Projection Pursuit Regression

JEROME H. FRIEDMAN and WERNER STUETZLE*

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. It is more general than standard stepwise and stagewise regression procedures, does not require the definition of a metric in the predictor space, and lends itself to graphical interpretation.

KEY WORDS: Nonparametric regression; Smoothing; Projection pursuit; Surface approximation.

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 $\{(\mathbf{x}_i, y_i): i = 1, 2, \ldots, n\}$. Typically, one assumes 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 R^{10}, and consider local averaging over hypercubical neighborhoods. If the dimensions of the neighborhood are chosen to cover 10 percent of the range of each coordinate, then it will (on the average) contain only $(.1)^{10}$ of the sample, and thus will nearly always be empty. If, on the other hand, one adjusts the neighborhood to contain 10 percent of the sample, it will cover (on the average) $(.1)^{1/10} \simeq 80$ percent of the range of each coordinate. This problem of sparsity basically limits the success of direct p-dimensional local averaging. In addition, these methods do not provide any comprehensible information about the nature of the regression surface.

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. Since the sample size limits the complexity of the models that can be used, these procedures will be successful to the extent that the regression surface can be approximated by models on levels of low complexity in 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 recursive 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 kernel and nearest-

* Jerome H. Friedman is Group Leader, Computation Research Group, Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94305. Werner Stuetzle is with the Stanford Linear Accelerator Center and is Assistant Professor, Department of Statistics, Stanford University, Stanford, CA 94305. The authors' work was supported by the Department of Energy under Contract DE-AC03-76SF00515. The authors gratefully acknowledge the help of Mark Jacobson during the early stages of this work and thank the editor for many helpful comments.

© Journal of the American Statistical Association
December 1981, Volume 76, Number 376
Theory and Methods Section

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. Univariate smoothing is discussed in Section 3; implementation specifics are considered in Section 4. In Section 5 we illustrate the procedure by applying it to several data sets. The merits of this procedure, relative to other nonparametric procedures, are discussed in Section 6. In Section 7 we relate projection pursuit regression to the projection pursuit technique for cluster analysis presented by Friedman and Tukey (1974).

2. THE ALGORITHM

The regression surface is approximated by a sum of empirically determined univariate functions S_{α_m} of linear combinations of the predictors:

$$\varphi(\mathbf{X}) = \sum_{m=1}^{M} S_{\alpha_m}(\boldsymbol{\alpha}_m \cdot \mathbf{X}), \tag{1}$$

where $\boldsymbol{\alpha}_m \cdot \mathbf{X}$ denotes the inner product. The approximation is constructed in an iterative manner.

1. Initialize current residuals and term counter

$$r_i \leftarrow y_i, \quad i = 1 \ldots n$$
$$M \leftarrow 0.$$

(We assume that the response is centered: $\sum y_i = 0$.)

2. Search for the next term in the model. For a given linear combination $Z = \boldsymbol{\alpha} \cdot \mathbf{X}$, construct a smooth representation $S_{\alpha}(Z)$ of the current residuals as ordered in ascending value of Z (see Sec. 3). Take as a *figure of merit* (criterion of fit) $I(\boldsymbol{\alpha})$ for this linear combination the fraction of so far unexplained variance that is explained by S_{α}:

$$I(\boldsymbol{\alpha}) = 1 - \sum_{i=1}^{n} (r_i - S_{\alpha}(\boldsymbol{\alpha} \cdot \mathbf{x}_i))^2 \Big/ \sum_{i=1}^{n} r_i^2. \tag{2}$$

Find the coefficient vector $\boldsymbol{\alpha}_{M+1}$ that maximizes $I(\boldsymbol{\alpha})$ (*projection pursuit*) $\boldsymbol{\alpha}_{M+1} = \max_{\alpha}{}^{-1} I(\boldsymbol{\alpha})$, and the corresponding smooth $S_{\alpha_{M+1}}$.

3. Termination. If the figure of merit is smaller than a user-specified threshold, stop. (The last term is not included in the model.) Otherwise, update the current residuals and the term counter

$$r_i \leftarrow r_i - S_{\alpha_{M+1}}(\boldsymbol{\alpha}_{M+1} \cdot \mathbf{x}_i), \quad i = 1 \ldots n$$
$$M \leftarrow M + 1$$

and go to Step 2.

