SLAC REPORT-355
STAN-LCS-106
UC-405
(M)

INTERPRETABLE PROJECTION PURSUIT*

SALLY CLAIRE MORTON

Stanford Linear Accelerator Center
Stanford University
Stanford, California 94309

OCTOBER 1989

Prepared for the Department of Energy
under contract number DE-AC03-76SF00515

Printed in the United States of America. Available from the National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Road, Springfield, Virginia 22161. Price: Printed Copy A06; Microfiche A01.

*Ph.D Dissertation

Abstract

The goal of this thesis is to modify projection pursuit by trading accuracy for interpretability. The modification produces a more parsimonious and understandable model without sacrificing the structure which projection pursuit seeks. The method retains the nonlinear versatility of projection pursuit while clarifying the results.

Following an introduction which outlines the dissertation, the first and second chapters contain the technique as applied to exploratory projection pursuit and projection pursuit regression respectively. The interpretability of a description is measured as the simplicity of the coefficients which define its linear projections. Several interpretability indices for a set of vectors are defined based on the ideas of rotation in factor analysis and entropy. The two methods require slightly different indices due to their contrary goals.

A roughness penalty weighting approach is used to search for a more parsimonious description, with interpretability replacing smoothness. The computational algorithms for both interpretable exploratory projection pursuit and interpretable projection pursuit regression are described. In the former case, a rotationally invariant projection index is needed and defined. In the latter, alterations in the original algorithm are required. Examples of real data are considered in each situation.

The third chapter deals with the connections between the proposed modification and other ideas which seek to produce more interpretable models. The

modification as applied to linear regression is shown to be analogous to a nonlinear continuous method of variable selection. It is compared with other variable selection techniques and is analyzed in a Bayesian context. Possible extensions to other data analysis methods are cited and avenues for future research are identified. The conclusion addresses the issue of sacrificing accuracy for parsimony in general. An example of calculating the tradeoff between accuracy and interpretability due to a common simplifying action, namely rounding the binwidth for a histogram, illustrates the applicability of the approach.

Acknowledgments

I am grateful to my principal advisor Jerry Friedman for his guidance and enthusiasm. I also thank my secondary advisors and examiners: Brad Efron, Persi Diaconis, Art Owen, Ani Adhikari, Joe Oliger; my teachers and colleagues: Kirk Cameron, David Draper, Tom DiCiccio, Jim Hodges, Iain Johnstone, Mark Knowles, Michael Martin, Daryl Pregibon, John Rolph, Joe Romano, Anne Sheehy, David Siegmund, Hal Stern; and my friends: Mark Barnett, Ginger Brower, Renata Byl, Ray Cowan, Marty Dart, Judi Davis, Glen Diener, Heather Gordon, Holly Haggerty, Curt Lasher, Arla LeCount, Alice Lundin, Michele Marincovich, Mike Strange and Joan Winters.

This work was supported in part by the Department of Energy, Grant DE-AC03-76F00515.

I dedicate this thesis to
my parents, sister, and brothers,
who inspire me by example.

Table of Contents

Figure Captions

List of Tables

Introduction

The goal of this thesis is to modify projection pursuit by trading accuracy for more interpretability in the results. The two techniques examined are exploratory projection pursuit (Friedman 1987) and projection pursuit regression (Friedman and Stuetzle 1981). The former is an exploratory data analysis method which produces a description of a group of variables. The latter is a formal modeling procedure which determines the relationship of a dependent variable to a set of predictors.

The common outcome of all projection pursuit methods is a collection of vectors which define the directions of the linear projections. The remaining component is nonlinear and is summarized pictorially rather than mathematically. For example, in exploratory projection pursuit the description contains the projection linear combinations and the histograms of the projected data points. In projection pursuit regression, the model contains the linear combinations and the smooths of the dependent variable versus the projected predictors.

The statistician is faced with a collection of vectors and a nonlinear graphic representation. Given a dataset of n observations of p variables each, suppose that q projections are made. The resulting direction matrix A is q X p, each row corresponding to a projection. The statistician must try to understand and explain these vectors, both singly and as a group, in the context of the original p variables and in relation to the nonlinear components. The purpose of this thesis is to illustrate a method for trading some of the accuracy in the description or

model in return for more interpretability or simplicity in the matrix A. The object is to retain the versatility and flexibility of this promising technique while increasing the clarity of the results.

In this dissertation, interpretability is used in a similar yet broader sense than parsimony, which may be thought of as a special case. The principle of parsimony is that as few parameters as possible should be used in a description or model. Tukey stated the concept in 1961 as 'It may pay not to try to describe in the analysis the complexities that are really present in the situation.' Several methods exist which choose more parsimonious models, such as Mallows' (1973) C_p in linear regression which balances the number of parameters and prediction error. Another example is the work of Dawes (1979), who restricts the parameters in linear models based on standardized data to be 0 or ± 1, calling the resulting descriptions 'improper' linear models. His conclusion is that these models predict almost as well as ordinary linear regressions and do better than clinical intuition based on experience. Throughout this thesis, accuracy is not measured in terms of the prediction of future observations but rather refers to the goodness-of-fit of the model or description to the particular data.

Parsimony, while considering solely the number of parameters in a model, shares the general philosophical goals of interpretability. These goals are to produce results which are

(i) easier to understand.

(ii) easier to compare.

(iii) easier to remember.

(iv) easier to explain.

The adjective 'simple' is used interchangeably with 'interpretable' as is 'complex' with 'uninterpretable' throughout this thesis. However, the new term interpretability is included in part to distinguish this notion from that of simplicity which receives extensive treatment in the philosophical and Bayesian literature.

The quantification of interpretability is a difficult problem. The concept is not easy to define or measure. As Sober (1975) comments, 'the diversity of our intuitions about simplicity ... presents a veritable chaos of opinion.' Fortunately, some situations are easier than others. In particular, linear combinations as produced by projection pursuit and many other data analysis methods lend themselves readily to the development of an interpretability index. This mathematical index can serve as a 'cognostic' (Tukey 1983), or a diagnostic which can be interpreted by a computer rather than a human, in an automatic search for more interpretable results.

Exploratory projection pursuit is considered first in Chapter 1. A short review of the original method motivates the simplifying modification. The algorithmic approach chosen is supported versus alternative strategies. Various interpretability indices are developed. The modification algorithm to be employed requires changes to the original. An example of the resulting interpretable exploratory projection pursuit method is examined. Projection pursuit regression is considered in a similar manner in Chapter 2. The differing goals of this second procedure compel changes in the interpretability index. The chapter concludes with an example.

Chapter 3 connects the new approach with established techniques of trading accuracy for interpretability. Extensions to other data analysis methods that might benefit from this modification are proposed. The thesis closes with a general application of the tradeoff between accuracy and interpretability. This work is an example of an approach which simplifies the complex results of a novel statistical method. The hope is that the framework described within will be used by others in similar circumstances.

Chapter 1

Interpretable Exploratory Projection Pursuit

In this chapter, interpretable exploratory projection pursuit is demonstrated. Section 1.1 presents the basic concepts and goals of the original exploratory projection pursuit technique and outlines the algorithm. An example which provides the motivation for the new approach is included. The modification is described and support for the strategy chosen is given in the next section. The new method requires that a simplicity index be defined, which is discussed in Section 1.3. Section 1.4 details the algorithm, and its application to the example is described in the final section.

1.1 The Original Exploratory Projection Pursuit Technique

Exploratory projection pursuit (Friedman 1987) is an extension and improvement of the algorithm presented by Friedman and Tukey in 1974. It is a computer intensive data analysis tool for understanding high dimensional datasets. The method helps the statistician look for structure in the data without requiring any initial assumptions, while providing the basis for future formal modeling.

