Testing in Additive and Projection Pursuit Models

Arun Kumar Kuchibhotla*
Indian Statistical Institute, Kolkata, India - karun3kumar@gmail.com

Abstract

Additive models and projection pursuit models are very useful popular nonparametric methods for fitting multivariate data. The flexibility of these models makes them very useful. Yet, this very property can sometimes lead to overfitting. Inference procedures like testing of hypothesis in these cases are not very well developed in the literature. This might be due to the complexity involved in estimation. In the present paper we introduce a bootstrap based technique which allows one to test the hypothesis of the adequacy of multiple linear regression model versus the nonparametric additive model and beyond. These tests are highly useful for practitioners since the simpler models are more interpretable. We will also introduce a new model which incorporates both the additive model and the multiple index model.

Keywords: Maximum Correlation Coefficient, Alternating Conditional Expectation, Projection Pursuit Regression, Bootstrap.

1 Introduction

In this paper we consider the additive model with a response variable Y and p predictors $X_1, X_2, \ldots, X_p$; the model is given by

$$g(Y) = \sum_{k=1}^{p} f_k(X_k) + \epsilon, \quad E[\epsilon|X] = 0 \quad a.e. \tag{1}$$

Here $g, f_1, \ldots, f_p$ are all unknown functions from $\mathbb{R}$ to $\mathbb{R}$. This will also be referred to as the Alternating Conditional Expectation (ACE) model. Projection pursuit model or multiple index model in this case is given by

$$Y = \sum_{k=1}^{M} f_k(\alpha_k^\top X) + \epsilon, \quad E[\epsilon|X] = 0 \quad a.e. \tag{2}$$

Here M is a natural number which is chosen based on some goodness of fit measure and α_k, $1 \leq k \leq M$ are all unit vectors in $\mathbb{R}^p$ with X representing $(X_1, \ldots, X_p)^\top$. This will also be referred to as the Projection Pursuit Regression (PPR) model.

The idea of the additive model can be understood better in the case $p = 1$. This can be dated back to Rényi (1959) who considered the problem of defining a measure of dependence between two random variables. The maximum correlation coefficient between random variables X and Y is defined as the supremum of the correlation coefficient between transformed X and transformed Y over all transformations, whenever the correlation is defined. That is,

$$S := \sup_{f,g} \mathrm{Corr}(f(X), g(Y)), \tag{3}$$

where supremum is taken over all functions f and g such that correlation is defined. It is clear from the definition that $S = 0$ if and only if the variables are independent. Also note that if there exist such transformations g_0 for Y and f_0 for X such that $S = \mathrm{Corr}(g_0(Y), f_0(X))$, then the regression of transformed Y on transformed X would give the best linear fit. This idea of estimating the optimal transformations gives rise to the additive model. For $p > 1$, an obvious extension considers the correlation between $g(Y)$ and the sum of $f_k(X_k)$ for $1 \leq k \leq p$.

The idea of multiple index model or, as it was first proposed, the projection pursuit model, stems from the fact that the additive model cannot represent the interactions between the response variables. Consider, for example, the model $Y = X_1X_2 + \epsilon$. By fitting an additive model to a simulated data from this model, we get the log function as the optimal transformation for each of these variables. But this can only be approximate because the error involved here is not multiplicative but additive. Noting that $4xy = (x+y)^2 - (x-y)^2$, we view the multiple index model to be more suitable for these data. This is by no means a coincidence. Results of Diaconis and Shahshahani (1984) show that almost any multivariate function can be approximated as closely as needed by the sum of functions of linear combinations of the covariates.
The new model that we consider which incorporates both the additive model and the multiple index model is given by

$$g(Y) = \sum_{i=1}^{M} f_k(\alpha_k^\top X) + \epsilon, \quad E[\epsilon|X] = 0 \quad a.e. \tag{4}$$

Here M is a natural number chosen on the basis of some goodness of fit measure and α_k, $1 \le k \le M$ are all unit vectors in $\mathbb{R}^p$ with X representing $(X_1, X_2, \ldots, X_p)^\top$. This new model will be referred to as GACE (Generalized ACE) model. What do we need a new model for? There are some distributions where $E[Y|X]$ cannot be approximated by the sum of functions of linear combinations of the covariates. In these cases, there may be a function g such that $E[g(Y)|X]$ may be approximated in that way. Even if the conditional expectation $E[Y|X]$ can be approximated, we might require a very large value of M. In such cases, applying a transformation to Y might lead to a simpler model with better interpretation.
The rest of the paper is organized as follows. In Section 2, we describe the alternating conditional expectation algorithm to estimate the unknown functions in additive model. In Section 3, we describe a set up for a simulated dataset and describe a real dataset. In Section 4, we describe tests of linearity for each variable for a general model. In Section 5, we describe the GACE model in detail and consider an estimation algorithm and a new testing procedure.

