PROJECTION PURSUIT REGRESSION

BY

JEROME H. FRIEDMAN, MARK JACOBSON and WERNER STUETZLE

TECHNICAL REPORT NO. 146
MARCH 1980

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
PROJECTION PURSUIT REGRESSION

by

JEROME H. FRIEDMAN, MARK JACOBSON and WERNER STUETZLE

TECHNICAL REPORT NO. 146
MARCH 1980

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
PROJECTION PURSUIT REGRESSION*

Jerome H. Friedman
Stanford Linear Accelerator Center
Stanford University
Stanford, California 94305

Mark Jacobson
Department of Statistics
Stanford University
Stanford, California 94305

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

(Presented at IMS Special Topics on Regression, Boulder, Colorado,
October 22-26, 1979)

ABSTRACT

A new method for nonparametric multiple regression is presented.
The procedure models the regression surface as a sum of general
smooth functions of linear combinations of the predictor variables
in an iterative manner. This procedure is more general than stan-
ard stepwise and stagewise regression procedures, is affine invar-
iant, does not require the definition of a metric, and lends itself
to graphical interpretation.

*Work supported by the Department of Energy under contract number
DE-AC03-76SF00515
(Submitted to Journal of the American Statistical Association)
1. INTRODUCTION

In the regression problem, one is given a p-dimensional random vector \( \mathbf{x} \), the components of which are called predictor variables, and a random variable \( y \), which is called the response. The aim of regression analysis is to estimate the conditional expectation of \( y \) given \( \mathbf{x} \) on the basis of a sample \( \{(x_i, y_i) : i = 1, 2, \ldots, n\} \). Typically, one makes the assumption that the functional form of the regression surface is known, reducing the problem to that of estimating a set of parameters. To the extent that this model is correct, such parametric procedures can be successful; unfortunately, model correctness is difficult to verify in practice, and an incorrect model can yield misleading results. For this reason, there is a growing interest in nonparametric methods which make only a few very general assumptions about the regression surface.

The most extensively studied nonparametric regression techniques (kernel, nearest-neighbor, and spline smoothing) are based on p-dimensional local averaging: the estimate of the regression surface at a point \( \mathbf{x}_0 \) is the average of the responses of those observations with predictors in a neighborhood of \( \mathbf{x}_0 \). These techniques can be shown to have desirable asymptotic properties (Stone, 1977). In high dimensional settings, however, they do not perform well for reasonable sample sizes. The reason is the inherent sparsity of high-dimensional samples. This is illustrated by the following simple example: Let \( \mathbf{x} \) be uniformly distributed over the unit hypercube in \( \mathbb{R}^p \), and consider local averaging over hypercubical neighborhoods. If the dimensions of the neighborhood are chosen to cover 10% of the range of each coordinate, then it will (on the average) contain only \((.1)^p\) of the sample, and thus will
nearly always be empty. If, on the other hand, one adjusts the neighborhood to contain 10% of the sample, it will cover (on the average) $(.1)^{1/10} \approx .8$ of the range of each coordinate. This problem of sparsity basically limits the success of direct p-dimensional local averaging.

The successful nonparametric regression procedures that have been proposed are based on successive refinement. A hierarchy of models of increasing complexity is formulated. The complexity of a model is the number of degrees of freedom used to fit it. The aim is to find the particular model that, when estimated from the data, best approximates the regression surface. The search usually proceeds through the hierarchy in a stepwise manner. At each step, the model of the subsequent level of the hierarchy that best fits the data is selected. The process stops when this fit does not yield significant improvement. Since the sample size limits the depth in the hierarchy from which the model can be chosen, these procedures will be successful to the extent that the regression surface can be approximated by models in the high levels of the hierarchy.

Applying this concept with a hierarchy of polynomial functions of the predictors leads to the stepwise, stagewise, and all-subsets polynomial regression procedures. These procedures have proven to be successful in many applications. Unfortunately, regression surfaces occurring in practice often are not represented well by low-order polynomials (e.g., surfaces with asymptotes); use of higher-order polynomials is limited by considerations of sample size and computational feasibility.

