Journal of Data Science **2**(2004), 329-346

Estimating Optimal Transformations for Multiple Regression Using the ACE Algorithm

Duolao Wang[1] and Michael Murphy[2]
[1]*London School of Hygiene and Tropical Medicine*
and [2]*London School of Economics*

Abstract: This paper introduces the alternating conditional expectation (ACE) algorithm of Breiman and Friedman (1985) for estimating the transformations of a response and a set of predictor variables in multiple regression that produce the maximum linear effect between the (transformed) independent variables and the (transformed) response variable. These transformations can give the data analyst insight into the relationships between these variables so that relationship between them can be best described and non-linear relationships can be uncovered. The power and usefulness of ACE guided transformation in multivariate analysis are illustrated using a simulated data set as well as a real data set. The results from these examples clearly demonstrate that ACE is able to identify the correct functional forms, to reveal more accurate relationships, and to improve the model fit considerably compared to the conventional linear model.

Key words: Alternating conditional expectation (ACE) algorithm, nonparametric regression, transformation.

1. Introduction

In regression analysis, we try to explain the effect of one or more independent variables (predictors or covariates) on a dependent variable (response). The initial stages of data analysis often involve exploratory analysis. Instead of imposing preconceived models, we seek insight into the nature of relationships in the data set and, if possible, the underlying phenomena that might have produced the observed data values. Unfortunately traditional multiple regression techniques are limited in this respect since they usually require a priori assumptions about the functional forms that relate the response and predictor variables.

The objective of fully exploring and explaining the effect of covariates on a response variable in regression analysis is facilitated by properly transforming the independent variables. A number of parametric transformations for continuous variables in regression analysis have been suggested (Box and Tidwell, 1962;

Kruskal, 1965; Mosteller and Tukey, 1977; Cook and Weisberg, 1982; Carroll and Ruppert, 1988; Royston, 2000.)

Box and Cox (1964) discussed response variable transformations with an emphasis on a parameterised family of power transformations that has become known as the Box-Cox family. Though often used in practice, the Box-Cox method is confined to transformations of the response variable and does not facilitate transformations of the covariates. Kruskal (1965) suggested a technique similar to the Box-Cox method that uses isotonic regression to estimate the transformation of the response variable. (Box and Tidwell (1962) studied parametric transformations of the predictors. Mosteller and Tukey (1977) developed a data-analytic approach. Others have studied parametric families of transformations and associated diagnostic methods, see Cook and Weisberg (1982) and Carroll and Ruppert (1988). Royston (2000) summarised the methods for parametrically modelling the effect of a continuous covariate in medicine and epidemiology.

Non-parametric curve-fitting techniques have also been proposed for variable transformations. Those include bin smoothers, high-degree and fractional polynomials Royston and Altman (1994), cubic regression splines Durrleman and Simon (1989), cubic smoothing splines Hastie and Tibshirani (1990), Green and Silverman (1994), kernel smoothers Bowman and Assalini (1997). Estimating the optimal transformation is the primary motivation for the use of non-parametric regression techniques, which make few assumptions about the regression surface (Friedman and Stuetzle, 1981; Breiman and Friedman, 1985; Hastie and Tibshirani, 1990). A comprehensive coverage of these methods and their applications was given by Härdle (1992). Non-parametric regression techniques are based on successive refinements by attempting to define the regression surface in an iterative fashion while remaining 'data-driven' as opposed to 'model-driven'. These non-parametric regression methods can be broadly classified into those which do not transform the response variable (such as Generalised Additive Models) and those which do (such as Alternating Conditional Expectations (ACE)).

In this paper, we introduce the ACE algorithm developed by Breiman and Friedman (1985) for estimating optimal transformations for both response and independent variables in regression and correlation analysis, and illustrate through two examples that usefulness of ACE guided transformation in multivariate analysis. The power of the ACE approach lies in its ability to recover the functional forms of variables and to uncover complicated relationships.

2. The ACE Algorithm

The general form of a linear regression model for p independent variables (predictors), say X_1, X_2,...,X_p, and a response variable Y is given by

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon \tag{2.1}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients to be estimated, and ϵ is an error term. The (2.1) therefore assumes that the response, Y, is a combination of linear effects of X_1, X_2,...,X_p and a random error component ϵ.

Conventional multiple regression requires a linear functional form to be presumed a priori for the regression surface, thus reducing the problem to that of estimating a set of parameters. This linear parametric approach can be successful provided the assumed model is appropriate. When the relationship between the response and predictor variables is unknown or inexact, linear parametric regression can yield erroneous and even misleading results. This is the primary motivation for the use of non-parametric regression techniques, which make few assumptions about the regression surface (Friedman and Stuetzle, 1981).

These non-parametric regression methods can be broadly classified into those which do not transform the response variable such as Generalised Additive Models and those which do such as the ACE, which is the focus of our discussion in this paper.

An ACE regression model has the general form:

$$\theta(Y) = \alpha + \sum_{i=1}^{p} \phi_i(X_i) + \epsilon$$

where θ is a function of the response variable, Y, and ϕ_i are functions of the predictors $X_i, i = 1, ..., p$. Thus the ACE model replaces the problem of estimating a linear function of a p-dimensional variable $\mathbf{X}$=$(X_1, X_2, \ldots, X_p)$ by estimating p separate one-dimensional functions, ϕ_i, and θ using an iterative method. These transformations are achieved by minimising the unexplained variance of a linear relationship between the transformed response variable and the sum of transformed predictor variables.