2 The Alternating Conditional Expectation Algorithm

Before describing the algorithm, we present a few definitions and results from Breiman and Friedman (1985) which will motivate the algorithm. First consider the case $p = 1$.

Definition: We say transformations f_0 and g_0 for X and Y are *optimal for correlation* if they satisfy,

$$S := \mathrm{Corr}(f_0(X), g_0(Y)) = \sup_{f,g} \mathrm{Corr}(f(X), g(Y))$$

Observe that S is always non-negative. Without loss of generality, one can assume that supremum is taken over all functions, whose expectations are zero and variances one. That is, $E(f(X)) = E(g(Y)) = 0$ and $\mathrm{Var}(f(X))=\mathrm{Var}(g(Y))=1$. Hence

$$\begin{aligned} S &:= E[f_0(X)g_0(Y)] \\ &= \sup_{f,g}\{E[f(X)g(Y)] : E[f(X)] = 0 = E[g(Y)] \text{ and } E[f^2(X)] = 1 = E[g^2(Y)]\} \end{aligned}$$

Definition: We say transformations α and β for X and Y are *optimal for regression* if they satisfy, $E[\beta^2(Y)] = 1$ and

$$e^2 := E[\alpha(X) - \beta(Y)]^2 = \inf_{f,g} E[f(X) - g(Y)]^2$$

Again we assume that infimum is taken over all functions whose expectation is zero. Hence,

$$\begin{aligned} e^2 &:= E[\alpha(X) - \beta(Y)]^2 \\ &= \inf_{f,g}\{E[f(X) - g(Y)]^2 : E[f(X)] = 0 = E[g(Y)] \text{ and } \mathrm{Var}(g(Y)) = 1\}. \end{aligned}$$

Using Cauchy-Schwarz inequality it is easy to prove that

$$E[E[f_0(X)|Y]|X] = S^2 f_0(X), \tag{5}$$

$$E[E[g_0(Y)|X]|Y] = S^2 g_0(Y). \tag{6}$$

The following result gives the relation between transformations that are optimal for regression and optimal for correlation.

Theorem 1 (Breiman and Friedman (1985), Theorem 5.1)**.** *The pair (f_0, g_0) are optimal for correlation if and only if $\alpha = Sf_0, \beta = g_0$ are optimal for regression. Furthermore, $e^2 = 1 - S^2$.*

Using this result, we find that the following relations will be satisfied by transformations optimal for regression,

$$\beta(Y) = \frac{E[\alpha(X)|Y]}{\sqrt{\mathrm{Var}(E[\alpha(X)|Y])}} \tag{7}$$

$$\alpha(X) = E[\beta(Y)|X] \tag{8}$$

Remark: It is easy to see that finding maximum correlation coefficient, in general, is hard. Hence for finding optimal transformations, it is better to use Equations (7) and (8) instead of Equations (5) and (6) because in the former we have three 'parameters' to estimate, one less than the latter. Henceforth, we discuss about estimating optimal transformations optimal for regression instead of for correlation.

For p covariates $X_1, X_2, \ldots, X_p$, one can define the optimal transformation as follows:

Definition: Let $g(Y), f_1(X_1), f_2(X_2), \ldots . f_p(X_p)$ be arbitrary mean zero transformations of corresponding variables. Also, let variance of $g(Y)$ be 1. Then the optimal transformations for regression are defined by those functions which minimize the fraction of unexplained variance. That is

$$e^2 := E[\beta(Y) - \sum_{i=1}^{p} \alpha_i(X_i)]^2 = \min_{g, f_1, \ldots . f_p} E[g(Y) - \sum_{i=1}^{p} f_i(X_i)]^2. \tag{9}$$

It is easy to see that the following relations are satisfied by the optimal transformations.