A hierarchy of piecewise constant (Sonquist, 1970) or piecewise linear (Breiman and Meisel, 1976, Friedman, 1979) models leads to recur-
sive partitioning regression. These procedures basically operate as follows: For a particular predictor and a value of this predictor, the predictor space is split into two regions, one projecting to the left and the other to the right of the value. A separate constant or linear model is fit to the sample points lying in each region. The particular predictor and splitting value are chosen to minimize the residual sum of squares over the sample. The procedure is then recursively applied to each of the regions so obtained.

These recursive partitioning methods can be viewed as local averaging procedures, but unlike in kernel and nearest-neighbor procedures, the local regions are adaptively constructed based on the nature of the response variation. In many situations, this results in dramatically improved performance. However, as each split reduces the sample over which further fitting can take place, the number of regions, and thus the number of separate models, is limited.

In this paper, we apply the successive refinement concept in a new way that attempts to overcome the limitations of polynomial regression and recursive partitioning. The procedure is presented in Section 2; implementation specifics are considered in Section 3. In Section 4, we illustrate the procedure by applying it to several data sets. We discuss the merits of this procedure relative to other nonparametric procedures in Section 5, and in Section 6 we offer a paradigm for multivariate analysis.
2. THE ALGORITHM

The regression surface is approximated by a sum of smooth functions of linear combinations of the predictors,

\[ \phi(x) = \sum_{m=1}^{M} S_m(\alpha_m \cdot x). \]

The approximation is constructed in a stepwise manner:

(1) For a given linear combination \( Z = \alpha \cdot x \), construct a smoothed representation \( S(Z) \) of the training responses \( \{y_i: i=1,2,\ldots,n\} \) as ordered in ascending value of \( Z \).

Define \( \alpha_m \) to be the coefficient vector of the linear combination that minimizes the sum of squares of deviations from the smooth, i.e.,

\[ \alpha_m = \min_{\alpha} \sum_{i=1}^{n} (y_i - S(\alpha \cdot x_i))^2, \]

and take \( S_m \) to be the corresponding smooth.

(2) Form the residuals from this smooth and substitute them for the responses: \( y_i \leftarrow r_i = y_i - S_m(\alpha_m \cdot x_i) \).

(3) Iterate (1) and (2) until the residual sum of squares does not decrease significantly.

This procedure directly follows the successive refinement concept outlined in the previous section: The models at the \( m \)th level of the hierarchy are sums of \( m \) smooth functions of arbitrary linear combinations of the predictors.

Standard additive models approximate the regression surface by a
sum of functions of the individual predictors. Such models are not completely general in that they cannot deal with interactions of predictors. Considering functions of linear combinations of the predictors removes this limitation. As an example, consider a simple interaction: \( Y = X_1 X_2 \). A standard additive model cannot represent this multiplicative dependence; however, if we define \( Z_1 = \frac{1}{\sqrt{2}} (X_1 + X_2) \), \( Z_2 = \frac{1}{\sqrt{2}} (X_1 - X_2) \), we obtain

\[ Y = \frac{1}{2} Z_1^2 - \frac{1}{2} Z_2^2 \],

a model additive in linear combinations of the predictors. The introduction of arbitrary linear combinations of predictors allows the representation of general regression surfaces.

3. IMPLEMENTATION

Implementation of the algorithm requires a suitable procedure for smoothing one-dimensional sequences and a means for determining the linear combination that yields the smallest residual sum of squares from the smooth.

Our choice of a smoother is guided by the fact that the model underlying traditional smoothing procedures (i.e., \( y_i = S(x_i) + \varepsilon_i \), \( \{\varepsilon_i\} \text{ iid, } S \text{ "smooth"} \)) is not appropriate. Our model seeks to explain response variability by not just one smoothed sequence, but by a sum of smooths of several sequencings of the response (as induced by the several linear combinations of the predictors). High variability encountered in a particular sequence may be caused by smooth dependence of the response on other linear combinations. As a consequence, we use a variable-bandwidth smoother with large bandwidth in regions of high variability. To reduce bias, especially at the ends of the sequence, we smooth by locally linear,
rather than locally constant, fitting. Furthermore, each observation is omitted from the local average that determines its smoothed value. This "cross-validation" makes the average squared residual a more realistic indicator of variability about the smooth (for example, it is not possible to make the average squared residual arbitrarily small by reducing the bandwidth). To protect against isolated outliers, each response is first replaced by the median of its two adjacent neighbors and itself ("running medians of three").