1.1.1 Introduction

Classical multivariate methods such as principal components analysis or discriminant analysis can be used successfully when the data is elliptical or normal

in nature and well-described by its first few moments. Exploratory projection pursuit is designed to deal with the type of nonlinear structure these older techniques are ill-equipped to handle. The method linearly projects the data cloud onto a line (one dimensional exploratory projection pursuit) or onto a plane (two dimensional exploratory projection pursuit). By reducing the dimensionality of the problem while maintaining the same number of datapoints, the technique overcomes the 'curse of dimensionality' (Bellman 1961). This malady is due to the fact that a huge number of points is required before structure is revealed in high dimensional space.

A linear projection is chosen for two reasons (Friedman 1987). First, the projection definition is easier to understand as it consists of one or two linear combinations of the original variables. Second, a linear projection does not exaggerate structure in the data as it is a smoothed shadow of the actual datapoints.

The goal of exploratory projection pursuit is to find projections which exhibit structure. Initially, the idea was to let the statistician choose interesting views interactively by eye (McDonald 1982, Asimov 1985). because the time required to perform an exhaustive search was prohibitive, the method was automated. A mathematical projection index which measures the structure of a view is defined and the space is searched via a computer optimization. Structure is no longer measured on a multidimensional human pattern recognition scale but rather on a univariate numerical one. This simplification consequently means that numerous possible indices may be defined. Thus the scheme, while making the analysis feasible for large datasets, requires the careful choice of a projection index.

As Huber (1985) points out, many classical techniques are forms of exploratory projection pursuit for specific projection index choices. For example, consider principal components analysis. Let Y be a random vector in R^p. The

definition of the i^{th} principal component is the solution of the maximization problem

$$\max_{\beta_i} \ \text{Var}[\beta_i^T Y]$$
$$\ni \ \beta_i^T \beta_i = 1$$
$$\text{and } \beta_i^T Y \text{ is uncorrelated with}$$
$$\text{all previous principal components} \quad .$$

Thus, principal components analysis is an example of one dimensional exploratory projection pursuit with variance (Var) as the projection index.

Variance is a global or general measure of how structured a view is. In contrast, the novelty and applicability of exploratory projection pursuit lies in its ability to recognize nonlinear or local structure. As remarked previously, many definitions of structure and corresponding projection index choices exist. All present indices, however, depend on the same basic premise. The idea is that though 'interesting' is difficult to define or agree on, 'uninteresting' is clearly normality.

Friedman (1987) and Huber (1985) provide theoretical support for this choice. Effectively, the normal distribution can be explored adequately using traditional methods which explain covariance structure. Exploratory projection pursuit is attempting to address situations for which these methods are not applicable.

The statistician must choose a method by which to measure distance from this normal origin that leads the algorithm to views she holds interesting. The distance metric choice is thus based on individual preference for the type of structure to be found. Desirable computing and invariance properties which have been neglected so far in this discussion also affect the decision. These considerations in the concrete context of a particular index are discussed as the original algorithm is detailed in the next subsection.

1.1.2 The Algorithm

Two dimensional exploratory projection pursuit is more interesting and useful than one dimensional, so the former is discussed solely. The two dimensional situation also raises intriguing problems when interpretable exploratory projection pursuit is considered which need not be addressed in the one dimensional case. For the present, the goal is to find one structured plane. Actually, the data may have several interesting views or local projection index optima and the algorithm should find as many as possible. Section 1.5.3 addresses this point.

The original algorithm presented by Friedman (1987) is reviewed in the abstract version due to Huber (1985), thereby simplifying the notation. Thus, though the data consists of n observations of length p, consider first a random variable $Y \in R^p$. The goal is to find linear combinations (β_1, β_2) which

$$\begin{aligned} \max_{\beta_1,\beta_2} \; & G(\beta_1^T Y, \beta_2^T Y) \\ \ni \; & \mathrm{Var}[\beta_1^T Y] = \mathrm{Var}[\beta_2^T Y] = 1 \\ & \text{and } \mathrm{Cov}[\beta_1^T Y, \beta_2^T Y] = 0 \end{aligned} \qquad [1.1]$$

where G is the projection index which measures the structure of the projection density. The constraints on the linear combinations ensure that the structure seen in the plane is not due to covariance (Cov) effects which can be dealt with by classical methods.

Initially, the original data Y is sphered (Tukey and Tukey 1981). The sphered variable $Z \in R^p$ is defined as

$$Z \equiv D^{-\frac{1}{2}} U^T (Y - E[Y]) \qquad [1.2]$$

with U and D resulting from an eigensystem decomposition of the covariance matrix of Y. That is,

$$\begin{aligned} \Sigma &= E[(Y - E[Y])(Y - E[Y])^T] \\ &= UDU^T \end{aligned} \qquad [1.3]$$

with U the orthonormal matrix of eigenvectors of Σ, and D the diagonal matrix of associated eigenvalues. The axes in the sphered space are the principal component directions of Y. The previous optimization problem [1.1] involving Y can be translated to one involving linear combinations α_1, α_2 and projections of Z:

$$\begin{aligned} \max_{\alpha_1,\alpha_2} \; & G(\alpha_1^T Z, \alpha_2^T Z) \\ \ni \; & \alpha_1^T \alpha_1 = \alpha_2^T \alpha_2 = 1 \\ & \text{and } \alpha_1^T \alpha_2 = 0 \quad . \end{aligned} \tag{1.4}$$

The fact that the standardization constraints imposed to exclude covariance structure, which the technique is not concerned with, are now geometric conditions reduces the computational work required (Friedman 1987). Thus, all numerical calculations are performed on the sphered data Z.

In subsequent notation, the parameters of a projection index G are $(\beta_1^T Y, \beta_2^T Y)$ though actually the value of the index is calculated for the sphered data [1.2]. The sphering, however, is merely a computational convenience and is invisible to the statistician. She associates the value of the index with the visual projection in the original data space.

After the maximizing sphered data plane is found, the vectors α_1 and α_2 are translated to the unsphered original variable space via

$$\begin{aligned} \beta_1 &= U D^{-\frac{1}{2}} \alpha_1 \\ \beta_2 &= U D^{-\frac{1}{2}} \alpha_2 \quad . \end{aligned} \tag{1.5}$$

Since only the direction of the vectors matter, these combinations are usually normed as a final step.

The variance and correlation constraints on β_1 and β_2 can be written as

$$\begin{aligned} \beta_i^T \Sigma \beta_i &= 1 \qquad i = 1, 2 \\ \beta_1^T \Sigma \beta_2 &= 0 \quad . \end{aligned} \tag{1.6}$$

Thus, β_1 and β_2 are orthogonal in the covariance metric while α_1 and α_2 are orthogonal geometrically.

The optimization method used to solve the maximization problem [1.4] is a coarse search followed by an application of steepest descent, a derivative-based optimization procedure. The initial survey of the index space via a coarse stepping approach helps the algorithm avoid deception by a small local maximum. The second stage employs the derivatives of the index to make an accurate search in the vicinity of a good starting point.

The numerical optimization procedure requires that the projection index possess certain computational properties. The index should be fast and stable to compute, and must be differentiable. These criteria surface again with respect to the interpretability index defined in Section 1.3.