This procedure directly follows the successive refinement concept outlined in the previous section: The models at the mth 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. For example, consider a simple interaction: $Y = X_1 X_2$. A standard additive model cannot represent this multiplicative dependence; however, Y can be expressed in the form (1), with $\boldsymbol{\alpha}_1 = (1/\sqrt{2})(1, 1)$, $\boldsymbol{\alpha}_2 = (1/\sqrt{2})(1, -1)$, $S_1(Z) = \frac{1}{2}Z^2$, $S_2(Z) = -\frac{1}{2}Z^2$. The introduction of arbitrary linear combinations of predictors allows the representation of general regression surfaces.

3. UNIVARIATE SMOOTHING

The purpose of smoothing a set of observations $\{y_i, z_i\}_{i=1}^n$, sequenced in ascending order of z, is to produce a decomposition $y_i = S(z_i) + r_i$, where S is a smooth function and the r_i are called residuals. The degree of smoothness of a function S can be formally defined (e.g., $\int S''^2(z)\, dz$), but for the purpose of this discussion an intuitive notion of smoothness will be sufficient. Many procedures for smoothing have been described (Tukey 1977; Cleveland 1979; Gasser and Rosenblatt 1979). They are based on the notion of local averaging:

$$S(z_i) = \underset{i-k \le j \le i+k}{\text{AVE}} (y_j),$$

with suitable adjustment for the boundaries. Here AVE can denote the mean, median, or other ways of "averaging." The parameter k defines the bandwidth of the smoother.

The assumption underlying traditional smoothing procedures is that the observed responses y_i are generated according to the model $y_i = f(x_i) + \epsilon_i$, ϵ_i iid, $E(\epsilon_i) = 0$, f smooth. The resulting smooth S is then taken as an estimate for f. Choosing too small a bandwidth will tend to increase the variance component of the mean squared error of the estimate, whereas too large a bandwidth may increase the bias. The optimum bandwidth will, of course, depend on f and the variance of ϵ, which are generally unknown. Formal methods for estimating the optimal bandwidth using cross-validation have been proposed (Wahba and Wold 1975). Often, however, the degree of smoothing is determined experimentally. One attempts to use as large a bandwidth as possible, subject to the smooth not lying systematically above or below the data in any region (oversmoothing).

Our design of a smoother is guided by the fact that the model underlying traditional smoothing procedures 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 predic-

tors). High local variability encountered in a particular sequence may be caused by smooth dependence of the response on other linear combinations. In order to preserve the ability of fitting this structure in further iterations, it is important to avoid accounting for it by spurious fits along existing directions. Consequently, we use a variable bandwidth smoother. An average smoother bandwidth is specified by the user. The actual bandwidth used for local averaging at a particular value of Z can be larger or smaller than the average bandwidth. Larger bandwidths are used in regions of high local variability of the response.

To reduce bias, especially at the ends of the sequence, we smooth by locally linear, rather than locally constant, fitting (Cleveland 1979). 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 (e.g., it is not possible to make the average squared residual arbitrarily small by reducing the bandwidth). To protect against isolated outliers, we use running medians of three (Tukey 1977) as a first pass in our smoother.

Our smoothing algorithm thus makes four passes over the data:

1. Running medians of three;
2. Estimating response variability at each point by the average squared residual of a locally linear fit with constant bandwidth;
3. Smoothing these variance estimates by a fixed-bandwidth moving average; and
4. Smoothing the sequence obtained by pass (1) by locally linear fits with bandwidths determined by the smoothed local variance estimates obtained in pass (3).

4. IMPLEMENTATION

For a particular linear combination, the smoother yields a residual sum of squares from the corresponding smooth. The optimal linear combination is sought by numerical optimization. Considerations governing the choice of the optimization algorithm are that (a) the function evaluations are expensive (each one requires several passes over the data); (b) the search usually starts far from the solution; and (c) the search can be restricted to the unit sphere in R^p. For these reasons we chose a Rosenbrock method (Rosenbrock 1960) modified to search on the unit sphere. The search is started at the best coordinate direction. On any given search there is no guarantee that the global optimum will be found. If the local optimum is not acceptable, the search is restarted at random directions. This guards against premature termination. If the local optimum is acceptable but not identical to the global optimum, no great harm is done because a new search is performed in the next iteration on an object function for which the previous optima have been deflated.

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 (*backfitting*). 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.