$$\beta(Y) = \frac{E\left[\sum_{i=1}^{p} \alpha_i(X_i)|Y\right]}{\sqrt{\mathrm{Var}\left(E\left[\sum_{i=1}^{p} \alpha_i(X_i)|Y\right]\right)}} \tag{10}$$

$$\alpha_i(X_i) = E[\beta(Y) - \sum_{j \neq i} \alpha_j(X_j)|X_i] \tag{11}$$

Equations (7), (8), (10) and (11) form the basis of the alternating conditional expectation algorithm given in Breiman and Friedman (1985).
The basic idea of the algorithm is to alternatively condition on variables until convergence is attained and hence the name alternating conditional expectation. It proceeds as follows:

1. Initialize. Set $\beta(Y) = (Y - E[Y])/\sqrt{\mathrm{Var}(Y)}$ (usual notation) and set all $\alpha_j(X_j) = 0$ (or set $\alpha_j(X_j) = E[Y|X_j]$).
2. Backfit. Using (11), find new functions $\alpha_j(X_j)$, for $j = 1, 2, \ldots, p$.
3. Compute. Use these new functions to compute new $\beta(Y)$, using (10) (standardizing by scaling with the standard deviation avoids getting a trivial solution).
4. Alternate. Perform steps 2 and 3 until e^2 converges to a minimum.

The algorithm can be used even when the predictor variables $X_1, X_2, \ldots, X_p$ are of mixed types, for example, they can be continuous, categorical, periodic. In that case the conditional expectations will be accordingly changed.

Remark: Breiman and Friedman (1985) suggest that if a particular estimate of transformation suggests a familiar functional form for the transformation, then the data can be pretransformed using this functional form and the ACE can be rerun. The ACE algorithm can be made semi-parametric also by introducing certain functional forms into algorithm. For example, if we want a variable, say X_1, to have a linear transformation, we introduce the assumption that

$$E[\beta(Y) - \sum_{j \neq 1} \alpha_j(X_j)|X_1] = a + bX_1$$

and in each iteration compute optimal a, b using simple linear regression. But if we introduce parametric (functional) transformations to many variables then a certain care has to be taken regarding the number of parameters estimated and the degrees of freedom. In case of samples where the conditional expectations cannot be calculated, we use smoothers as estimates of the conditional expectation in the ACE algorithm.

In case of a categorical variable, say Z, the estimate of conditional expectation $E[Z|X_1 = x]$, is given by, $\sum_{x_k=x} z_k / \sum_{x_k=x} 1$. We use the supersmoother (for real data examples and simulated studies) in alternating conditional expectation but any smoother satisfying certain regularity conditions can be used. The supersmoother proposed by Friedman and Stuetzle (1982) is based on local linear k-NN (span k nearest neighbour) fits in a variable neighbourhood of the estimation point x. 'Local cross-validation' is applied to estimate the optimal span as a function of the predictor variable. We refer to Breiman and Friedman (1985) and Friedman (1984) for more details.
Prediction in case of additive models is fairly straightforward. One simple way to get a prediction is to estimate the function part of the right hand side of Equation (1) and then invert $\hat{g}$. This gives us the predicted value of y. But if the function $\hat{g}$ is not invertible this method does not work well. There is a variant of ACE called predictive ACE proposed by Owen and Friedman for making better prediction in ACE model. Since multiple linear regression model, ACE, PPR and GACE models are all nested, comparisons between the models based on residual sum of squares or the coefficient of determination (R^2) are not very useful. We use a cross-validation based measure Q for comparison of the models. We define

$$Q := 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

as a measure of predictive ability of a model, where $\hat{y}_i$ is obtained by fitting the model on the remaining $n-1$ observations.

3 Numerical Study

We now present a simulated model and a real dataset on which we will apply the testing procedure to be discussed in the next section.

3.1 An Example of Simulated Data

We consider 6 random variables $Y, X_1, X_2, X_3, X_4, X_5$ related by the model,

$$Y = \log(4 + |X_1| + X_2^2 + \sin(2\pi X_3) + X_4 + X_5^3 + \epsilon), \tag{12}$$

with X_1, X_2, X_3, X_4, X_5 all generated independently from $U(-1, 1)$ and $\epsilon \sim N(0, 1)$. When ACE is applied on this simulated data, we get transformations as shown in the following figure.
Linear regression on this simulated data gave an adjusted R-Square 0.4309. Applying ACE on this data gave R-Squared of 0.989244 which is expected. Figure (3.1) shows that ACE estimates are very close to the actual ones. Note that the model is not the same as,

$$Y = \log(4 + |X_1| + X_2^2 + \sin(2\pi X_3) + X_4 + X_5^3) + \epsilon$$

which explains why minimizing $E[g(Y) - \sum_{i=1}^{p} f_i(X_i)]^2$ is better than minimizing $E[Y - g(\sum_{i=1}^{p} f_i(X_i))]^2$.