1.1.3 The Legendre Projection Index

Friedman's (1987) Legendre index exhibits these properties. He begins by transforming the sphered projections to a square with the definition

$$R_1 \equiv 2\Phi(\alpha_1^T Z) - 1$$

$$R_2 \equiv 2\Phi(\alpha_2^T Z) - 1$$

where Φ is the cumulative probability density function of a standard normal random variable. Under the null hypothesis that the projections are normal and uninteresting, the density $p(R_1, R_2)$ is uniform on the square $[-1, 1]$ X $[-1, 1]$. As a measure of nonnormality, he takes the integral of the squared distance from the uniform

$$G_L(\beta_1^T Y, \beta_2^T Y) \equiv \int_{-1}^{1}\int_{-1}^{1} \left[p(R_1, R_2) - 1/4\right]^2 dR_1 dR_2 \quad . \qquad [1.7]$$

He expands the density $p(R_1, R_2)$ using Legendre polynomials, which are orthogonal on the square with respect to a uniform weight function. This action, along with subsequent integration taking advantage of the orthogonality

relationships, yields an infinite sum involving expected values of the Legendre polynomials in R_1 and R_2, which are functions of the random variable Y. In application, the expansion is truncated and sample moments replace theoretical moments. Thus, instead of $E[f(Y)]$ for a function f, the sample mean over the n observations $y_1, y_2, \ldots, y_n$ is calculated.

This index has been shown to find structure in the center of distribution, rather than in the tails. Thus, it identifies projections which exhibit clustering rather than heavy-tailed densities. G_L also has the property of 'affine invariance' (Huber 1985), which means that it is invariant under scale changes and location shifts of Y. The index has this characteristic as it is based on the sphered variables Z. Since exploratory projection pursuit should not be drawn in by covariance structure, this property is desirable for any projection index.

The Legendre index is also stable and fast to compute. Research continues in the area of projection indices but so far alternatives have proved less computationally feasible. Other indices use different sets of polynomials to take advantage of different weight functions or use alternate methods of measuring distance from normality as discussed in Section 1.4.2.

In the following section, an example of the original algorithm using G_L is discussed in order to provide the motivation for the proposed modification. After considering how to modify the method in order to make the results more interpretable, the projection index clearly needs a new theoretical property which G_L does not have. Consequently, a new index is defined in Section 1.4.2.

1.1.4 The Automobile Example

The automobile dataset (Friedman 1987) consists of ten variables collected on 392 car models and reported in 'Consumer Reports' from 1972 to 1982:

Y_1 : gallons per mile (fuel inefficiency)
Y_2 : number of cylinders in engine
Y_3 : size of engine (cubic inches)
Y_4 : engine power (horse power)
Y_5 : automobile weight
Y_6 : acceleration (time from 0 to 60 mph)
Y_7 : model year
Y_8 : American $(0,1)$
Y_9 : European $(0,1)$
Y_{10} : Japanese $(0,1)$

The second variable has 5 possible values while the last three are binary, indicating the car's country of origin. As Friedman suggests, these variables are gaussianized, which means the discrete values are replaced by normal scores after any repeated observations are randomly ordered. The object is to ensure that their discrete marginals do not overly bias the search for structure.

All the variables are standardized to have zero mean and unit variance before the analysis. In addition, the linear combinations which define the maximizing plane (β_1, β_2) are normed to length one. As a result, the coefficient for a variable in a combination represents its relative importance.

The definition of the solution plane is

$$\beta_1 = (-0.21, -0.03, -0.91,\ 0.16,\ \ 0.30, -0.05, -0.01,\ \ 0.03,\ \ 0.00, -0.02)^T$$

$$\beta_2 = (-0.75, -0.13,\ \ 0.43,\ \ 0.45, -0.07,\ \ 0.04, -0.15, -0.03,\ \ 0.02, -0.01)^T$$

That is, the horizontal coordinate of each observation is -0.21 times fuel inefficiency (standardized) -0.03 times the number of cylinders (gaussianized and standardized) and so on. The scatterplot of the points is shown in Figure 1.1.

In the projection scatterplot, the combinations (β_1, β_2) are the orthogonal horizontal and vertical axes and each observation is plotted as $(\beta_1^T Y, \beta_2^T Y)$. These

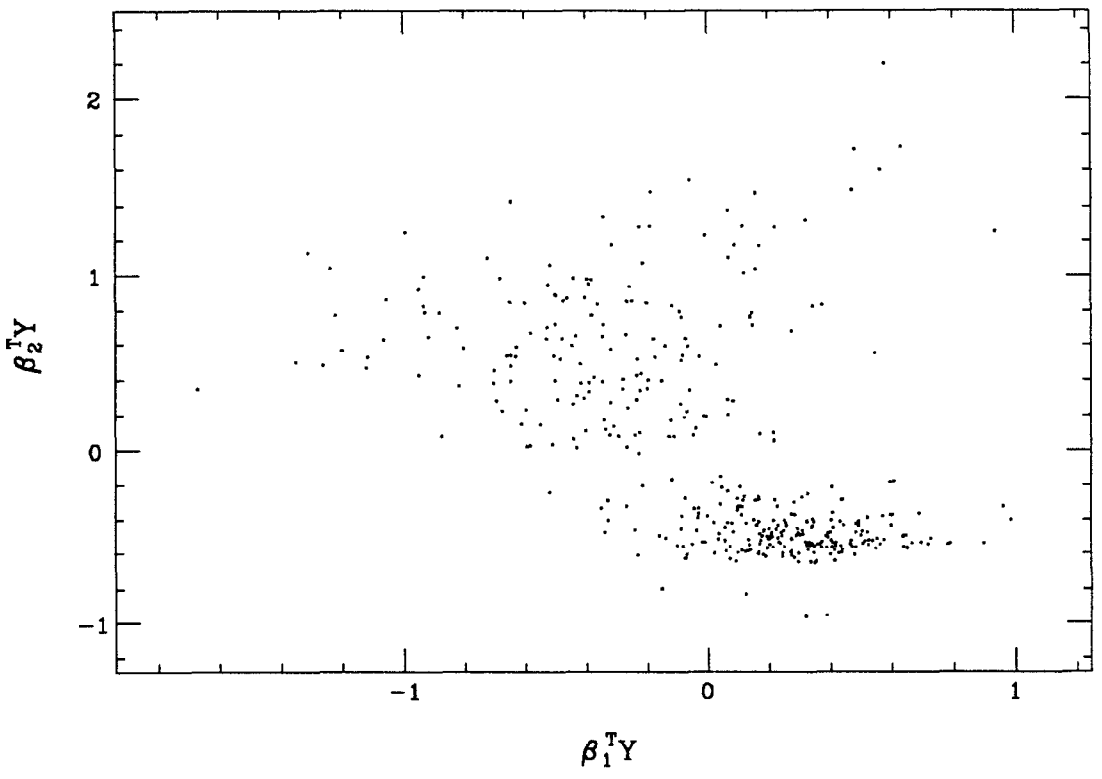


Fig. 1.1 Most structured projection scatterplot of the automobile data according to the Legendre index.

vectors are orthogonal in the covariance metric due to the constraint [1.6]. However, in the usual reference system, the orthogonal axes correspond to the variables. For example, the x axis is Y_1, the y axis is Y_2, the z axis is Y_3, and so on. The combinations are not orthogonal in this reference frame. This departure from common graphic convention is discussed further in Section 1.4.3.

The first step is to look at the structure of the projected points. In this case, the points are clustered along the vertical axis at low values and straggle out to the upper right corner with a few outliers to the left. The obvious concern is whether this structure actually exists in the data or is due to sampling fluctuation. The value of the index G_L for this particular view is 0.35 and the question is whether this value is significant. Friedman (1987) approximates the answer to this question by generating values of the index for the same number

of observations and dimensions (n and p) under the null hypothesis of normality. Comparison of the observed value with these generated ones gives an idea of how unusual the former is. Sun (1989) delves deeper into the problem, providing an analytical approximation for a critical value given the data size and chosen significance level. The structure found in this example is significant.