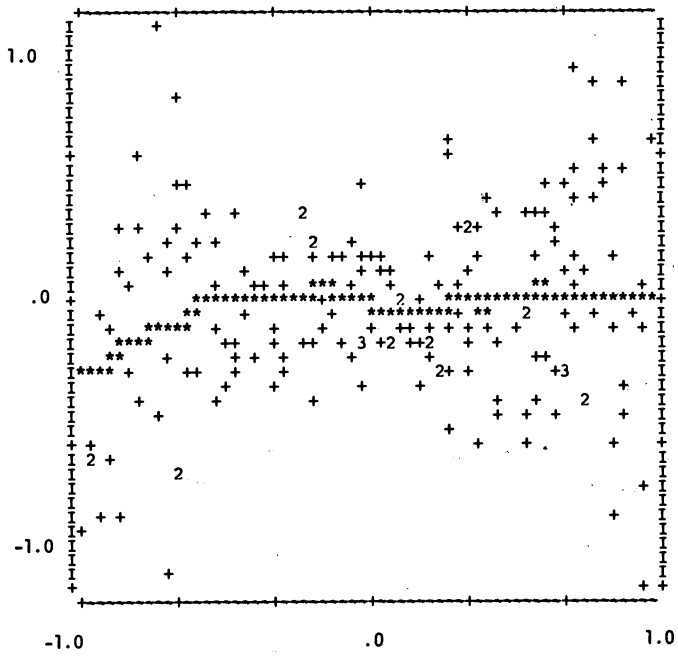
In some situations it may be useful to restrict the search for solution directions to the set of predictors (*projection selection*) rather than allowing for linear combinations. Although the resulting additive model cannot represent completely general regression surfaces, it is still more general than 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. One could also run projection selection, followed by projection pursuit, thereby separating the additive and interactive parts of the model. Another strategy would be to run projection pursuit and get some easily interpreted linear combinations (as in Sec. 5, second example, with $X_1 - X_2$, X_4, X_5) and then run projection selection on these directions to see how much is lost. Forming a parametric model based on these directions is another possibility.

5. 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 on request.) For all three examples the iteration was terminated when the figure of merit for the next term was less than .1. The average bandwidth of the one-dimensional smoother was taken to be 30 percent for the first two examples and 10 percent for the third. All predictors were standardized to have median zero and interquartile range one. (Widely different scales can cause problems for the numerical optimizer.)

The first example is artificially constructed to be especially simple in order 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. Figure 1b shows Y plotted against the first linear combination $Z_1 = \boldsymbol{\alpha}_1 \cdot \mathbf{X}$, $\boldsymbol{\alpha}_1 = (.71, .70)$, found by projection pursuit, with the corresponding smooth $S_{\alpha_1}(\boldsymbol{\alpha}_1 \cdot \mathbf{X})$. Figure 1c shows the residuals $r_1 = Y - S_{\alpha_1}(\boldsymbol{\alpha}_1 \cdot \mathbf{X})$ plotted against the second linear combination $Z_2 = \boldsymbol{\alpha}_2 \cdot \mathbf{X}$, $\boldsymbol{\alpha}_2 = (.72, -.69)$, together with $S_{\alpha_2}(\boldsymbol{\alpha}_2 \cdot \mathbf{X})$. Figure 1d shows the residuals $r_2 = Y - S_{\alpha_1}(\boldsymbol{\alpha} \cdot \mathbf{X}) - S_{\alpha_2}(\boldsymbol{\alpha}_2 \cdot \mathbf{X})$ plotted against the third linear combination with the corresponding smooth. This projection was not accepted because the figure of merit was below the threshold. (Note that the figure of merit, as defined in equation (2), measures the improvement in

*Figure 1a. $Y = X_1X_2 + \epsilon$, $\epsilon \sim N(0, .04)$, vs. X_2 (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)*



goodness of fit.) It is evident from inspection of Figure 1d that this projection does not substantially contribute to the model. The pure quadratic shapes of S_{α_1} and S_{α_2}, together with the corresponding coefficient vectors α_1 and α_2, reveal that PPR has essentially expressed the model $Y = X_1X_2$ in the additive form $Y = \frac{1}{4}(X_1 + X_2)^2 - \frac{1}{4}(X_1 - X_2)^2$.