For a given data set consisting of a response variable Y and predictor variables $X_1, \ldots, X_p$, the ACE algorithm starts out by defining arbitrary measurable mean-zero transformations $\theta(Y)$, $\phi_1(X_1), \ldots, \phi_p(X_p)$. The error variance ($\epsilon^2$) that is not explained by a regression of the transformed dependent variable on the sum of transformed independent variables is (under the constraint, $E[\theta^2(Y)] = 1$)

$$\epsilon^2(\theta, \phi_1, ..., \phi_p) = E\{[\theta(Y) - \sum_{i=1}^{p} \phi_i(X_i)]\}^2$$

The minimisation of ϵ^2 with respect to $\phi_i(X_i), \ldots, \phi_p(X_p)$ and $\theta(Y)$ is carried out through a series of single-function minimisations, resulting in the following equations

$$\phi_i(X_i) = E[\theta(Y) - \sum_{j \neq i}^{p} \phi_j(X_j)|X_i]$$

$$\theta(Y) = E[\sum_{i=1}^{p} \phi_i(X_i)|Y]/||E[\sum_{i=1}^{p} \phi_i(X_i)|Y]|| \tag{2.2}$$

Two basic mathematical operations involved in here are conditional expectations and iterative minimisation and hence, the name *alternating conditional expectations*. The final $\phi_i(X_i), i = 1, \ldots, p$, and $\theta(Y)$ after the minimisation are estimates of the optimal transformation $\phi_i^*(X_i), i = 1, \ldots, p$, and $\theta^*(Y)$. In the transformed space, the response and predictor variables are related as follows

$$\theta^*(Y) = \sum_{i=1}^{p} \phi_i^*(X_i) + e^*$$

where e^* is the error not captured by the use of the ACE transformations and is assumed to have a normal distribution with zero mean. The minimum regression error, e^*, and maximum multiple correlation coefficient, ρ^*, are related by $e^{*2} = 1 - \rho^{*2}$.

These optimal ACE transformations are derived solely from the given data and do not require a priori assumptions of any functional form for the response or predictor variables and thus provide a powerful tool for exploratory data analysis. Moreover, the ACE algorithm can handle variables other than continuous predictors such as categorical (ordered or unordered), integer and indicator variables. These present no additional computational complications. For categorical variables, the ACE transformations can be regarded as estimating optimal scores for each value level of the variable and therefore may be used to combine groups in a parsimonious way.

3. Simulated Example

The ACE algorithm for multiple linear regression can be implemented using the **ace** function in the S-PLUS statistical package (Venables and Ripley, 2002).

In this section, we apply the ACE technique to a synthetic example – i.e., case for which we know the correct answers – to demonstrate how the ACE algorithm can be used to identify the functional relationship between dependent and independent variables.

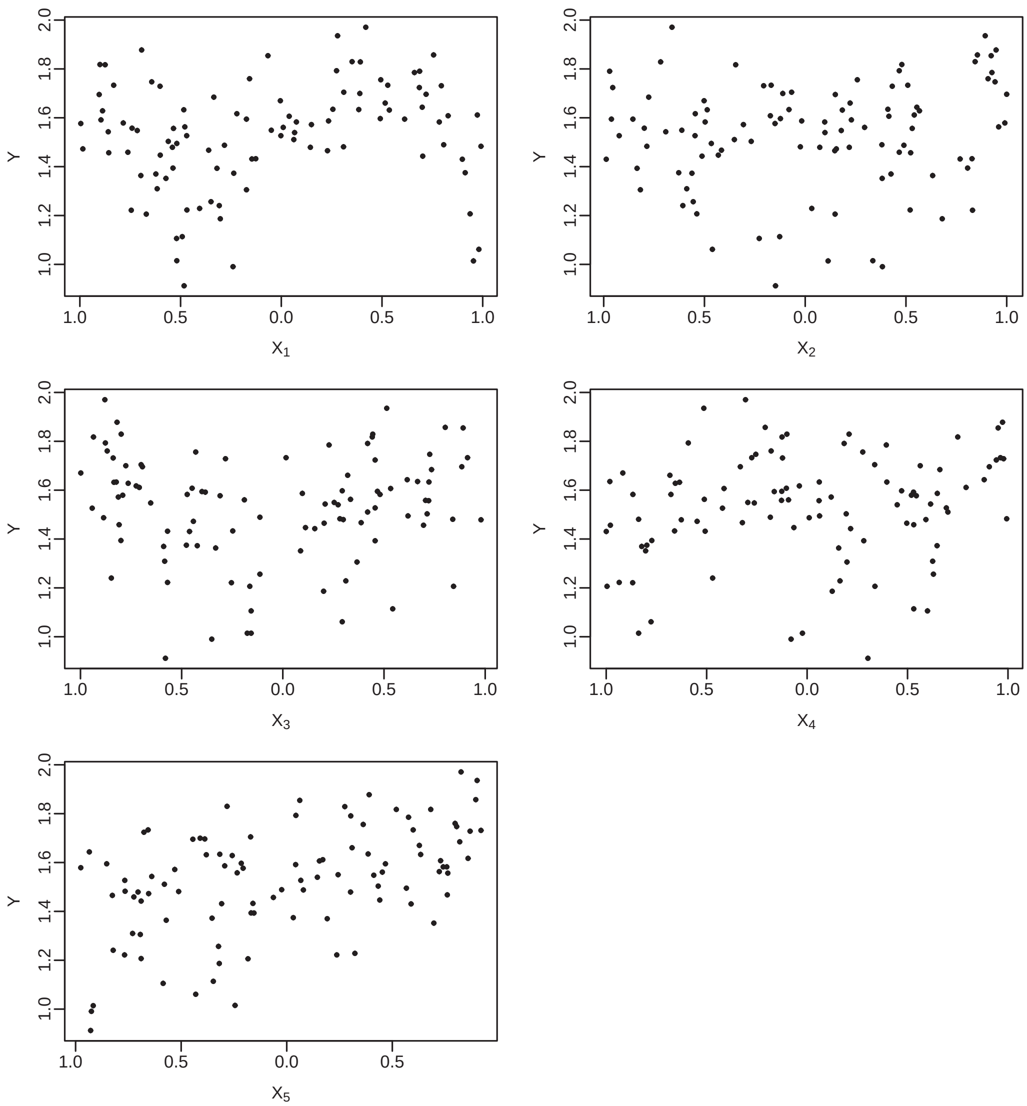

Figure 1. Scatterplots of simulated dataset

Our synthetic example is a multivariate case with five predictor and 100 observations generated from the following model