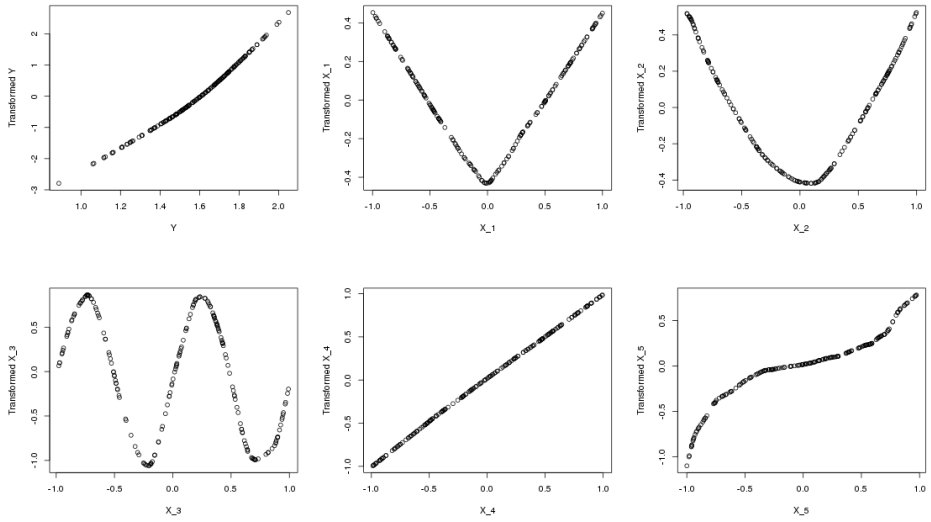

Figure 1: Model is $Y = \log(4 + |X_1| + X_2^2 + \sin(2\pi X_3) + X_4 + X_5^3 + \epsilon)$

3.2 A Real Data Example

The first real data that we consider is about the European rabbit Oryctolagus Cuniculus which is a major pest in Australia. A reliable method of age determination for rabbits caught in the wild would be of importance in ecological studies. We got a dataset with 71 data points; see Dudzinski and Mykytowycz (1961) for more details. The scatter plot of eye lens weight on the age of the rabbit is as shown in Figure (3.2). Clearly, age and eye lens weight are not (just) linearly related. There is a clear functional relation between age and eye lens weight. Applying ACE on these data gives the transformations as shown in Figure 3.2. The scatter plot between transformed variables is almost linear. In this case multiple linear regression gave a R-squared value of 0.7605. But the application of ACE leads to an R-squared value of 0.9875936. Clearly, the variable age seems to have the transformation from the log family.

4 Tests of Linearity

Considering the simplicity of linear regression it is necessary to test whether transforming the variables is at all necessary. Statistically speaking, we would like to test whether the optimal transformation for a variable, say X_1, is linear. Since we have no distributional assumptions in the model, we choose to develop a test by the use of bootstrapping. We briefly describe the bootstrap methodology for testing of hypothesis.
For any proposed hypothesis, let the rejection region be specified as $T > \tau$, for some τ based on the level of the test α. In case of analytical difficulty, one can use the bootstrap to estimate the distribution of T, and thus the value of τ, to perform the test. The procedure is as follows.

1. Compute the test statistics T and also the unknowns in the model satisfying the null hypothesis.
2. Take B subsamples each of size R with replacement and then calculate the value of T as calculated in step 1.
3. Now we have B values from the distribution of T under the null. For large enough B, we can find the estimate of critical value, $\hat{\tau}$, based on the empirical distribution of T. Calculation of the p-value can also be done using the empirical distribution, $P(T > T_{obs.})$.