Figure 1b. Y vs. First Solution Linear Combination $\alpha_1 \cdot \boldsymbol{X}$, α_1 = (.71, .70)

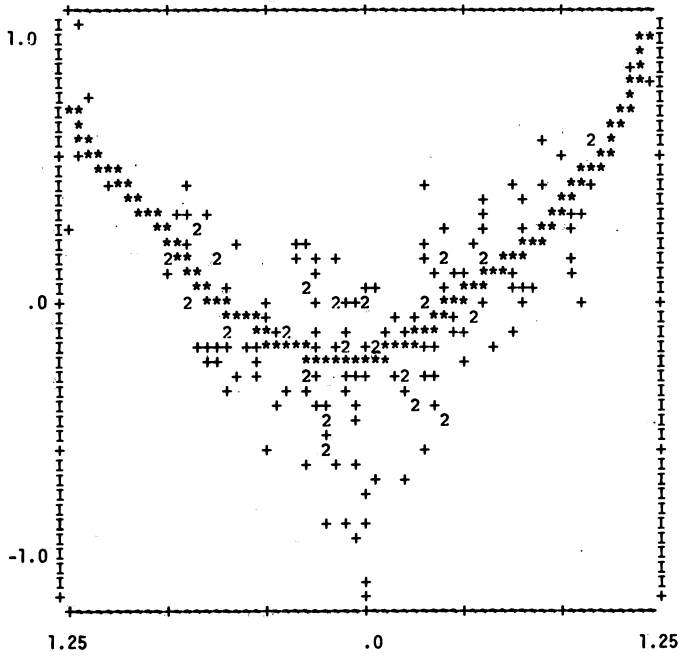
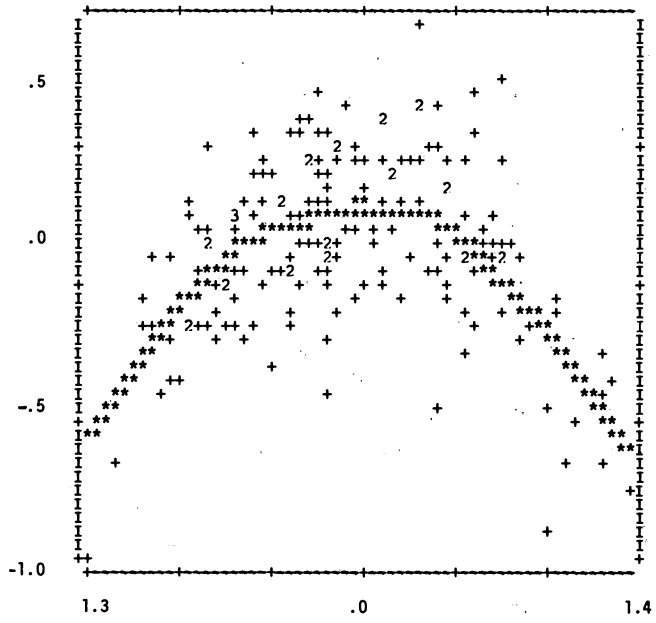
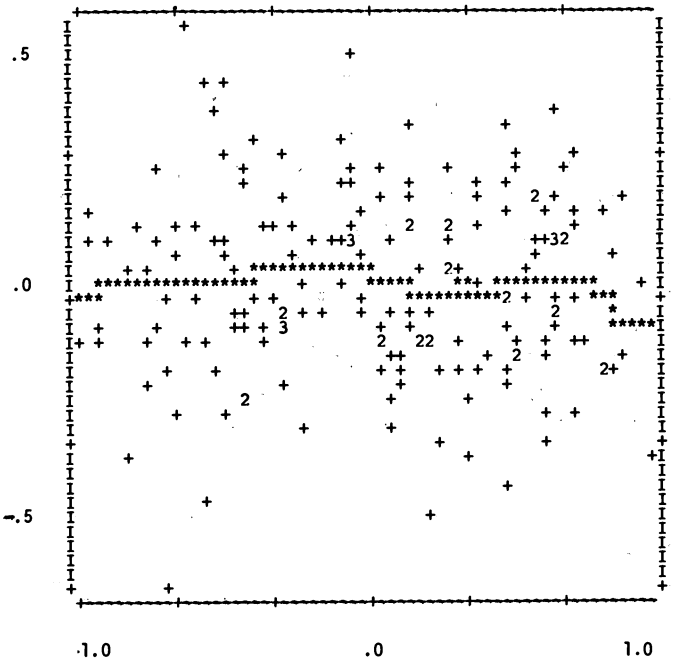


Figure 1c. Residuals From First Solution Smooth vs. Second Solution Linear Combination $\alpha_2 \cdot \boldsymbol{X}$, α_2 = (.72, −.69)



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 Street, San Francisco, CA 94109). 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), and wind direction at 4:00 A.M. (X_4) and 4:00 P.M. (X_5)

Figure 1d. Residuals From First Two Solution Smooths vs. Third Solution Linear Combination $\alpha_3 \cdot \boldsymbol{X}$, α_3 = (−.016, .99)



at the San Jose measuring station. Three projections were accepted. Figures 2a through 2c show the three final smooths (after backfitting) plotted against their corresponding linear combinations. The points plotted are obtained by adding the residuals from the final model to each smooth. The first projection (Fig. 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 (Fig. 2b) and the corresponding direction (essentially X_4) show a much smaller pollutant dependence on 4:00 A.M. wind direction. The third projection (Fig. 2c) suggests an additional dependence on the 4:00 P.M. wind direction, but the effect, if any, is clearly small.