$$Y = \log[4 + \sin(4X_1) + |X_2| + X_3^2 + X_4^3 + X_5 + .1\epsilon] \tag{3.1}$$

where X_1, X_2, X_3, X_4 and X_5 are independently drawn from a uniform distribution $U(-1, 1)$ and ϵ is independently drawn from a standard normal distribution $N(0, 1)$. If we are simply given the data values for Y, X_1, X_2, X_3, X_4 and X_5 without any knowledge of functional relationship in (3.1), we might then try to

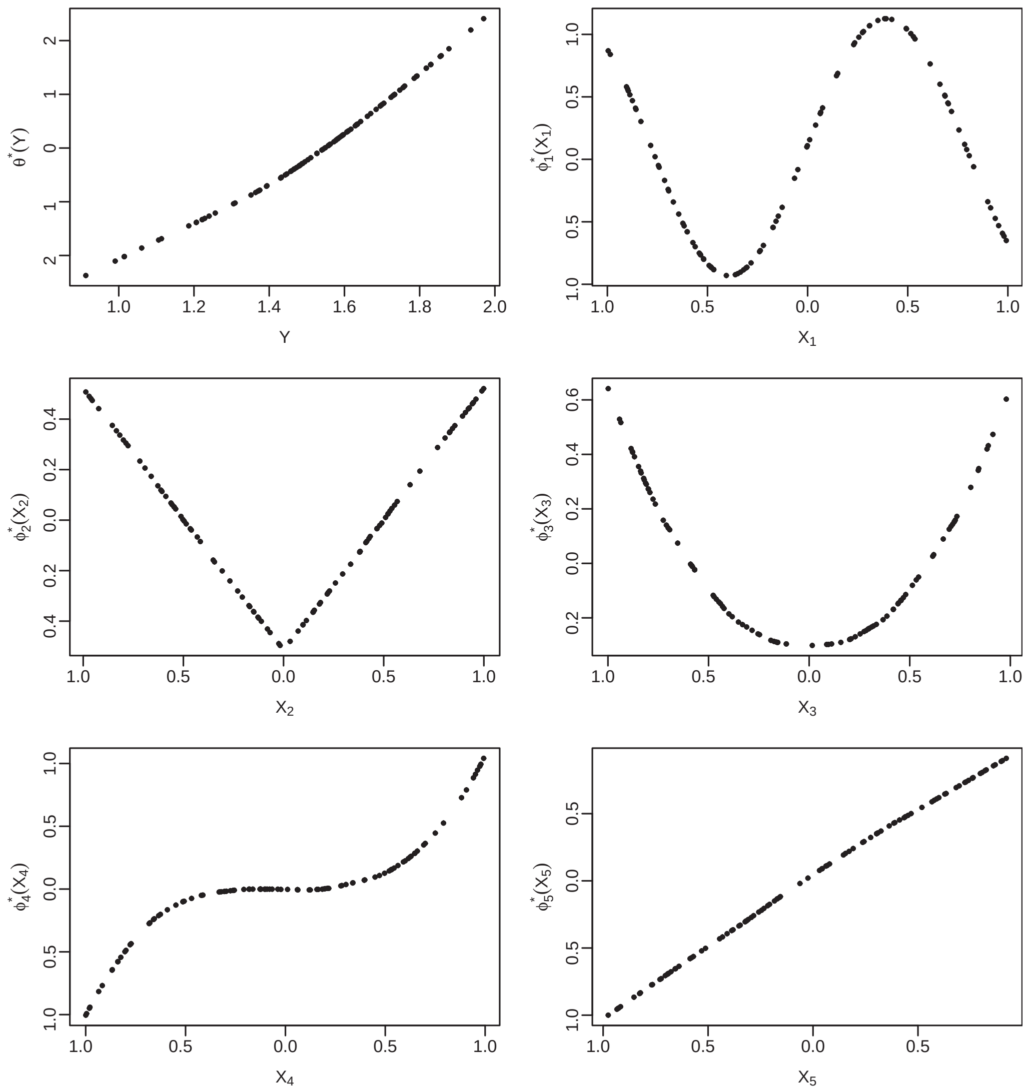

Figure 2: ACE optimal transformations of simulated dataset

plot Y individually against the predictors X_1, X_2, X_3, X_4 and X_5 so as to gain some insight into the pair-wise relationships, yielding the graphs in Figure 1. Note that the plots in Figure 1 do not reveal any obvious functional forms for either the dependent variable or predictors, even though X_1, X_2, X_3, X_4 and X_5 are statistically independent. Under such circumstances, direct application of linear regression is not appropriate.