Our smoothing algorithm makes four passes over the data:

(1) Running medians of three.

(2) Estimation of the response variability at each point by the average squared residual of a local linear fit with constant bandwidth.

(3) Smoothing of these variance estimates by a fixed-bandwidth moving average.

(4) Smoothing of the sequence (1) by local linear fits with bandwidths determined by the smoothed local variance estimates (3).

The basic parameter of the procedure is the characteristic bandwidth—the percentage of the observations over which local averaging takes place. This parameter, set by the user, defines smoothness in the notion of a "smooth function".

For a particular linear combination, this smoother yields a residual sum of squares from the corresponding smooth. The optimal linear combination is sought by numerical optimization (Projection Pursuit, Friedman and Tukey, 1974). Considerations governing the choice of the optimization
algorithm are:

- the function evaluations are expensive (each one requires several passes over the data);
- the search usually starts far from the solution;
- the search can be restricted to the unit sphere in $\mathbb{R}^p$.

For these reasons, we have chosen a Rosenbrock method (Rosenbrock, 1960) modified to search on the unit sphere.

Projection pursuit regression can be implemented with or without readjustment of the smooths along previously determined linear combinations when a new linear combination has been found. In the terminology of linear regression, this would correspond to the difference between a stepwise and a stagewise procedure. We have implemented the stepwise version.

In some situations, it may be useful to restrict the search for solution directions to the set of predictors ("projection selection") rather than allowing for linear combinations. Although the resulting additive model cannot represent completely general regression surfaces, it is still more general than linear regression in allowing for general smooth functions rather than only linear functions of the predictors. Projection selection is computationally less expensive than full projection pursuit and the resulting models are often more easily interpreted. Another strategy is to run projection selection, followed by projection pursuit, thereby separating the additive and interactive parts of the model.
4. EXAMPLES

In this section, we present and discuss the results of applying projection pursuit regression (PPR) to three data sets. (A FORTRAN program implementing the PPR procedure is available from the authors upon request). For all three examples, the iteration was terminated when the next step did not improve the goodness-of-fit by more than 10%. The characteristic bandwidth of the one-dimensional smoother was taken to be 30% for the first two examples and 10% for the third. All predictors were standardized to have median zero and interquartile range one. (Although PPR is affine invariant, widely different scales can cause problems for the numerical optimizer.)

The first example is artificially constructed to illustrate how PPR models interactions between predictors. A sample of 200 observations was generated according to the simplest interaction model $Y = X_1X_2 + \epsilon$ with $(X_1, X_2)$ uniformly distributed in $(-1,1) \times (-1,1)$ and $\epsilon \sim N(0,.04)$. Figure 1a shows $Y$ plotted against $X_2$ with the corresponding smooth. ($Y$ is plotted on the vertical axis, $X_2$ on the horizontal axis. The "+" symbols represent data points, numbers indicate more than one data point. The smooth is represented by "*" symbols.) Figure 1b shows $Y$ plotted against the first linear combination $Z_1 = \alpha_1 \cdot X$, $\alpha_1 = (.71,.70)$, found by projection pursuit, with the corresponding smooth $S_1(\alpha_1 \cdot X)$. Figure 1c shows the residuals $r_1 = Y - S_1(\alpha_1 \cdot X)$ plotted against the second linear combination $Z_2 = \alpha_2 \cdot X$, $\alpha_2 = (.72,-.69)$, together with $S_2(\alpha_2 \cdot X)$. Figure 1d shows the residuals $r_2 = Y - S_1(\alpha_1 \cdot X) - S_2(\alpha_2 \cdot X)$ plotted against the third linear combination with the corresponding smooth. This projection was not accepted because the improvement in goodness-of-fit measure was
below the improvement threshold. It is evident from inspection of Figure 1d that this projection does not substantially contribute to the model. The pure quadratic shapes of $S_1$ and $S_2$, together with the corresponding coefficient vectors $\alpha_1$ and $\alpha_2$, reveal that PPR has expressed the model $Y = X_1 x_2$ in the additive form $Y = \frac{1}{4} (X_1 + x_2)^2 - \frac{1}{4} (x_1 - x_2)^2$.