In order to illustrate PPR on highly structured data, which are 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 π 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 through 3d show the final smooths (after backfitting) for the first three of the nine accepted projections. As in Figures 2a through 2c, we show the residuals from the final model added to the final smooths. Note the substantial change in the first smooth due to backfitting, which readjusts for later projections. 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 percent of the variance.

Figure 2a. Air Pollution (suspended particulate matter)—First Solution Smooth S_{α_1}, α_1 = (.83, −.55, .0, .0, .10), With Residuals Added

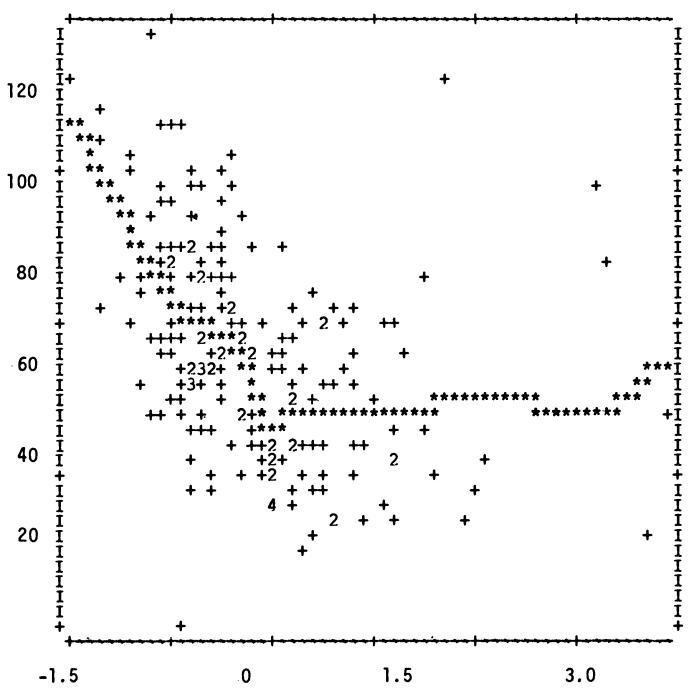
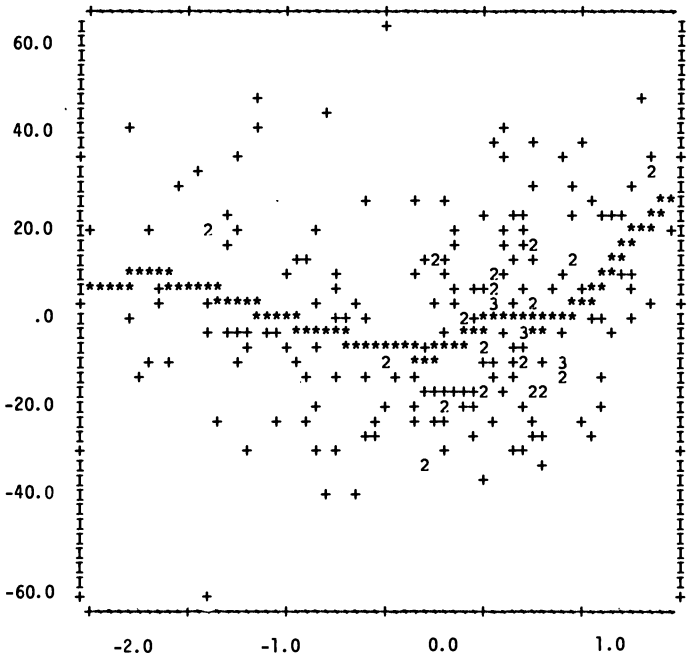


Figure 2b. Air Pollution—Second Solution Smooth S_{α_2}, α_2 = (.16, .29, .17, .91, .16), With Residuals Added



6. DISCUSSION

Although simple in concept, projection pursuit regression overcomes many limitations of other nonparametric regression procedures. The sparsity limitation of kernel