If we regress Y on X_1, X_2, X_3, X_4 and X_5 using the Ordinary Least Squares (OLS) method, we obtain the estimated OLS equation that has an adjusted R^2

of .3708 and F-statistic of 12.67 on 5 and 94 degrees of freedom ($P < 0.0001$). The predictors X_1, X_4 and X_5 are statistically significant ($P < 0.0001$) but the probabilities of null tests of zero coefficients for X_2 and X_3 are 0.0527 and 0.5635, respectively. Rearranging (3.1) as:

$$\exp(Y) = 4 + \sin(4X_1) + |X_2| + X_3^2 + X_4^3 + X_5 + .1\epsilon$$

we know that the optimal transformations of dependent and independent variables will have the following forms up to a linear transformation:

$$\begin{aligned}
\theta^*(Y) &= \exp(Y) \\
\phi_1^*(X_1) &= \sin(3X_1) \\
\phi_2^*(X_2) &= |X_2| \\
\phi_3^*(X_3) &= X_3^2 \\
\phi_4^*(X_4) &= X_4^3 \\
\phi_5^*(X_5) &= X_5
\end{aligned} \tag{3.2}$$

To check if the ACE algorithm can recover these functions, we applied the algorithm to this simulated data set and the results are plotted in Figure 2. Clearly, ACE is able to recover the corresponding functions in (3.2).

A regression of the transformed dependent variable on all the transformed covariates results in all parameter coefficients of the independent variables being positive and close to 1:

$$\theta^*(Y) = 0.9989\phi_1^*(X_1) + .9959\phi_2^*(X_2) + .9999\phi_3^*(X_3) + 1.0006\phi_4^*(X_4) + 1.0000\phi_5^*(X_5)$$

which is very close estimate of

$$\theta^*(Y) = \phi_1^*(X_1) + \phi_2^*(X_2) + \phi_3^*(X_3) + \phi_4^*(X_4) + \phi_5^*(X_5)$$

indicating that the optimal parametric transformations have achieved. The ACE transformed variables has an adjusted R^2 of 0.9904, considerably better than the value of 0.3708 obtained using OLS. Note that in theory ACE cannot produce a worse fit than ordinary regression, because if no transformations are found to be necessary (i.e., the ordinary regression model is appropriate), then ACE would simply suggest nearly linear transformations for all the variables.

As with similar methods, ACE will generally not perform as well with empirical data as the simulated example here for reasons which include: (1) the dependent variable will usually have a lower association with independent variables; (2) some predictors are likely to be highly correlated; (3) sizeable error terms tend to exist; (4) there are some unobserved predictors which have been omitted; and (5) some superfluous variables may be included in the regression

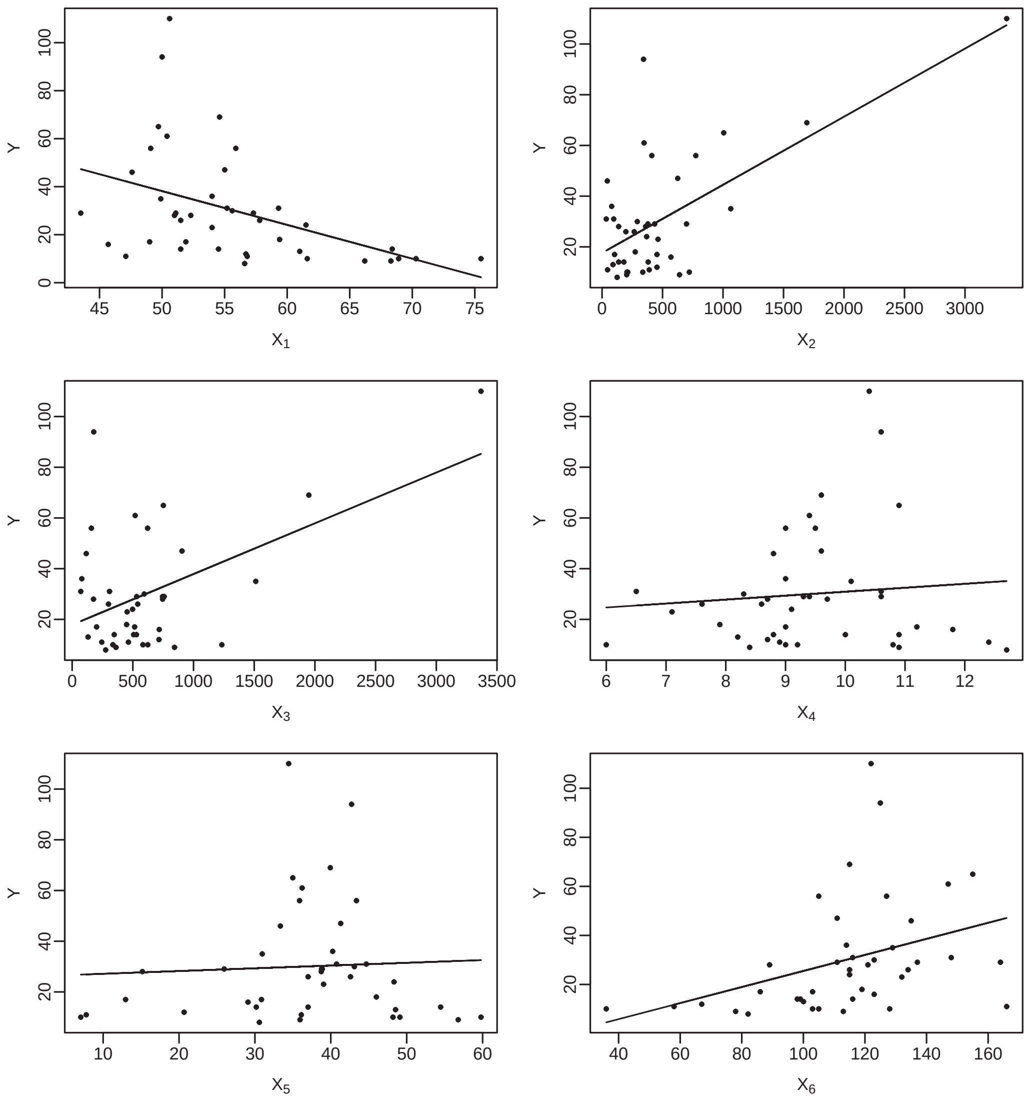

Figure 3: Scatterplots of US air pollution data

model (although this situation can be partially taken care of by using a stepwise variable selection procedure, such procedures should be used with caution). For this simulated dataset, a backward stepwise procedure retains all five independent variables when using ACE, but discards X_2 and X_3 when using the OLS method.