Given the clustering exhibited along the vertical axis, the second step is to attempt to interpret the linear combinations which define the projection plane. Though by projecting the data from ten to two dimensions, exploratory projection pursuit has reduced the visual dimension of the problem, the linear projections must still be interpreted in the original number of variables. The structure is represented by a set of points in two dimensions but understanding what these points represent in terms of the original data requires considering all ten variables.

An infinite number of pairs of vectors (β_1, β_2) exist which satisfy the constraints [1.6] and define the most structured plane. In other words, the two vectors can be spun around the origin in the plane rigidly via an orthogonal rotation and they still satisfy the constraints and maintain the structure found. The orientation of the scatterplot is inconsequential to its visual representation of the structure.

These facts lead to the principle that a plane should be defined in the simplest or most interpretable way possible. Given that the data is standardized and the linear combinations have length one, the coefficients represent the individual contribution of each variable to the combination. Friedman (1987) attempts to find the simplest representation of the plane by spinning the vectors until the variance of the squared coefficients of the vertical coordinate are maximized. This action forces the coefficients of the second combination β_2 to differ as much as possible. As a consequence, variables are forced out of the combination. This more interpretable and parsimonious vector has fewer variables in it. Such a vector is easier to understand, compare, remember and explain.

The goal of the present work is to expand this approach. A precise definition of interpretability is discussed in Section 1.3. Criteria which involve both combinations are considered. More importantly, not only is rotation of the vectors in the plane allowed but also the solution plane may be rocked in p dimensions slightly away from the most structured plane. The resulting loss in structure is acceptable only if the gain in interpretability is deemed sufficient.

1.2 The Interpretable Exploratory Projection Pursuit Approach

As shown in the preceding example, exploratory projection pursuit produces the scatterplot $(\beta_1^T Y, \beta_2^T Y)$, the value of the index $G_L(\beta_1^T Y, \beta_2^T Y)$, and the linear combinations (β_1, β_2). The scatterplot's nonlinear structure may be visually assessed but attaching much meaning to the actual numerical value of the index is difficult. As remarked earlier, the linear combinations must still be interpreted in terms of the original number of variables. In fact, a mental attempt to understand these vectors may involve 'rounding them by eye' and dropping variables which have 'too small' coefficients.

The question to be considered is whether the linear combinations can be made more interpretable without losing too much observed structure in the projection. The idea of the present modification is to trade some of the structure found in the scatterplot in return for more comprehensibility or parsimony in (β_1, β_2).

1.2.1 A Combinatorial Strategy

If initially interpretability is linked with parsimony, a combinatorial method of variable selection might be considered. The analogous approach in linear regression is all subsets regression. A similar idea, principal variables, has been applied to principal components (Krzanowski 1987, McCabe 1984). This method restricts the linear transformation matrix to a specific form, for example each row has two ones and all other entries zero. The result is a variable selection method with each component formed from two of the original variables. Both variable

selection methods, all subsets and principal variables, are discrete in nature and only consider whether a variable is in or out of the model.

Applying a combinatorial strategy to exploratory projection pursuit results in the following number of solutions, each of which consists of a pair of variable subsets. Given p variables and the fact that each variable is either in or out of a combination, 2^p possible single subsets of variables exist. The combinations are symmetric, meaning that the (β_1, β_2) plane is that the same as the (β_2, β_1) plane. Thus, the number of pairs of subsets with unequal members is $\binom{2^p}{2}$. However, this count does not include the 2^p pairs with both subsets identical, which are permissible solutions. It does include $2^p - p$ degenerate pairs in which one subset is empty, or both subsets consist of the same single variable. These degenerate pairs do not define planes. The total number of solutions is

$$\binom{2^p}{p} + 2^p - 2^p - p = 2^{2p-1} - 2^{p-1} - p \quad . \qquad [1.8]$$

Some planes counted may not be permissible due to the correlation constraint as discussed in Section 1.4.3, so the actual count may be slightly less. However, the principal term remains of the order of 2^{2p-1}. The number of subsets grows exponentially with p and for each subset the optimization must be completely redone. Unlike all subsets regression, no smart search methods exist for eliminating poor contenders which do not produce interesting projections. Due to the time required to produce a single exploratory projection pursuit solution, this combinatorial approach is not feasible.

1.2.2 A Numerical Optimization Strategy

Given that a numerical optimization is already being done, an approach is to consider whether a variable selection or interpretability criterion can be included in the optimization. The objective function used is

$$\max_{\beta_1,\beta_2} \; (1-\lambda)\frac{G(\beta_1^T Y, \beta_2^T Y)}{\max G} \; + \; \lambda\, S(\beta_1, \beta_2) \qquad [1.9]$$

for $\lambda \in [0,1]$. This function is the weighted sum of the projection index G contribution and an interpretability index S contribution, indexed by the interpretability parameter λ. The interpretability or simplicity index S is defined to have values $\in [0,1]$. Analogously, the value of the projection index is divided by its maximum possible value in order that its contribution is also $\in [0,1]$.

The interpretable exploratory projection pursuit algorithm is applied in an iterative manner. First, find the best plane using the original exploratory projection pursuit algorithm. Set max G equal to the index value for this most structured plane. For a succession of values $(\lambda_1, \lambda_2, \ldots, \lambda_I)$, such as $(0.1, 0.2, \ldots, 1.0)$, solve [1.9] with $\lambda = \lambda_i$. In each case, use the previous λ_{i-1} solution as a starting point.

One way to envision the procedure is to imagine beginning at the most structured solution and being equipped with an interpretability dial. As the dial is turned, the value of λ increases as the weight on simplicity, or the cost of complexity, is increased. The plane rocks smoothly away from the initial solution. If the projection index is relatively flat near the maximum (Friedman 1987), the loss in structure is gradual. When to stop turning the dial is discussed in Section 1.5.3.

The additive functional form choice for the objective function [1.9] is made by drawing a parallel with roughness penalty methods for curve-fitting (Silverman 1984). In that context, the problem is a minimization. The first term is a goodness-of-fit criterion such as the squared distance between the observed and fitted values, and the second is a measure of the roughness of the curve such as the integrated square of the curve's second derivative. As the fit improves, the curve becomes more rough. The negative of roughness, or smoothness, is comparable to interpretability.

An alternate idea is to think of solving a series of constrained maximization subproblems

$$\begin{aligned} \max_{\beta_1,\beta_2} \quad & G(\beta_1^T Y, \beta_2^T Y) \\ \ni \quad & S(\beta_1, \beta_2) \geq c_i \end{aligned} \qquad [1.10]$$

for values of c_i such as $(0.0, 0.1, \ldots, 1.0)$. This problem may be rewritten as an unconstrained maximization using the method of Lagrangian multipliers. Reinsch (1967) describes the relationship between [1.9] and [1.10] and consequently the relationship between c_i and λ_i.

The amount of work required by the functional approach [1.9] is linear in I. The computational savings for this numerical method versus a combinatorial one can be substantial. The inner loop, namely finding an exploratory projection pursuit solution is the same for either. The outer loop, however, is reduced from [1.8] to I.

1.3 The Interpretability Index

The interpretability index S measures the simplicity of the pair of vectors (β_1, β_2). It has a minimum value of zero at the least simple pair and a maximum value of one at the most simple. Like the projection index G, it needs to be differentiable and fast to compute.