Figure 2c. Air Pollution—Third Solution Smooth S_{α_3}, α_3 = (.16, .21, .01, −.05, .96), With Residuals Added

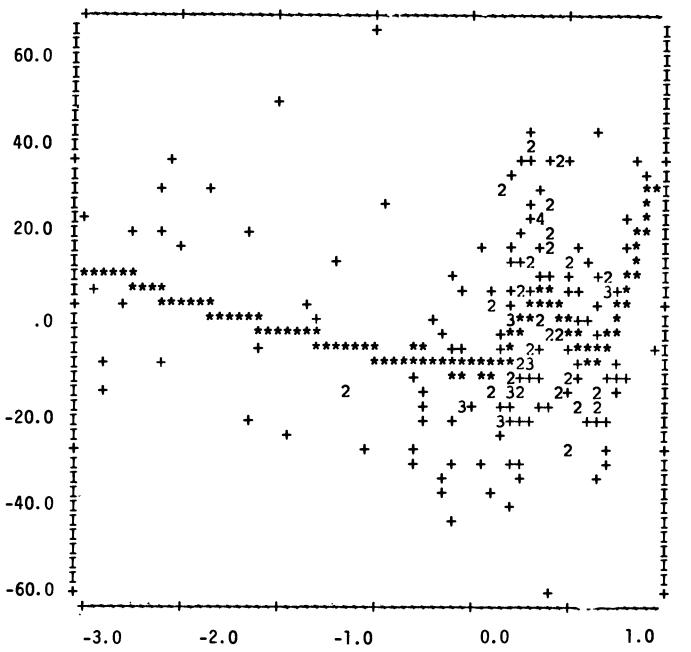
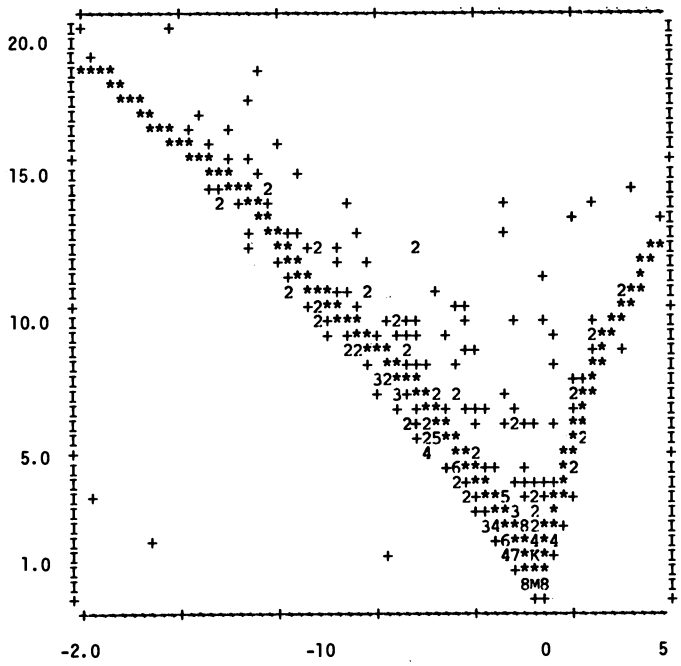


Figure 3a. Combined Energy of Three π Mesons $E_{3\pi}$ (particle physics data) vs. First Solution Linear Combination, α_1 = (.83, .54, .0, −.16, .0, .0), With Corresponding Smooth Found on the First Iteration



and nearest-neighbor techniques is not encountered since all estimation (smoothing) is performed in a univariate setting. PPR does not require specification of a metric in the predictor space. 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

Figure 3b. Particle Physics Data—First Solution Smooth S_{α_1}, α_1 = (.83, .54, .0, −.16, .0, .0), With Residuals Added