4. Example: US Air Pollution Data

In this section, we apply *ACE* to air pollution data for 41 US cities, collected by Sokal and Rohlf (1981) from several US government publications (the data

Table 1: Comparison of R^2 values for the US air pollution data

Dependent Variable	Independent Variable	Correlation R^2
Y	X_1	0.1880
Y	X_2	0.4157
Y	X_3	0.2438
Y	X_4	0.0090
Y	X_5	0.0029
Y	X_6	0.1366
Y	$X_1, X_2, X_3, X_4, X_5, X_6$	0.6695
$\theta^*(Y)$	$\sum_{i=1}^{6} \phi_i^*(X_i)$	0.9428

are also available in Handetal (1994)). There is a single dependent variable, Y, the annual mean concentration of sulphur dioxide, in micrograms per cubic metre. The data generally relate to means for the three years 1969-71. The values of six explanatory variables are also recorded, two of which relate to human ecology, and four to climate as follows:

X_1: Average annual temperature in degrees Farenheit
X_2: Number of manufacturing enterprises employing 20 or more workers
X_3: Population size (1970 census) in thousands
X_4: Average annual wind speed in miles per hour
X_5: Average annual precipitation in inches
X_6: Average number of days with precipitation per year

The main question of interest is how the pollution level, as measured by sulphur dioxide concentration, is determined by these six explanatory variables using multiple regression.

Figure 3 shows scatterplots of Y against X_1,X_2,X_3,X_4,X_5,and X_6. A linear regression of Y on the individual variables yields a maximum multiple correlation R^2=0.4157 with X_2 (number of manufacturing enterprises) and a minimum correlation R^2=0.0029 with X_5 (average annual precipitation), see Table 1. The correlation coefficient is statistically significant at the 5% level for only four of six variables, X_1, X_2, X_3, and X_6.
We first present the conventional linear regression model of sulphur dioxide concentration on the six predictor variables:

$$\begin{aligned} Y &= 111.7285 - 1.2679X_1 + 0.0649X_2 - 0.0393X_3 \\ &\qquad - 3.1814X_4 + 0.5124X_5 - 0.0521X_6 \end{aligned} \tag{4.1}$$

with $R^2 = 0.6695$. Statistical P values for zero coefficient test are 0.0491, 0.0002, 0.0138, 0.0887, 0.1669 and 0.7500 for X_1, X_2, X_3, X_4, X_5, and X_6, respectively.

We then applied the ACE agorithm to the data. The optimal transformations for Y and the six independent variables are shown in Figure 4. Regression of the transformed response on the transformed independent variables yields the following estimated equation:

$$\begin{aligned}\theta^*(Y) = {} & 1.0123\phi_1^*(X_1) + 1.0089\phi_2^*(X_2) + 1.0479\phi_3^*(X_3) \\ & +1.1610\phi_4^*(X_4) + 1.0630\phi_5^*(X_5) + 1.1907\phi_6^*(X_6)\end{aligned} \tag{4.2}$$

with $R^2 = 0.9469$ and statistical P values for zero terms ($H_0 : \phi_i^*(X_i) = 0$) are all less than 0.0001.