1.3.1 Factor Analysis Background

For two dimensional exploratory projection pursuit, the object is simplify a 2 X p matrix

$$\begin{pmatrix} \beta_1^T \\ \beta_2^T \end{pmatrix} .$$

Consider a general q X p matrix Ω with entries ω_{ij}, corresponding to q dimensional exploratory projection pursuit. What characteristics does an interpretable matrix have? When is one matrix more simple than another? Researchers have

considered such questions with respect to factor loading matrices. The goal in factor analysis is to explain the correlation structure in the variables via the factor model. The solution is not unique and the factor matrix often is rotated to make it more interpretable. Comments are made regarding the geometric difference between this rotation and the interpretable rocking of the most structured plane in Section 1.4.4. Though the two situations are different, the philosophical goal is the same and so factor analysis rotation research is used as a starting point in the development of a simplicity index S.

Intuitively, the interpretability of a matrix may be thought of in two ways. 'Local' interpretability measures how simple a combination or row is individually. In a general and discrete sense, the more zeros a vector has, the more interpretable it is as less variables are involved. 'Global' interpretability measures how simple a collection of vectors is. Given that the vectors are defining a plane and should not collapse on each other, a simple set of vectors is one in which each row clearly contains its own small set of variables and has zeros elsewhere.

Thurstone (1935) advocated 'simple structure' in the factor loading matrix, defined by a list of desirable properties which were discrete in nature. For example, each combination (row) should have at least one zero and for each pair of rows, only a few columns should have nonzero entries in both rows. In summary, his requirements correspond to a matrix which involves, or has nonzero entries for, only a subset of variables (columns). Combinations should not overlap too much or have nonzero coefficients for the same variables and those that do should clearly divide into subgroups.

These discrete notions of interpretability must be translated into a continuous measure which is tractable for computer optimization. Local simplicity for a single vector is discussed first and the results are extended to a set of two vectors.

1.3.2 The Varimax Index For a Single Vector

Consider a single vector $\omega = (\omega_1, \omega_2, \ldots, \omega_p)^T$. In a discrete sense, interpretability translates into as few variables in the combination or as many zero entries as possible. The goal is to smooth the discrete count interpretability index

$$D(\omega) \equiv \sum_{i=1}^{p} I\{\omega_i = 0\} \qquad [1.11]$$

where $I\{\cdot\}$ is the indicator function.

Since the exploratory projection pursuit linear combinations represent directions and are usually normed as a final step, the index should involve the normed coefficients. In addition, in order to smooth the notion that a variable is in or out of the vector, the general unevenness of the coefficients should be measured. The sign of the coefficients is inconsequential. In conclusion, the index should measure the relative mass of the coefficients irrespective of sign and thus should involve the normed squared quantities

$$\frac{\omega_i^2}{\omega^T\omega} \qquad i = 1, \ldots, p \quad .$$

In the 1950's, several factor analysis researchers arrived separately at the same criterion (Gorsuch 1983, Harman 1976) which is known as 'varimax' and is the variance of the normed squared coefficients. This is the criterion which Friedman (1987) used as discussed in Section 1.1.3. The corresponding interpretability index is denoted by S_v and is defined as

$$S_v(\omega) \equiv \frac{p}{p-1} \sum_{i=1}^{p} \left(\frac{\omega_i^2}{\omega^T\omega} - \frac{1}{p} \right)^2 \quad . \qquad [1.12]$$

The leading constant is added to make the index value be $\in [0, 1]$. Fig. 1.2 shows the value of the varimax index for a linear combination ω in two dimensions

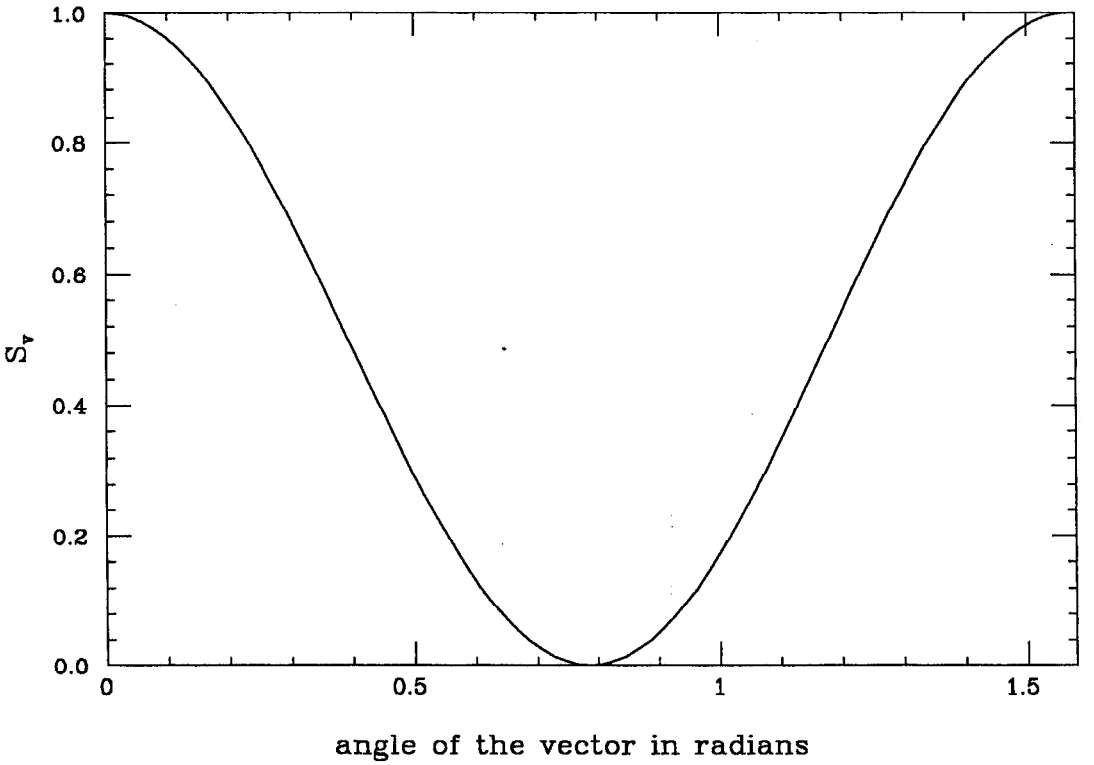


Fig. 1.2 Varimax interpretability index for $q = 1$, $p = 2$. The value of the index for a linear combination ω in two dimensions is plotted versus the angle of the direction in radians over the range $[0, \pi]$.

$(p = 2)$ versus the angle of the direction $\arctan(\omega_2/\omega_1)$ of the linear combination ω.