The hypothesis that we introduced at the beginning of section can be stated as,

$$H_0 : \alpha_1(X_1) = a_1 + b_1 X_1 \text{ for some } a_1, b_1, \quad \text{versus} \quad H_1 : \alpha_1(X_1) \text{ is not linear.}$$

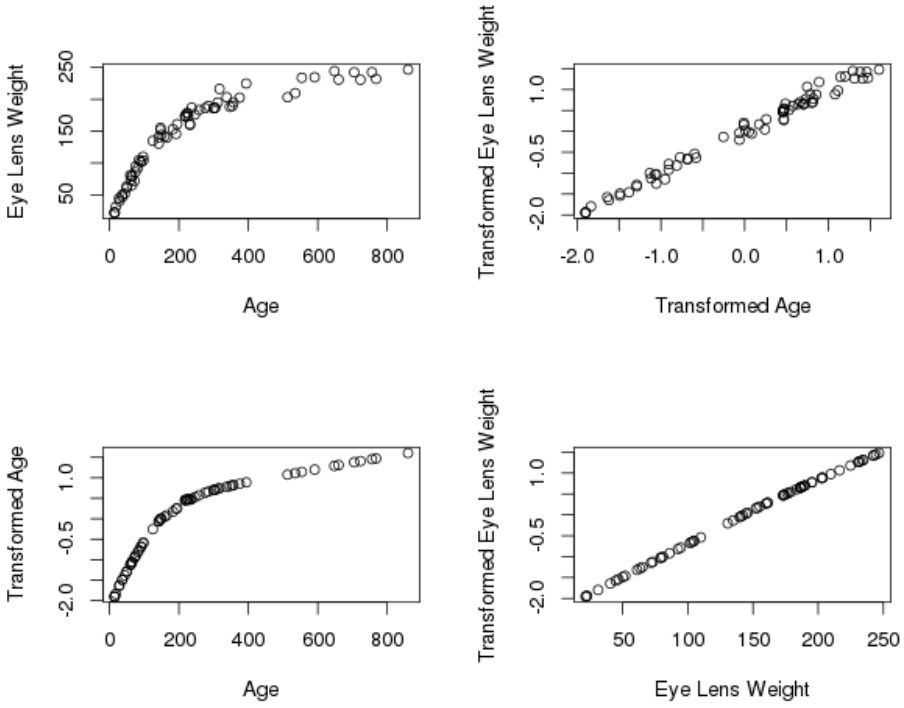

Figure 2: Eye Lens Weight versus Age of rabbit

This hypothesis can also be written in the form of models as,

$$H_0 : \beta(Y) = a_1 + b_1 X_1 + \sum_{i=2}^{p} \alpha_i(X_i) + \epsilon, \quad \text{versus} \quad H_1 : \beta(Y) = \sum_{i=1}^{p} \alpha_i(X_i) + \epsilon.$$

This can also be viewed as model selection. We can use any measure of goodness-of-fit which can discriminate between the null and the alternative. We propose to use either the ratio of the residual sum of squares or the ratio of the prediction sum of squares. The first measure is

$$T_1 = SSE_{H_0}/SSE_{H_1},$$

where $SSE_H = \sum_{j=1}^{p}(\hat{\beta}(Y_i) - \sum_{i=1}^{p} \hat{\alpha}_i(X_{ij}))^2$, hat functions computed under the hypothesis H. The prediction sum of squares based statistic is given by

$$T_2 = PSS_{H_0}/PSS_{H_1},$$

where PSS_H is the prediction sum of squares under that hypothesis given by, $PSS_H = \sum_{i=1}^{n}(Y_i - \hat{Y}_{iH})^2$. We generate subsamples from the data and then by fitting the model under both the hypothesis for each subsample, we estimate the distribution functions of T_1 and T_2. Using the level constraint, we estimate the cut-off's C_1 and C_2. Note here that when we take subsamples, they satisfy the hypothesis that the actual sample satisfies. Hence, when T_1 is calculated in this way for each subsample, we are estimating the distribution of T_1 under the true hypothesis which need not be the null hypothesis. Hence, the estimate we get from the bootstrap in this way may not converge to the actual cut-off value C_1.
To overcome this problem, we need to find a way to generate the value of the test statistics under the null independently of the hypothesis. For this, we follow the approach suggested by Davidson and MacKinnon (2004) and Martin (2007). This approach is as follows,

1. Fit the unrestricted full model and find the residuals, $e_1, \ldots, e_n$. Also fit the null model and find the estimates of parameters of line and estimates of other functions.
2. Construct resamples, $(y_1^*, \mathbf{x}_1), (y_2^*, \mathbf{x}_2), \ldots, (y_n^*, \mathbf{x}_n)$ under null hypothesis, where y_i^* is calculated from the model $\hat{\beta}(y_i) = \hat{a}_1 + \hat{b}_1 X_{1i} + \sum_{j=1}^{p} \hat{\alpha}_j(X_{ji}) + e_i^*$ for resampled $e_1^*, \ldots, e_n^*$ obtained from $e_1, \ldots, e_n$ using hat functions obtained from null model.
3. Find the value of the test statistic T_1 for each of these resamples by fitting the models under null and alternative for resampled observations.