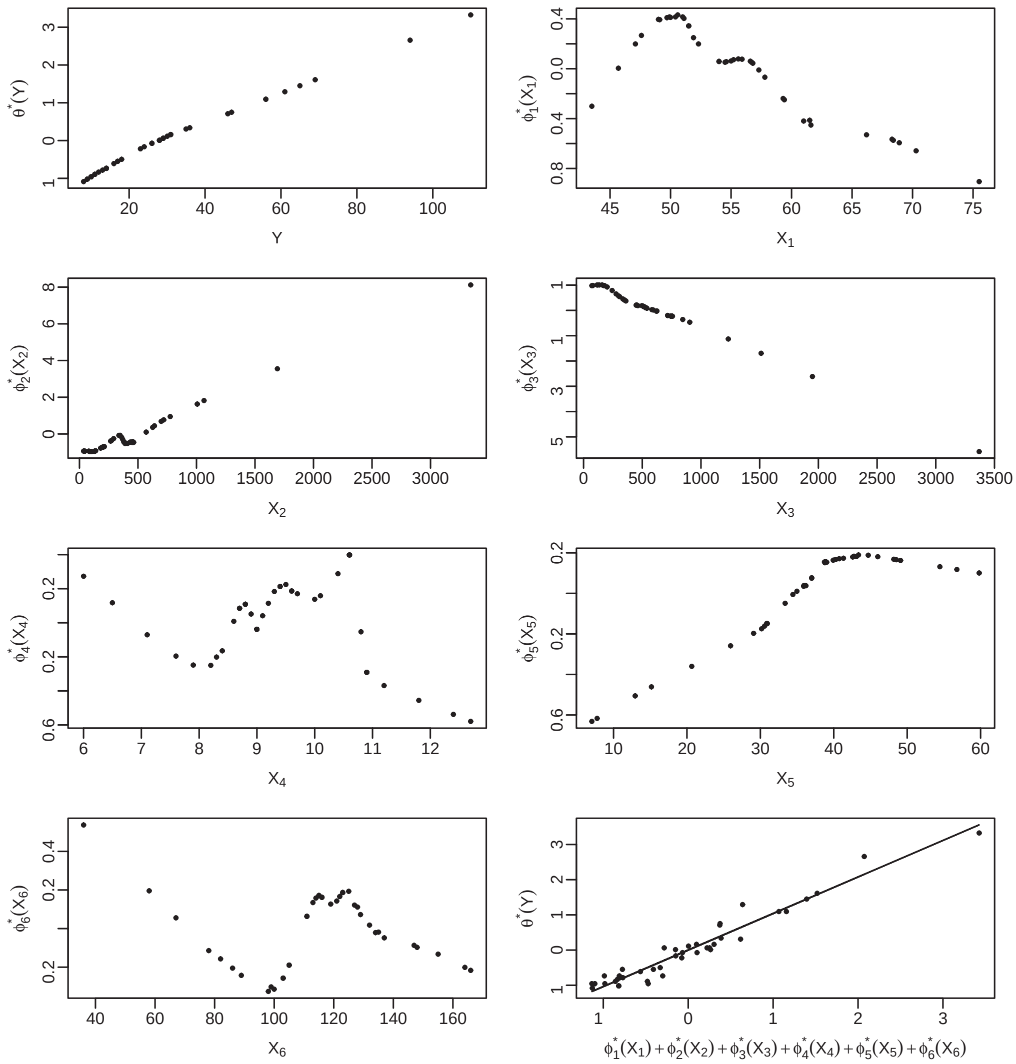

Figure 4: ACE optimal transformations of US air pollution data

We draw the following conclusions from the results above. First, the estimated regression model of (4.2) from the ACE transformed variables has an R^2 value of 0.9469 which is close to 1 and much better than 0.6695 for the estimated model of (4.1) obtained using the original untransformed variables. The ACE transformation is optimal and outperforms conventional regression by a large margin in this case.

Second, all the transformed independent variables ($\phi_1^*(X_1)$, $\phi_2^*(X_2)$,..., $\phi_6^*(X_6)$) are statistically significantly related to the transformed dependent variable $\theta^*(Y)$ at $P < 0.0001$. This is in contrast with the finding from the estimated linear regression model in (4.1) that only X_1, X_2 and X_3 are statistically significant at the 5% level.

Third, of the six independent variables, only X_2 (number of manufacturing enterprises) and X_3 (population size) have a linear effect on sulphur dioxide concentration (after controlling for the other predictors, population size has a net negative effect on the sulphur dioxide concentration. This is largely due to the high correlation between X_2 and X_3 (r=0.9553), as they are both indicators of size of the city.). The irregular patterns of the estimated transformations for the other four independent variables indicate complicated relationships between the concentration of sulphur dioxide and its predictors. One of major problems with conventional linear regression is its inability to detect and uncover such relationships. For example, from model (4.1) we would conclude that X_4 is not a significant predictor of sulphur dioxide concentration, but ACE analysis shows that it is significant in non-linear form. Treating such a non-linear relationship as linear one will distort the true relationship between a predictor and response.

Having implemented the ACE procedure and plotted the ACE results in Figure 4, we may attempt to find the closed functional forms for these six determinants of sulphur dioxide concentration. There are often a number of potential candidates for transformation of a variable suggested by the ACE plot that fit the data well according to statistics such as proportion of variance explained, or goodness-of-fit Chi-square. There are a number of statistical selection criteria for choosing of the best transformation (Akaike, 1973; Raftery, 1995), of which the Bayesian Information Criterion (*BIC*) has been widely used (Raftery, 1995). In a linear regression model, *BIC* is often approximated by:

$$BIC = N \log(1 - R^2) + p \log N$$

where R^2 is the square of the multiple correlation coefficient, p is the number of independent variables in the model of interest (not including the intercept) and N is the sample size. The smaller the *BIC*, the better the model.

As seen in Figure 4 and mentioned earlier, $\theta^*(Y)$, $\phi_2^*(X_2)$ and $\phi_3^*(X_3)$ seem to be linear, so no transformation is needed for these three variables. The remaining four transformed variables tend to be linear in different regions. So we use linear splines to approximate and parameterise those four transformations. We calculated the *BIC* statistic for various change points so that we could determine the best cut points (nodes) for the proposed linear spline functions. Table 2 displays the parameterisation of transformed variables and their coefficient estimates from regression analysis of the parameterised variables.

Table 2: Parameter estimates from regression analysis of parameterised variables for the US air pollution data

i	$\hat{\phi}_i^*(X_i)$	β_{ij}	$S.E.(\beta_{ij})$	T	P
	Intercept	-214.9484	156.7822	-1.3710	0.1813
1	$\beta_{11}\min(X_1, 48)$	6.9969	2.8786	2.4310	0.0217
	$+\beta_{12}\max(X_2, 48)$	-1.1340	0.9859	-1.1500	0.2598
2	$\beta_2 X_2$	0.0620	0.0129	4.7891	0.0001
3	$\beta_3 X_3$	-0.0439	0.0127	-3.4612	0.0017
4	$\beta_{41}\min(X_4, 8)$	-1.3794	6.9971	-0.1973	0.8451
	$+g_1(X_4)$	6.1285	2.9690	2.0644	0.0484
	$+\beta_{43}\max(0, X_4 - 10.6)$	-14.8222	5.0014	-2.9641	0.0061
5	$\beta_{51}\min(X_5, 42)$	0.5720	0.5373	1.0646	0.2962
	$+\beta_{52}\max(X_5, 42)$	-0.0889	1.2106	-0.0734	0.9420
6	$\beta_{61}\min(X_6, 100)$	-0.5184	0.4223	-1.2275	0.2299
	$+\beta_{62}g_2(X_6)$	0.8466	0.3326	2.5457	0.0167
	$+\beta_{63}\max(0, X_6 - 125)$	-0.3716	0.2235	-1.6635	0.1075