In the second example, PPR was applied to air pollution data. The data (213 observations) were taken from the contaminant and weather summary of the Bay Area Pollution Control District (Technical Services Division, 993 Ellis Str., San Francisco). In this example, we study the relation between the amount of suspended particulate matter ($Y$) and predictor variables mean wind speed ($X_1$), average temperature ($X_2$), insolation ($X_3$), wind direction at 4:00 A.M. ($X_4$) and 4:00 P.M. ($X_5$) at the San Jose measuring station. Figures 2a-2c show the three accepted smooths $S_1(\alpha_1 \cdot X)$, $S_2(\alpha_2 \cdot X)$ and $S_3(\alpha_3 \cdot X)$ plotted against their corresponding solution linear combinations. The points are obtained by adding the residuals from the final model to each smooth. The first projection (Figure 2a) shows that a good indicator of suspended particulate matter is (standardized) temperature minus wind speed. For small values of this indicator, the amount of pollution is seen to be roughly constant; for higher values, there is a strong linear dependence. The second smooth (Figure 2b) and the corresponding direction (essentially $X_4$) show a much smaller pollutant dependence on 4:00 A.M. wind direction. The third projection suggests an additional dependence on 4:00 P.M. wind direction, but the effect, if any, is clearly small.

In order to illustrate PPR on highly structured data, common in the physical sciences, we apply it to data taken from a particle physics
experiment (Ballam, et. al., 1971). This data set (500 observations) is described in Friedman and Tukey (1974). Here we study the combined energy of the three $\pi$ mesons ($Y$) as a function of the six other variables.

Figure 3a shows $Y$ plotted against the first linear combination and the corresponding smooth found in the first iteration. Figures 3b-3d show the final smooths for the first three of the 9 accepted projections. As in Figures 2a-2c, we show the residuals from the final model added to the final smooths. Note the substantial change in the first smooth due to readjustment for subsequent projections (see paragraph 5 of Section 3). Note also the striking nonlinearity in Figures 3c and 3d and the high degree of structuring in the data expressed by the fact that the model explains over 99% of the variance.
5. DISCUSSION

Although simple in concept, projection pursuit regression overcomes many limitations of other nonparametric regression procedures. The sparsity limitation of kernel and nearest-neighbor techniques is not encountered, since all estimation (smoothing) is performed in a univariate setting. PPR is affine invariant and does not require specification of a metric. Unlike recursive partitioning, PPR does not split the sample, thereby allowing, when necessary, more complex models. In addition, interactions of predictors are directly considered.

One can view linear regression, projection selection, and full projection pursuit as a group of regression procedures ordered in ascending generality. Linear regression models the regression surface as a sum of linear functions of the predictors. Projection selection allows for nonlinearity by modeling with general smooth functions of the predictors. Full projection pursuit allows for interactions by modeling with general smooth functions of linear combinations of the predictors.

PPR is computationally quite feasible. For increasing sample size \( n \), dimensionality \( p \), and number of iterations \( M \), the computation required to construct the model grows as \( Mnp \log n \).

As seen in the examples, an important feature of PPR is that the results of each iteration can be represented graphically, facilitating interpretation. This pictorial output can be used to adjust the parameters of the procedure (smoother bandwidth and termination criterion).

Under- or over-smoothing is easily detected visually. Whether a
particular projection affects a significant improvement in the model can be judged subjectively by viewing its smooth and the corresponding residuals. One can also employ a more formal procedure based on cross validation (see Stone, 1979).

The PPR procedure can clearly be applied to the residuals from any initial model. If the initial model does not fit the data well, PPR will so indicate by augmenting it.

All stepwise procedures have difficulties modeling regression surfaces that cannot be represented well by models in the high levels of their hierarchies. In the case of PPR, such surfaces are those that cannot be represented well by a sum of general smooth functions of linear combinations of the predictors (e.g., bell-shaped regression surfaces in $\mathbb{R}^D$).