Fig. 1.3 shows the varimax index for vectors ω of length one in three dimensions $(p = 3)$. Only vectors with all positive components are plotted due to symmetry. Fig. 1.3 shows the value of the index as the vertical coordinate versus the values of (ω_1, ω_2). The value of ω_3 is known and does not need to be graphed. Fig. 1.4 shows contours for the surface in Fig. 1.3. These contours are just the curves of points ω which satisfy the equation formed when the left side of [1.12] is set equal to a constant value. As the value of the interpretability index is increased, the contours move away from $(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$ toward the three points $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$ and $e_3 = (0, 0, 1)$. The centerpoint is

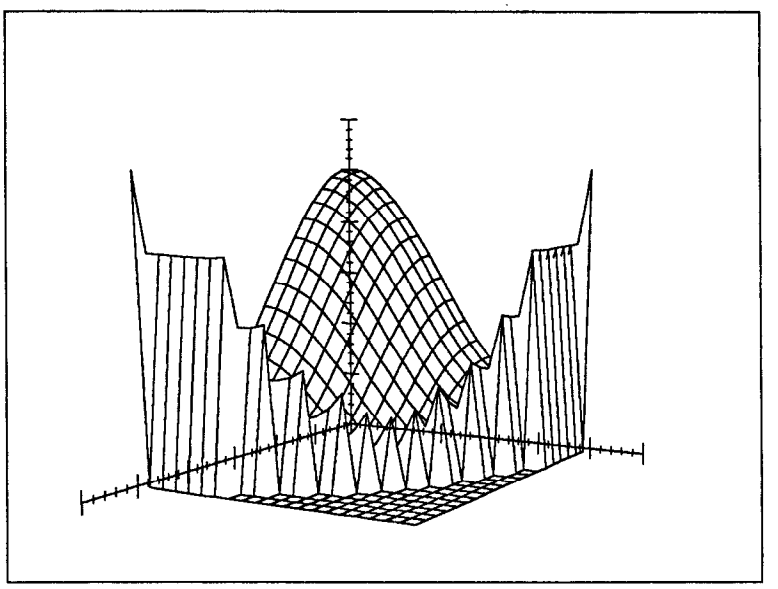


Fig. 1.3 Varimax interpretability index for $q = 1$, $p = 3$. The surface of the index is plotted as the vertical coordinate versus the first two coordinates (ω_1, ω_2) of vectors of length one in the first quadrant.

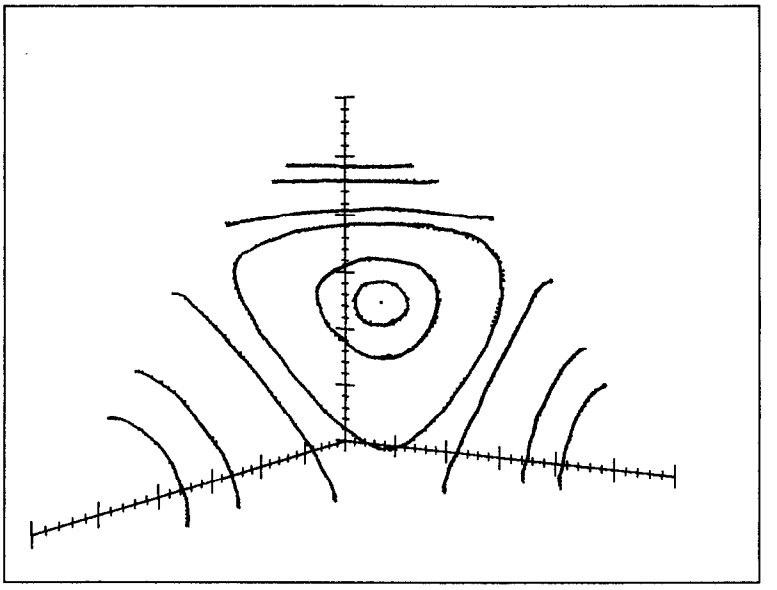


Fig. 1.4 Varimax interpretability index contours for $q = 1$, $p = 3$. The axes are the components $(\omega_1, \omega_2, \omega_3)$. The contours, from the center point outward, are those points which have varimax values $S_v(\omega)$ =(0.0, 0.01, 0.05, 0.2, 0.3, 0.6, 0.8).

the contour for $S_v(\omega) = 0.0$ and the next three joined curves are contours for $S_v(\omega) = 0.01, 0.05, 0.2$. The next three sets of lines going outward toward the e_i's are contours corresponding to $S_v(\omega) = 0.3, 0.6, 0.8$.

Since ω is normed, the varimax criterion S_v is equivalent to the 'quartimax' criterion, which derives its name from the fact it involves the sum of the fourth powers of the coefficients. S_v is also the coefficient of variation squared of the squared vector components.

1.3.3 The Entropy Index For a Single Vector

The vector of normed squared coefficients has length one and all entries are positive, similar to a multinomial probability vector. The negative entropy of a set of probabilities measures how nonuniform the distribution is (Rényi 1961). If a vector is more simple, the more uneven or distinguishable its entries are from one another. Thus a second possible interpretability index is the negative entropy of the normed squared coefficients or

$$S_e(\omega) \equiv 1 + \frac{1}{\ln p} \sum_{i=1}^{p} \frac{\omega_i^2}{\omega^T \omega} \ln \frac{\omega_i^2}{\omega^T \omega} \quad .$$

The usual entropy measure is slightly altered to have values $\in [0, 1]$. The two simplicity measures S_v and S_e share four common properties.

Property 1. Both are *maximized* when

$$\frac{\omega}{\sqrt{\omega^T \omega}} = \pm e_j \qquad j = 1, \ldots, p$$

where the $e_j, j = 1, \ldots, p$ are the unit axis vectors. Thus, the maximum value occurs when only one variable is in the combination.

Property 2. Both are *minimized* when

$$\frac{\omega}{\sqrt{\omega^T\omega}} = (\pm\frac{1}{\sqrt{p}}, \pm\frac{1}{\sqrt{p}}, \ldots, \pm\frac{1}{\sqrt{p}})$$

or when the projection is an equally weighted average. The argument could be made that an equally weighted average is in fact simple. However, in terms of deciding which variable most clearly affects the projection, it is the most difficult to interpret.

Property 3. Both are *symmetric* in the coefficients ω_i. No variable counts more than any other.

Property 4. Both are *strictly Schur-convex* as defined below. The following explanation follows Marshall and Olkin (1979).

Definition. Let $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_p)$ and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)$ be any two vectors $\in R^p$. Let $\zeta_{[1]} \geq \zeta_{[2]} \geq \ldots \geq \zeta_{[p]}$ and $\gamma_{[1]} \geq \gamma_{[2]} \geq \ldots \geq \gamma_{[p]}$ denote their components in decreasing order. The vector ζ *majorizes* γ ($\zeta \succ \gamma$) if

$$\sum_{i=1}^{j} \zeta_{[i]} \geq \sum_{i=1}^{j} \gamma_{[i]} \qquad j = 1, \ldots, p-1$$
$$\sum_{i=1}^{p} \zeta_{[i]} = \sum_{i=1}^{p} \gamma_{[i]} \quad .$$

The above definition holds $\iff \gamma = \zeta P$ where P is a doubly stochastic matrix, that is P has nonnegative entries, and column and row sums of one. In other words, if γ is a smoothed or averaged version of ζ, it is majorized by ζ. An example of a set of majorizing vectors is

$$(1, 0, \ldots, 0) \succ (\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0) \succ \ldots \succ (\frac{1}{p-1}, \ldots, \frac{1}{p-1}, 0) \succ (\frac{1}{p}, \ldots, \frac{1}{p}) \quad .$$

Definition. A function $f : R^p \mapsto R$ is *strictly Schur-convex* if

$$\zeta \succ \gamma \Longrightarrow f(\zeta) \geq f(\gamma)$$

with strict inequality if γ is not a permutation of ζ.

This type of convexity is an extension of the usual idea of Jensen's Inequality. Basically, if a vector ζ is more spread out or uneven than γ, then $S(\zeta) > S(\gamma)$. This intuitive idea of interpretability now has an explicit mathematical meaning. The two indices S_v and S_e rank all majorizable vectors in the same order. Using the theory of Schur-convexity, a general class of simplicity indices could be defined.