Note: $g_1(X_4) = \beta_{42}\max(0, \min(X_4 - 8, 10.6 - 8))$, $g_2(X_6) = \max(0, \min(X_6 - 100, 125 - 100))$

There are a number of observations from Table 2. First, the general patterns for the estimated effects of independent variables are similar to those observed in Figure 4. Second, the estimated model has an R^2 of 0.8324, much larger than that of 0.6695 for the OLS estimated model in (4.1) but still far from that of 0.9469 for the ACE estimated model in (4.2), indicating the parameterisation of the transformed variables in Table 2 is still not fully satisfactory. Third, $\hat{\phi}_4^*(X_4)$ and $\hat{\phi}_6^*(X_6)$ become statistically significant determinants of Y. In fact, the P values for joint F test for zero coefficients for their linear spline components ($H_0 : \beta_{41} = \beta_{42} = 0$ for X_4 and $H_0 : \beta_{51} = \beta_{52} = \beta_{53} = 0$ for X_6) are 0.0363 and 0.0249, respectively. The results contrast those from the OLS results in (4.1), which found that X_4 and X_5 were not significant. Finally, regarding the effect of X_1, the OLS analysis shows a marginally significant negative effect for X_1 (P=0.0491), whereas Table 2 shows that $\hat{\phi}_1^*(X_1)$ consists of two linear splines, $6.9969\min(X_1, 48) - 1.1340\max(X_1, 48)$, the former term being significant at the

5% level, but the latter not significant (suggesting a saturation effect). $\hat{\phi}_1^*(X_1)$ provides a slightly better fit of the data than the simple linear X_1, as the P values of the F test of their effects are 0.0310 and 0.0491, respectively.

5. Limitations of the ACE Algorithm Approach

We have demonstrated that the ACE algorithm provides a largely automatic method for estimating optimal transformations of response and predictor variables in multiple regression. The transformations generated by ACE can facilitate the identification of appropriate, and possibly meaningful, functional forms. Even when the transformed variable cannot be well-approximated as in Figure 4, examination of the ACE transformation plots may lead to new insights into the relationships between response and predictor variables. Another feature of ACE is its ability to incorporate multiple and mixed variables, both continuous (such as temperature) and categorical (such as occupation) within a common framework. The ACE algorithm can also be easily extended to generalised linear models (Raftery and Richardson, 1996). While the ACE approach has many advantages as a data analysis tool, we now turn to some important issues related to its application.

5.1 Sensitivity to variable ordering

ACE results depend on the order in which the predictor variables are entered into the analysis (in practice, this is the order they are in the $\mathbf{X}$ matrix), see (2.2). We use the US air pollution data and, for illustrative purposes, we use only six of the possible 720 permutations of the six predictor variables $X_1, X_2, X_3, X_4, X_5, X_6$ by changing only the order of first three variables X_1, X_2, X_3. We then apply ACE to them and present the R^2 values in Table 3, showing that the R^2 values are sensitive to the order of predictor variables, ranging from 0.9000 for the order $X_3, X_2, X_1, X_4, X_5, X_6$, to 0.9525 with X_1 and X_2 reversed.

Table 3: Sensitivity test of R^2 values to ordering and outlier in ACE transformation for the US air pollution data

Order	All Data	Excluding Outlier
$X_1, X_2, X_3, X_4, X_5, X_6$	0.9469	0.9102
$X_1, X_3, X_2, X_4, X_5, X_6$	0.9431	0.9213
$X_2, X_1, X_3, X_4, X_5, X_6$	0.9272	0.9122
$X_2, X_3, X_1, X_4, X_5, X_6$	0.9349	0.9292
$X_3, X_1, X_2, X_4, X_5, X_6$	0.9525	0.9320
$X_3, X_2, X_1, X_4, X_5, X_6$	0.9000	0.9082

This sensitivity might be due to the choice of smoother use by ACE (Hastie and Tibshirani, 1990). However, this property of *ACE* has important implications for its application. The authors of *ACE* suggested that several different orderings should be tried once a subset of the predictor variables has been selected (Breiman and Friedman, 1985). In the case of the US air pollution data, it seems that it is better to transform the variables in the order of $X_3, X_1, X_2, X_4, X_5, X_6$ rather than the order of $X_1, X_2, X_3, X_4, X_5, X_6$ as we used in the analysis (although this will not necessarily lead to similar magnitude differences in the results if functional transformations are made).

5.2 Robustness to outliers

Another inherent problem with the ACE procedure is that it may be highly sensitive to extreme outlying data values and highly influential points in both the response and predictor variables (Breiman and Friedman, 1985; Tibshirani, 1988). To illustrate this property, we use the US air pollution data again, in which Chicago has a very large number of manufacturing enterprises and an extreme value of sulphur dioxide concentration. Chicago is confirmed to be an outlier by the Cook statistic, and removed from the analysis. Application of ACE to this reduced dataset results in R^2 being 0.9102, compared with R^2=0.9469 when Chicago was included in the analysis. The above example demonstrates that even a single influential point can have some impact on the ACE results. The R^2 correlations for the sample without Chicago are also displayed in Table 3. The results show that R^2 values have been slightly reduced for all the permutations except for the order of $X_3, X_2, X_1, X_4, X_5, X_6$ compared with the results of Table 2, and that variation of the R^2 values are considerably less in the reduced than in the original data set.