6. THE PROJECTION PURSUIT PARADIGM

Although presented here in the context of regression, the projection pursuit concept can serve as a paradigm for a wide class of problems in multivariate analysis. As in regression, the goal is to describe multivariate "structure" by a combination of univariate models. The regression algorithm of Section 2 is one application of the projection pursuit paradigm:

1. Select an initial model and a global goodness-of-fit measure.

2. Project the current model and data on a one-dimensional subspace. Modify the model to agree with the data in this projection. The subspace is chosen to yield the
largest improvement in goodness-of-fit.

(3) Repeat (2) until no significant improvement is possible.

The goodness-of-fit measure and the method of model modification depend on the type of problem. For the regression problem, goodness-of-fit is measured by the residual sum of squares; the model is modified by adding the smooth of the residuals along the optimal projection. Applications of the projection pursuit paradigm to classification and density estimation are currently under investigation.
REFERENCES

Ballam, J., Chadwick, G.B., Guiragossian, Z.C.G., Johnson, W.B.,
reactions $\pi^- p \rightarrow \pi^- \pi^+ p$ and $\pi^+ p \rightarrow \pi^+ \pi^- p$ at 16 GeV/C. Phys. Rev., 4, pp 1946-1947.


multiple regression. In Smoothing Techniques for Curve Estimation


search strategy. Institute for Social Research, Univ. of Michigan,
Ann Arbor.

Stone, C.J. (1977). Nonparametric regression and its applications (with

Stone, C.J. (1979). Admissible selection of an accurate and parsimonious
normal linear regression model. UCLA Report.
FIGURE CAPTIONS

FIGURE 1a: \[ Y = X_1 X_2 + \varepsilon, \varepsilon \sim N(0, 0.04), \text{ vs. } X_2. \]

FIGURE 1b: Y vs. first solution linear combination (projection).

FIGURE 1c: Residuals from first projection smooth vs. second solution linear combination.

FIGURE 1d: Residuals from first two projection smooths vs. third solution linear combination.

FIGURES 2a-2c: Air pollution (suspended particulate matter) final model smooths (with residuals from model added) vs. corresponding solution linear combinations.

FIGURE 3a: Combined energy of three \( \pi \) mesons (particle physics data) vs. first solution linear combination.

FIGURES 3b-3d: First three (out of nine) final model smooths (with residuals added) vs. corresponding solution linear combinations.
PROJECTION 2 TRIAL 1
SOLUTION AVERAGE SQUARED RESIDUAL = 0.7218 - 0.6921
SOLUTION PROJECTION AND SMOOTH.

AXIS

PROJ ON X

PROJ ON Y

AXIS

FIGURE 1c
FIGURE 2a
PROJECTION 3 SOLUTION SMOOTH WITH RESIDUALS.
SOLUTION AXIS= 0.1655 0.2132 0.0059 -0.0463 0.9618

FIGURE 2c
SOLUTION AVERAGE SQUARED RESIDUAL = 3.86610
SOLUTION AXIS =
0.8275 0.5369 0.0041 -0.1641 0.0 0.0
SOLUTION PROJECTION AND SMOOTH.

PROJ
ON X
AXIS
1200000020114123415246655836784540012834339949432250032011
21.1097 I+
20.5307 I+
20.0564 I+
19.5298 I+
19.0032 I+
18.4766 I+
17.9500 I+
17.4234 I+
16.8968 I+
16.3702 I+
15.8436 I+
15.3170 I+
14.7904 I+
14.2638 I+
13.7372 I+
13.2106 I+
12.6840 I+
12.1573 I+
11.6307 I+
11.1041 I+
10.5775 I+
10.0509 I+
9.5243 I+
9.0077 I+
8.4711 I+
8.0445 I+
7.5179 I+
7.0257 I+
6.5233 I+
6.0209 I+
5.5185 I+
5.0281 I+
4.5280 I+
4.0257 I+
3.5248 I+
3.0238 I+
2.5229 I+
2.0218 I+
1.5208 I+
1.0197 I+
0.5207 I+
0.0454 I+

FIGURE 3b
Figure 3c
FIGURE 3d