1.3.4 The Distance Index For a Single Vector

Besides the variance interpretation, S_v measures the squared distance from the normed squared vector to the point $(\frac{1}{p}, \frac{1}{p}, \ldots, \frac{1}{p})$, which might thus be called the least simple or most complex point. Let the notation for the Euclidean norm of any vector ζ be

$$\|\zeta\| \equiv \left(\zeta^T\zeta\right)^{\frac{1}{2}} = \left(\zeta_1^2 + \zeta_2^2 + \ldots \zeta_p^2\right)^{\frac{1}{2}} \quad .$$

If ν_ω is defined to be the squared and normed version of ω,

$$\nu_\omega \equiv (\frac{\omega_1^2}{\omega^T\omega}, \frac{\omega_2^2}{\omega^T\omega}, \ldots, \frac{\omega_p^2}{\omega^T\omega}) \quad , \qquad [1.13]$$

and ν_c is the most complex point, then

$$S_v(\omega) = \frac{p}{p-1}\|\nu_\omega - \nu_c\|^2 \quad .$$

This index can be generalized. In contrast to having one most complex point, an alternate index can be defined by considering a set $V = \{\nu_1, \ldots, \nu_J\}$ of m simple points. This set must have the properties

$$\begin{aligned} &\nu_{ji} > 0 \qquad j = 1, \ldots, J \qquad i = 1, \ldots, p \\ &\sum_{i=1}^{p} \nu_{ji} = 1 \qquad j = 1, \ldots, J \quad . \end{aligned} \tag{1.14}$$

The ν_i's are the squares of vectors on the unit sphere $\in R^p$ as are ν_ω and ν_c. An example is $V = \{e_j : j = 1, \ldots, p\}$, with $J = p$. In the event that this index is used with exploratory projection pursuit, the statistician could define her own set of simple points rather than be restricted to the choice of ν_c used in S_v.

This distance would be large when ω is not simple, so the interpretability index should involve the negative of it. An example is

$$S_d(\omega) = 1 - c \min_{\nu_j : j=1,\ldots,J} \|\nu_\omega - \nu_j\|^2 \quad . \tag{1.15}$$

The constant c is calculated so that the values are $\in [0, 1]$. Any distance norm can be used and an average or total distance could replace the minimum.

If V is chosen to be the e_j's and the minimum Euclidean norm is used, the distance index becomes

$$S_d^*(\omega) \equiv 1 - \frac{p}{p-1}\left[\sum_{i=1}^{p}\left(\frac{\omega_i^2}{\omega^T\omega}\right)^2 - 2\frac{\omega_k^2}{\omega^T\omega} + 1\right]$$

where k corresponds to the maximum $\mid \omega_i \mid$, $i = 1, \ldots, p$. The minimization does not need to be done at each step, though the largest absolute coefficient in the vector must be found. Analogous results are obtainable for similar choices of V such as all permutations of two $\frac{1}{2}$ entries and $p - 2$ zeros $((\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0), (\frac{1}{2}, 0, \frac{1}{2}, 0, \ldots, 0), \ldots)$ corresponding to simple solutions of two variables each.

Since the e_j's maximize S_v and are the simple points associated with S_d^*, the relationship between the two indices proves interesting. Algebra reveals

$$S_d^*(\omega) = -S_v(\omega) + \frac{2}{p-1}\left(p\frac{\omega_k^2}{\omega^T\omega} - 1\right) \quad .$$

The minimum of the second term occurs when

$$\frac{\omega_k^2}{\omega^T\omega} = \frac{1}{p}$$

or all coefficients are equal and $S_d^*(\omega) = S_v(\omega) = 0$. The maximum occurs when

$$\frac{\omega_k^2}{\omega^T\omega} = 1$$

and $S_d^*(\omega) = S_v(\omega) = 1$. The relationship between the interim values varies.

The difficulty with the distance index S_d^* is that its derivatives are not smooth and present problems when a derivative-based optimization procedure is used. Given that the entropy index S_e and the varimax index S_v share common properties and the latter is easier to deal with computationally, the varimax approach is generalized to two dimensions.

1.3.5 The Varimax Index For Two Vectors

The varimax index S_v can be extended to measure the simplicity of a set of q vectors $\omega_j = (\omega_{j1}, \omega_{j2}, \ldots, \omega_{jp})$, $j = 1, \ldots, q$. In the following, the varimax index for one combination [1.12] is called S_1. In order to force orthogonality between the squared normed vectors, the variance is taken across the vectors and summed over the variables to produce

$$S_v(\omega_1, \ldots, \omega_q) = \sum_{i=1}^{p}\sum_{j=1}^{q}\left(\frac{\omega_{ji}^2}{\omega_j^T\omega_j} - \frac{\sum_{j=1}^{q}\frac{\omega_{ji}^2}{\omega_j^T\omega_j}}{q}\right)^2 \quad . \qquad [1.16]$$

If the sums were reversed and the variance was taken across the variables and summed over the vectors, the unnormed index would equal

$$S_1(\omega_1) + S_1(\omega_2) + \cdots + S_1(\omega_q) \qquad [1.17]$$

with each element the one dimensional simplicity [1.12] of the corresponding vector. The previous approach [1.16] results in a cross-product term.

For two dimensional exploratory projection pursuit, $q = 2$ and the index, with appropriate norming, reduces to

$$S_v(\omega_1, \omega_2) = \frac{1}{2p}\Big[(p-1)S_1(\omega_1) + (p-1)S_1(\omega_2) + 2\Big] - \sum_{i=1}^{p} \frac{\omega_{1i}^2}{\omega_1^T\omega_1}\frac{\omega_{2i}^2}{\omega_2^T\omega_2} \quad , \quad [1.18]$$

with appropriate norming. The first term measures the local simplicities of the vectors while the second is a cross-product term measuring the orthogonality of the two normed squared vectors. This cross-product forces the vectors to 'squared orthogonality', so that different groups of variables appear in each vector. S_v is maximized when the normed squared versions of ω_1 and ω_2 are e_k and e_l, $k \neq l$ and minimized when both are equal to $(\frac{1}{p}, \frac{1}{p}, \ldots, \frac{1}{p})$.

1.4 The Algorithm

In this section, the algorithm used to solve the interpretable exploratory projection pursuit problem [1.9] with interpretability index S_v is discussed. The general approach of the original exploratory projection pursuit algorithm is followed but two changes are required, as described in the first three subsections.

1.4.1 Rotational Invariance of the Projection Index

As discussed in Section 1.1.4, the orientation of a scatterplot is immaterial and a particular observed structure should have the same value of the projection index G no matter from which direction it is viewed. An alternative way of stating this is that the projection index should be a function of the plane, not of the way the plane is represented (Jones 1983). The interpretable projection pursuit algorithm should always describe a plane in the simplest way possible as measured by S_v. Given these two facts, any projection index used by the algorithm should have the property of *rotational invariance.*

Definition. A projection index G is *rotationally invariant* if

$$G(\beta_1^T Y, \beta_2^T Y) = G(\eta_1^T Y, \eta_2^T Y)$$

where (β_1, β_2) is (η_1, η_2) rotated through θ or

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = Q \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$$

with Q the orthogonal rotation matrix associated with the angle θ or

$$Q = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad . \qquad [1.19]$$

Rotational invariance should not be confused with affine invariance. As remarked in Section 1.1.3, the latter property is a welcome byproduct of sphering.

Friedman's Legendre index G_L [1.7] is not rotationally invariant. This property is not required for his algorithm. As described in Section 1.1.4, he simplifies the axes after finding the most structured plane by maximizing the single vector varimax index S_1 for the second combination. He does not allow the solution to rock away from the most structured plane in order to increase interpretability.

Recall that the first step in calculating G_L is to transform the marginals of the projected sphered data to a square. Under the null hypothesis the projection is $N(0, I)$, which means the projection scatterplot looks like a disk and the orientation of the axes is immaterial. Intuitively, however, if the null hypothesis is not true, the placement of the axes affects the marginals and thus the index value.