In some extreme cases, outliers could significantly distort the estimated transformations. The authors of *ACE* warned that the algorithm should be used with a great deal of caution in the suspected presence of extreme outliers (Breiman and Friedman, 1985). It is advised that the algorithm should be used in the context of the ever-growing collection of modern tools for regression such robust procedures and diagnostic techniques for identifying influential points.

5.3 Normality and homoscedasticity assumptions

There are three basic assumptions in a linear regression analysis: linearity, homoscedasticity, and normality. Transformations in a regression analysis may be needed to overcome violation of one or more of these requirements, but not, of course, the problem of omitted variables. The goal of the *ACE* algorithm is to find the transformations that maximise the multiple linear correlation of the predictors with the response variable, i.e. the dependence of the response variable on the independent variables is maximised. This is only one of three goals of an optimal transformation for regression. Our examples suggests that when the transformation is optimal in terms of linearity, in practice, it also likely to be close to optimal in terms of normality and homoscedasticity. However, this is certainly not always the case. It should be emphasised here that there is no guarantee that the residuals in the ACE transformed model will be normally distributed with stable variance. Although ACE is a potent and versatile approach for maximising correlations, it suffers from some anomalies when viewed as a regression tool, especially in low-correlation settings. A modification of ACE designed primarily for such regression problems was proposed by Tibshirani (1988) and differs from the original ACE algorithm in that it chooses $\theta(Y)$ to achieve a special asymptotic variance stabilising feature. The goal here is to estimate transformations θ and ϕ_i which have the following properties:

$$E\{\theta^*(Y)|X_1, X_2, \ldots, X_p\} = \sum_{i=1}^{p} \phi_i(X_i)$$

$$Var(\theta(Y)|\{\sum_{i=1}^{p} \phi_i(X_i)\}) = \text{constant}$$

The transformation θ is assumed to be strictly monotone (and thus invertible) and the conditional expectations are approximated using scatterplot supersmoother smoothing (Friedman and Stuetzle, 1982). The resulting algorithm is called additivity and variance stabilisation (AVAS). In this paper, we discuss and apply only the ACE algorithm.

6. Summary

We have introduced the ACE algorithm, a non-parametric automatic transformation method that produces the maximum multiple correlation of a response and a set of predictor variables. The approach solves the general problem of establishing the linearity assumption required in regression analysis, so that the relationship between response and independent variables can be best described and existence of non-linear relationship can be explored and uncovered. An examination of these results can give the data analyst insight into the relationships between these variables, and suggest if transformations are required.

We have described the implementation of the ACE algorithm and the interpretation of its output. The ACE plot is very useful for understanding complicated relationships and it is an indispensable tool for effective use of the ACE results. It provides a straightforward method for identifying functional relationships between dependent and independent variables. There will often be a number of potential candidates for transformation of a variable suggested by the ACE plot that fit the data well according to R^2. To select the best transformation, we have introduced the *BIC* statistical selection criterion. With the ACE plot and *BIC*, we can derive the most appropriate functional relationships between the response variable and its predictors.

Using a simulated dataset and an actual dataset, we have demonstrated the usefulness of *ACE* guided transformation in multivariate analysis. The power of the *ACE* approach lies in its ability to recover the functional forms of variables and to uncover complicated relationships. In addition, the *ACE* guided parameterisation of variables will often improve the model fit considerably compared with the conventional linear model.

Although *ACE* provides a largely automated approach to estimating optimal transformations, it does not mean that the *ACE* results should be trusted blindly and used dogmatically, additional information and experience of the data analyst remain important. It should be emphasised that the success of the *ACE* algorithm, like other modern statistical methods, relies on the quality of the data and underlying association between the response and independent variables.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, (Edited by B.N. Petrov and F. Csaki). Akademiai Kiado, 267-281.

Box, G. E. P. and Cox, D. R. (1964). Analysis of transformation. *Journal of Royal Statistical Society*, B **26**, 211-252.

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

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford University Press.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80**, 580-78.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall.

Cook, R. P. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall.

Durrleman, S. and Simon, R. (1989). Flexible regression-models with cubic-splines. *Statistics in Medicine* **8**, 551-561.

Friedman, J. H.and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* **76**, 817-823.

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

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman and Hall.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. K. and Ostrowski, E. (1994). *A Handbook of Small Data Sets* Chapman and Hall.

Härdle, W. (1992). *Applied Nonparametric Regrerssion* Cambridge University Press.

Hastie, T and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.

Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotonic transformations of the data. *Journal of Royal Statistical Society*, B **27**, 251-263.

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

Raftery, A. E. (1995). Bayesian model selection in social research (with Discussion). *Sociological Methodology* **25**, 111-195.

Raftery, A. E. and Richardson, S. R. (1996). Model selection for generalised linear models via GLIB, with application to epidemiology. In *Bayesian Biostatistics* (Edited by D. A. Berry and D. K. Strangl), Marcel Dekker. 321-354.

Royston, P. (2000). A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. *Statistics in Medicine* **19**, 1831-1847.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics* **43**, 429-467.

Sokal, R. R. and Rohlf, F. J. (1981). *Biometry : The Principles and Practice of Statistics in Biological Research* W.H. Freeman.

Tibshirani, R. (1988). Estimating optimal transformations for regression via additivity and variance stabilization. *Journal of American Statistical Association*, **83**, 394-405.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S-PLUS*. Springer-Verlag.

Received March 4, 2003; accepted September 3, 2003.

Duolao Wang
Medical Statistics Unit
London School of Hygiene and Tropical Medicine
Keppel Street
London WC1E 7HT, UK
duolao.wang@lshtm.ac.uk

Michael Murphy
Population Studies

London School of Economics
Houghton Street
London WC2A 2AE, UK