Since these values are all obtained from the model under the null hypothesis, we get an estimate of distribution function of T_1 under H_0. Now, by estimating critical values and p-values, one can do a bootstrap test. Efron and Tibshirani (1986) have used this approach to find standard error of estimates of the optimal transformations in ACE. The method described above is not specific to the ACE model, This can be applied for comparing any two models. In particular, this method can be applied to the PPR model and also the GACE model. See Section 5.1 for a different approach.
Results for the test statistic T_1 applied to simulated data with sample size 300 is given in Table 1. Here the linearity test for the variable X_4 got accepted while all others got rejected. In simulations, this new simulation method has shown significant improvement compared to the previous one. The power can also be calculated by taking different alternatives in the same manner. It should be noted that this process involves

Table 1: Tests of linearity for Simulated Data

Variable	Statistic	Est. 5% Cut-Off	Est. p-value
Y	2.738687	1.032586	< 0.0001
1	8.688747	1.026901	< 0.0001
2	9.720563	1.027773	< 0.0001
3	37.07122	1.022618	< 0.0001
4	0.986757	1.015667	0.597
5	2.626931	1.022109	< 0.0001

obtaining y_i^* which can go completely wrong if transformation for Y is not invertible. Hence this test works properly only for invertible transformations of Y.
When applied to the real dataset, this test gave the following results (Table 2), which are consistent with the scatterplot.

Table 2: Tests of linearity for Real Data

Variable	Statistic	Est. 5% Cut-Off	Est. p-value
Y	1.114493	1.180262	0.129
1	4.531953	1.276061	< 0.0001

Remark: Similarly one can test whether optimal transformation for a variable belongs to some parametric family of functions. Also, one can extend the same methodology to test combined linearity of optimal transformations for different variables, like,

$$H_0 : \alpha_1(X_1) = a_1 + b_1 X_1,\ \alpha_2(X_2) = a_2 + b_2 X_2 \quad \text{versus} \quad H_1 : \text{Any of them is non-linear.}$$

This test needs to be done sequentially for interpreting it correctly. It is yet to be seen if the order in which these sequential tests are done matter. Using this, one can test if at all any transformations are needed for any of the variables. This test is the same as that of testing whether the ACE model is any better than the multiple linear regression for a data set. The same test (test criterion) can also be used to test for variable significance in the model in case of multiple covariates. Hence this testing procedure can be used to test for variable selection.

5 The GACE model

In full generality, one can consider fitting the model,

$$\beta(Y) = \alpha(X_1, X_2, \ldots, X_p) + \epsilon, \quad E[\epsilon|X] = 0. \tag{13}$$

Here we want α, β which minimize unexplained variance given by,

$$\frac{E[g(Y) - f(X_1, X_2, \ldots, X_p)]^2}{E[g^2(Y)]}.$$

With an additional constraint $E[g^2(Y)] = 1$, the transformations should satisfy,

$$E[\beta(Y) - \alpha(X_1, X_2, \ldots, X_p)]^2 = \min_{g,f} E[g(Y) - f(X_1, X_2, \ldots, X_p)]^2. \tag{14}$$

It is easy to prove that α, β satisfy,

$$\alpha(X_1, X_2, \ldots, X_p) = E[\beta(Y)|X_1, X_2, \ldots, X_p], \tag{15}$$

$$\beta(Y) = E[\alpha(X_1, X_2, \ldots, X_p)|Y]/\sqrt{\mathrm{Var}(E[\alpha(X_1, X_2, \ldots, X_p)|Y])}. \tag{16}$$

We follow the ACE algorithm and give an iterative algorithm for obtaining estimates of α, β. The algorithm can be written as follows.

1. Set $\beta(Y) = \frac{Y - E[Y]}{\sqrt{\mathrm{Var}(Y)}}$ and $\alpha(X_1, X_2, \ldots, X_p)$ as multiple linear regression line obtained by regressing Y on $X_1, X_2, \ldots, X_p$.
2. Use Equation (15) to get an estimate of $\alpha(X_1, X_2, \ldots, X_p)$.
3. Use Equation (16) to get an estimate of $\beta(Y)$.
4. Repeat steps 2 and 3 until minimization of

$$e^2 = E[\beta(Y) - \alpha(X_1, X_2, \ldots, X_p)]^2$$

is attained.