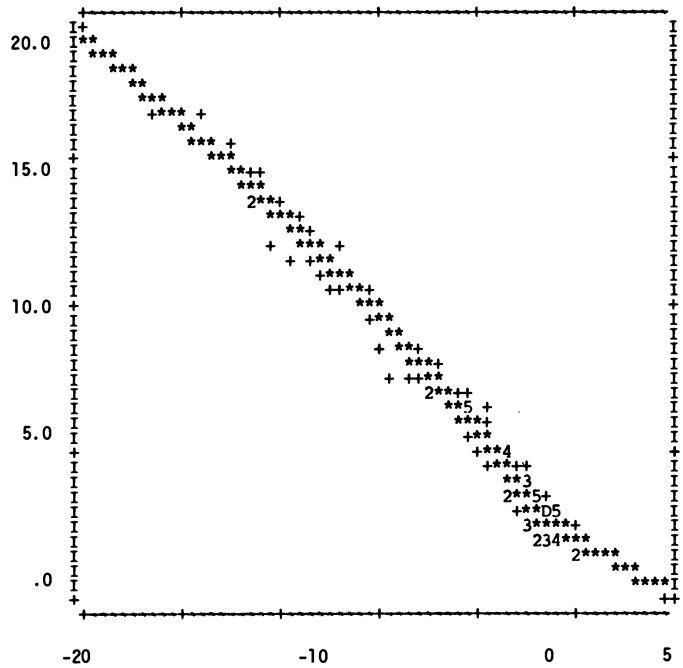


Figure 3c. Particle Physics Data—Second Solution Smooth S_{α_2}, α_2 = (.0, .82, −.05, .0, −.33, .46), With Residuals Added

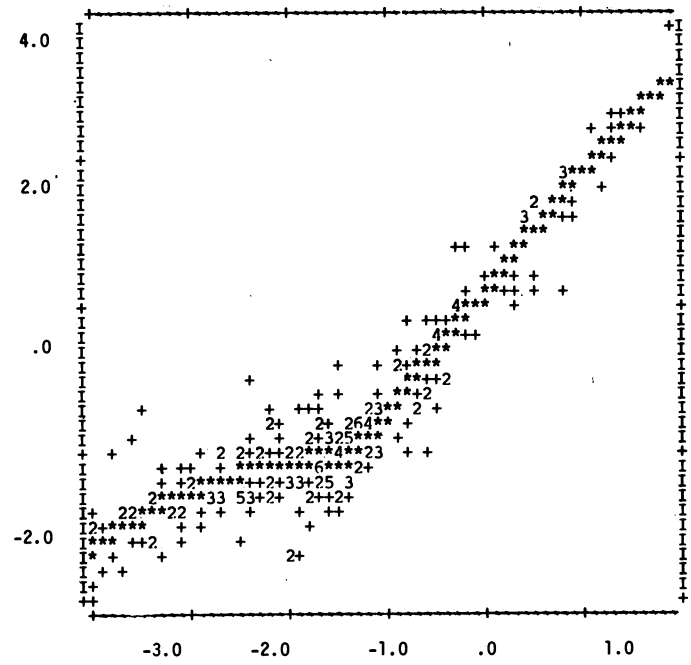
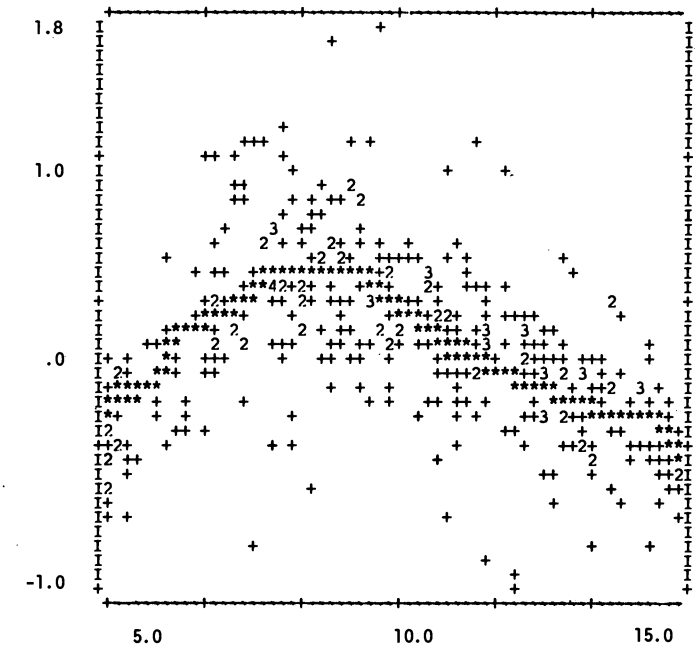


Figure 3d. Particle Physics Data—Third Solution Smooth S_{α_3}, α_3 = (.14, .39, .69, .51, −.16, .26), With Residuals Added



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 $Mpn \log(n)$.

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 main parameters of the procedure, that is, average smoother bandwidth and termination threshold.

