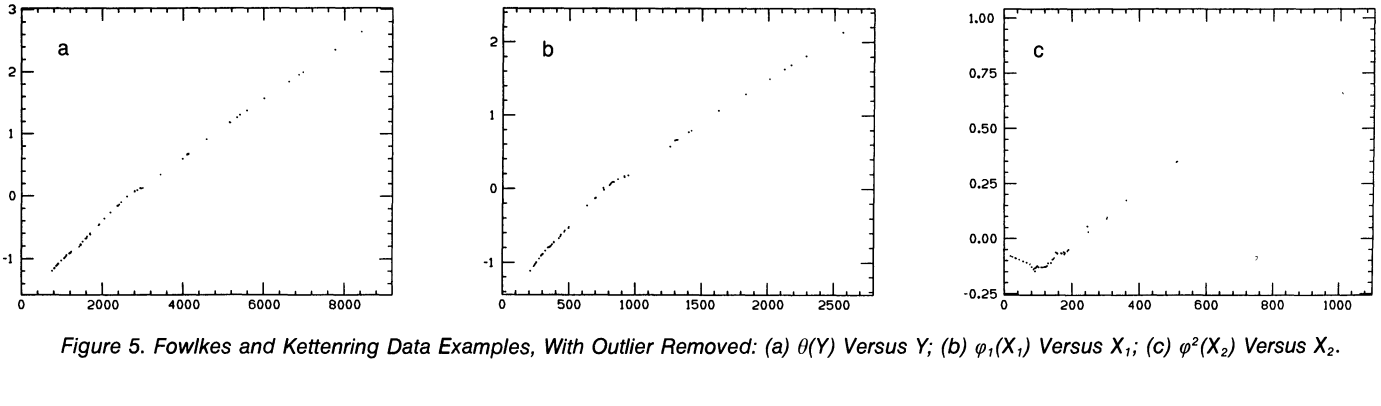

Figure 5. Fowlkes and Kettenring Data Examples, With Outlier Removed: (a) $\theta(Y)$ Versus Y; (b) $\varphi_1(X_1)$ Versus X_1; (c) $\varphi^2(X_2)$ Versus X_2.

are entered, to some extent on the starting transformations, and on small changes in the data set. In such situations it is wise to use a very tight iteration convergence threshold and to try several variable orderings and sets of starting transformations to see if it makes a substantial difference. Considerable care must be taken in placing strong interpretations on the particular transformations found by ACE unless some assessment of their variability is made. Bootstrapping suggests itself as an expensive but nevertheless effective tool for this purpose.

The output of ACE should never be used blindly. Sometimes it can be used to signal anomolies in the data—for example, the appropriateness of a (single) regression model. One such signature is an estimated response transformation $\theta(Y)$ that is significantly nonmonotonic. The example provided by Pregibon and Vardi (figure 2 in their comment) is a nice illustration of this. Although one might be motivated to restrict the response transformation to be monotonic, it is wise to first run the unrestricted procedure. A strongly nonmonotonic $\theta(Y)$ should signal further investigation.

We thank Pregibon and Vardi for bringing the work of Czári and Fischer to our attention. We have added it to our citations. Finally, we thank Buja and Kass, Fowlkes and Kettenring, and Pregibon and Vardi for their valuable contributions. Through them we have learned more about the ACE method, learned to judge its output with a bit more caution, and modified our procedure (and software) to make it more useful.

9. EPILOG

As seen in the comments, the ACE methodology offers a fertile and exciting field for exploration. For instance, Fowlkes and Kettenring point out a number of possible extensions of ACE to other interesting multivariate problems. We note that besides a continuing study of the eigenvalue–eigenfunction problem, Buja is also collaborating with Stuetzle on a principal components version of ACE. One or another of us has collaborated on extensions of the ACE methodology to discriminant analysis (Breiman and Ihaka 1985), factor analysis (Koyak 1985), and a version called PACE (predictive ACE), which directly predicts Y as a function of $\sum_{i=1}^{p} \varphi_i(X_i)$ (Friedman and Owen 1985). Other extensions have been made to time-series analysis (Owen 1983), generalized linear models (Hastie and Tibshirani 1984), and sparse contingency tables (Marhoul 1984). It has been applied to seismic data (Brillinger and Preisler 1984), failure data (Eason et al. 1984), and, in unpublished research that we know of, nuclear test data and economic time series.

ADDITIONAL REFERENCES

Breiman, L., and Ihaka, R. (1985), "Nonlinear Discriminant Analysis via Scaling and ACE," unpublished manuscript (submitted to the *Journal of the American Statistical Association*).

Brillinger, D. R., and Preisler, H. K. (1984), "An Exploratory Analysis of the Joyner–Boore Attenuation Data," *Bulletin of the Seismological Society of America*, 74, 1441–1450.

Eason, E. D., Nelson, E. E., and Patterson, S. D. (1984), "Stress Corrosion Cracking in Steam Turbine Disks," technical report, Failure Analysis Associates, Palo Alto, CA.

Friedman, J. H. (1984), "A Variable Span Smoother," Report LCS005, Stanford University, Dept. of Statistics.

Friedman, J. H., and Owen, A. (1985), "Predictive ACE," unpublished manuscript.

Hastie, T., and Tibshirani, R. (1984), "Generalized Additive Models," Report LCS002, Stanford University, Dept. of Statistics.

Koyak, R. (1985), "Nonlinear Dimensionality Reduction," unpublished Ph.D. thesis, University of California, Berkeley, Dept. of Statistics.

Marhoul, J. C. (1984), "A Model for Large Sparse Contingency Tables," Report LCSD013, Stanford University, Dept. of Statistics.

Owen, A. (1983), "Optimal Transformations for Autoregressive Time Series Models," Report ORION 020, Stanford University, Dept. of Statistics.