As was discussed in Section 2, we need to replace all the conditional expectations in the algorithm by estimates obtained from the data. These expectations are not functions of one variable. We need to estimate $E[Y|X_1, \ldots, X_p]$, which is a function from $\mathbb{R}^p$ to $\mathbb{R}$. The most extensively studied nonparametric estimates (kernel estimate, k-nearest neighbor, and spline smoothing) based on local averaging do not perform well for reasonable sample sizes because of the "curse of dimensionality". The PPR model is known to by-pass this curse of dimensionality resulting to good estimators even in higher dimensions. For different estimation procedures on Projection Pursuit Regression we refer to Chen (1991). See also Huber (1985) for a discussion on projection pursuit techniques. By the use of the PPR for estimating conditional expectation, we can rewrite the model we started with as the GACE model introduced in Equation (4) in Section 1. The following tables show that generalized ACE outperforms ACE in case of interaction between variables (In the simulated models we take $X_1, X_2 \sim U(-1, 1)$ and $\epsilon \sim N(0, 1)$).

Model 2 is considered, because right side function in PPR model cannot represent $e^{x_1 x_2}$. There is a variant of the PPR which can include variables of mixed types. Hence in place of the classical PPR estimate of conditional expectation, one can use the PPR for mixed variable types. See Laghi and Lizzani (1999) for more details. Using this variant the GACE model can also be used with covariates of mixed types.

5.1 Testing in the GACE model

Having built the ACE and GACE models, one can test to determine which model is correct. At first, test for multiple linear regression versus ACE and then test for ACE versus GACE. The first test can be done by using the testing procedure mentioned in Section 4, doing the test sequentially for all the variables. If this

Table 3: R square values

Model	R^2 GACE	R^2 ACE	R^2 Regression
$Y = X_1X_2 + \epsilon$	0.93991880	0.70530659	0.01187
$Y = \exp(X_1X_2) + \epsilon$	0.95895668	0.67364599	0.01091
$\log(Y) = \sin(2\pi X_1 + 2\pi X_2) + \epsilon$	0.9779115	0.1044552	0.01841
Real Data 3	0.9010383	0.8597254	0.7633

Table 4: Values of measure of prediction Q

Model	GACE	ACE	Regression
$Y = X_1X_2 + \epsilon$	0.90111441	-0.16947620	-0.01003546
$Y = \exp(X_1X_2) + \epsilon$	0.91478150	0.18471331	-0.01667949
$\log(Y) = \sin(2\pi X_1 + 2\pi X_2) + \epsilon$	0.911475179	-1.806580280	-0.008369527
Real Data 3	0.7668139	0.7244917	-1.4728816

test fails to reject then we need not test for ACE versus GACE. For testing ACE versus GACE, we can follow the same testing procedure. In this section, following Zheng (1996), we develop a new testing procedure. We want to test,

$$H_0 : \beta(Y) = \sum_{i=1}^{p} \alpha_i(X_i) + \epsilon, \quad \text{versus} \quad H_1 : \beta(Y) = \alpha(X_1, X_2, \ldots, X_p) + \epsilon.$$

Consider $u_i = \beta(y_i) - \sum_{j=1}^{p} \alpha_j(x_{ji})$, under actual model $E[u_i|\mathbf{x}_i] = 0$. Also, observe that, $E[u_iE[u_i|\mathbf{x}_i]p(\mathbf{x}_i)] = E[E^2[u_i|\mathbf{x}_i]p(\mathbf{x}_i)] \geq 0$ and equals zero only under actual model. Hence, if null hypothesis is true, then $E[u_iE[u_i|\mathbf{x}_i]p(\mathbf{x}_i)] = 0$ only under the null. Hence we can use a sample analogue of this expectation using the nonparametric estimate of the conditional expectation as a test statistic for testing ACE versus GACE. This test is shown to be consistent in Zheng (1996). See also Fan and Li (1996). Testing for models in this order after fitting the different models will give an idea of whether any transformations are needed for variables. Testing for correctness of any model can be done in this way. A test for GACE versus multiple linear regression can also be performed in this manner.