The average bandwidth should be chosen as large as possible, subject to the avoidance of oversmoothing. In any projection, whether the smooth systematically deviates from the data is easily detected by visual inspection. Whether a particular projection affects a significant improvement in the model can be judged subjectively by viewing its smooth and the corresponding residuals. Lack of a systematic tendency of the smooth indicates that including this projection into the model would only increase the variance, while not reducing the bias. One can also employ a more formal procedure based on cross-validation (see Stone 1981).

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 the model.

All stepwise procedures have difficulties modeling regression surfaces that cannot be well represented by models of low complexity in their hierarchy. Because models in PPR are sums of functions, each varying only along a single linear combination of the predictors, PPR has difficulties modeling regression surfaces that vary with equal strength along all possible linear combinations.

7. PROJECTION PURSUIT PROCEDURES

The idea of projection pursuit is not a new one. Interpreting high-dimensional data through the use of well-chosen lower-dimensional projections is a standard procedure in multivariate data analysis. The choice of a projection is usually guided by an appropriate figure of merit. If the goal is to preserve interpoint distances as well as possible, then the appropriate figure of merit is the variance of the projected data, leading to projection on the largest principal component. If the purpose is to separate two Gaussian samples with equal covariance matrices, the figure of merit is the error rate of a one-dimensional classification rule in the projection, leading to linear discriminant analysis. In both cases the figure of merit is especially simple and the solution can be found by linear algebra. In a similar spirit, Friedman and Tukey (1974) suggest detecting clusters by searching for clustered projections. Their figure of merit measuring the degree of clustering in a projection (P index) is too complex to be optimized by linear algebra. Instead, the optimal projection was sought by numerical optimization; this was referred to as projection pursuit. As multivariate structure often will not be completely reflected in one projection, it is important to remove structure already discovered (deflate previous optima of the figure of merit), allowing the algorithm to find additional interesting projections. Friedman and Tukey suggest splitting the data into clusters, once a clustered projection has been found, and applying the procedure to the data in each of the clusters separately.

Projection pursuit regression follows a similar prescription. It constructs a model of the regression surface based on projections of the data on planes spanned by the response Y and a linear combination $\boldsymbol{\alpha} \cdot \mathbf{X}$ of the predictors. Here the figure of merit for a projection is the fraction of variance explained by a smooth of Y versus $\boldsymbol{\alpha} \cdot \mathbf{X}$. Structure is removed by forming the residuals from the smooth and substituting them for the response. The model at each iteration is the sum of the smooths that were previously subtracted and thus incorporates the structure so far found.

[*Received February 1980. Revised April 1981.*]

REFERENCES

BALLAM, J., CHADWICK, G.B., GUIRAGOSSIAN, Z.C.G., JOHNSON, W.B., LEITH, D.W.G.S., and MORIGASU, J. (1971), "Van Hove Analysis of the Reactions $\pi^- p \to \pi^- \pi^- \pi^+ p$ and $\pi^+ p \to \pi^+ \pi^+ \pi^-$ at 16 GeV/C," *Physics Review*, 4, 1946–1947.

BREIMAN, L., and MEISEL, W.S. (1976), "General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models," *Journal of the American Statistical Association*, 71, 301–307.

CLEVELAND, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

FRIEDMAN, J.H. (1979), "A Tree-Structured Approach to Nonparametric Multiple Regression," in *Smoothing Techniques for Curve Estimation*, eds. Th. Gasser and M. Rosenblatt, New York: Springer-Verlag, 5–22.

FRIEDMAN, J.H., and TUKEY, J.W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers*, C-23, 881–890.

GASSER, T., and ROSENBLATT, M. (eds.) (1979), "Smoothing Techniques for Curve Estimation," in *Lecture Notes in Mathematics 757*, New York: Springer-Verlag.

ROSENBROCK, H.H. (1960), "An Automatic Method for Finding the Greatest or Least Value of a Function," *Computer Journal*, 3, 175–184.

SONQUIST, J. (1970), "Multivariate Model Building: The Validation of a Search Strategy," Report, Institute for Social Research, University of Michigan, Ann Arbor.

STONE, C.J. (1977), "Nonparametric Regression and Its Applications" (with discussion), *Annals of Statistics*, 5, 595–645.

——— (1981), "Admissible Selection of an Accurate and Parsimonious Normal Linear Regression Model," *Annals of Statistics*, 9, in press.

TUKEY, J.W. (1977), *EDA Exploratory Data Analysis*, Reading, Mass.: Addison–Wesley.

WAHBA, G., and WOLD, S. (1975), "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation," *Communications in Statistics*, 4, 1–17.