Empirically, this lack of rotational invariance is evident. In the automobile example from Section 1.1.4, the most structured plane was found to have $G_L(\beta_1^T Y, \beta_2^T Y) = 0.35$. The projection index values for the scatterplot as the axes are spun through a series of angles are shown in Table 1.1. A new index, which seeks to maintain the computational properties and to find the same structure as G_L, is developed in the next subsection.

θ	0.0	$\frac{\pi}{20}$	$\frac{\pi}{10}$	$\frac{3\pi}{20}$	$\frac{\pi}{5}$	$\frac{\pi}{4}$	$\frac{3\pi}{10}$	$\frac{7\pi}{20}$	$\frac{2\pi}{5}$	$\frac{9\pi}{20}$	$\frac{\pi}{2}$
$G(\beta_1^T Y, \beta_2^T Y)$	0.35	0.36	0.34	0.32	0.30	0.29	0.29	0.29	0.30	0.32	0.35

Table 1.1 Most structured Legendre plane index values for the automobile data. Values of the Legendre index for different orientations of the axes in the most structured plane are given.

1.4.2 The Fourier Projection Index

The Legendre index G_L is based on the Cartesian coordinates of the projected sphered data $(\alpha_1^T Z, \alpha_2^T Z)$. The index is based on knowing the distribution of these coordinates under the null hypothesis of normality. Polar coordinates are the natural alternative given that rotational invariance is desired. The distribution of these coordinates is also known under the null hypothesis. The expansion of the density via orthogonal polynomials is done in a manner similar to that of G_L.

The polar coordinates of a projected sphered point $(\alpha_1^T Z, \alpha_2^T Z)$ are

$$R \equiv (\alpha_1^T Z)^2 + (\alpha_2^T Z)^2$$
$$\Theta \equiv \arctan\left(\frac{\alpha_2^T Z}{\alpha_1^T Z}\right) \quad .$$

Actually, the usual polar coordinate definition involves the radius of the point rather than its square but the notation is easier given the above definition of R.

Under the null hypothesis that the projection is $N(0, I)$, R and Θ are independent and

$$R \sim \text{Exp}\left(\frac{1}{2}\right)$$
$$\Theta \sim \text{Unif}[-\pi, \pi] \quad .$$

The proposed index is the integral of the squared distance between the density of (R, Θ) and the null hypothesis density, which is the product of the exponential and uniform densities,

$$G(\beta_1^T Y, \beta_2^T Y) \equiv \int_{-\pi}^{\pi} \int_0^{\infty} \left[p_{R,\Theta}(u, v) - f_R(u) f_\Theta(v)\right]^2 du dv \quad . \qquad [1.20]$$

The density $p_{R,\Theta}$ is expanded as the tensor product of two sets of orthogonal polynomials chosen specifically for their weight functions and rotational properties. The weight functions must match the densities f_R and f_Θ in order to utilize the orthogonality relationships and the polynomials chosen for the Θ portion of the expansion must result in rotational invariance. By definition, R is not affected by rotation. Friedman (1987) uses Legendre polynomials for both Cartesian coordinates as his two density functions are identical (Unif[-1,1]), the Legendre weight function is Lebesgue measure, and his algorithm does not require rotational invariance. Hall (1989) considered Hermite polynomials and developed a one dimensional index. The following discussion combines aspects of the two authors' approaches. Throughout the discussion, i, j, and k are integers.

The set of polynomials for the R portion is the Laguerre polynomials which are defined on the interval $[0, \infty)$ with weight function $w(u) = e^{-u}$. The polynomials are

$$\begin{aligned} L_0(u) &= 1 \\ L_1(u) &= u - 1 \\ L_i(u) &= (u - 2i + 1)L_{i-1}(u) - (i-1)^2 L_{i-2}(u) \quad . \end{aligned} \tag{1.21}$$

The associated Laguerre functions are defined as

$$l_i(u) \equiv L_i(u)e^{-\frac{1}{2}u} \quad .$$

The orthogonality relationships between the polynomials are

$$\int_0^\infty l_i(u)l_j(u)du = \delta_{ij} \qquad i, j \geq 0 \text{ and } i \neq j$$

where δ_{ij} is the Kronecker delta function.

Any piecewise smooth function $f : R \mapsto R$ may be expanded in terms of these polynomials as

$$f(u) = \sum_{i=0}^{\infty} a_i l_i(u)$$

where the a_i are the Laguerre coefficients

$$a_i \equiv \int_0^\infty L_i(u)e^{-\frac{1}{2}u} f(u)du \quad .$$

The smoothness property of f means the function has piecewise continuous first derivatives and the Fourier series converges pointwise. If a random variable W has density f, the coefficients can be written as

$$a_i = E_f\left[l_i(W)\right] \quad .$$

The Θ portion of the index is expanded in terms of sines and cosines. Any piecewise smooth function $f : R \mapsto R$ may be written as

$$f(v) = \frac{a_0}{2} + \sum_{k=1}^{\infty} [a_k \cos(kv) + b_k \sin(kv)]$$

with pointwise convergence.

The orthogonality relationships between these trigonometric functions are

$$\int_{-\pi}^{\pi} \cos^2(kv)dv = \int_{-\pi}^{\pi} \sin^2(kv)dv = \pi \qquad k \geq 1$$
$$\int_{-\pi}^{\pi} \cos(kv)\sin(jv)dv = 0 \qquad k \geq 0 \text{ and } j \geq 1$$
$$\int_{-\pi}^{\pi} dv = 2\pi \quad .$$

The a_k and b_k are the Fourier coefficients

$$a_k \equiv \frac{1}{\pi}\int_{-\pi}^{\pi} \cos(kv)f(v)dv = \frac{1}{\pi}E_f[\cos(kW)]$$
$$b_k \equiv \frac{1}{\pi}\int_{-\pi}^{\pi} \sin(kv)f(v)dv = \frac{1}{\pi}E_f[\sin(kW)] \quad .$$

The density $p_{R,\Theta}$ can be expanded via a tensor product of the two sets of polynomials as

$$p_{R,\Theta}(u,v) = \sum_{i=0}^{\infty} l_i(u)\Big(\frac{a_{i0}}{2} + \sum_{k=1}^{\infty}[a_{ik}\cos(kv) + b_{ik}\sin(kv)]\Big) \quad .$$

The a_{ik} and b_{ik} are the coefficients defined as

$$a_{ik} \equiv \frac{1}{\pi}E_p[l_i(R)\cos(k\Theta)] \qquad i,k \geq 0$$
$$b_{ik} \equiv \frac{1}{\pi}E_p[l_i(R)\sin(k\Theta)] \qquad i \geq 0 \text{ and } k \geq 1 \quad .$$

The null distribution, which is the product of the exponential density and the uniform density over $[-\pi, \pi]$ is

$$f_R(u)f_\Theta(v) = \left(\frac{1}{2}e^{-\frac{1}{2}u}\right)\left(\frac{1}{2\pi}\right) = \frac{1}{4\pi}l_0(u) \quad .$$

The index [1.20] becomes

$$\int_{-\pi}\int_0^\infty \left[\sum_{i=0}^{\infty} l_i(u)\left(\frac{a_{i0}}{2} + \sum_{k=1}^{\infty}\left[a_{ik}\cos(kv) + b_{ik}\sin(kv)\right]\right) - \frac{1}{4\pi}l_0(u)\right]^2 dudv \quad .$$

The further condition that $p_{R,\Theta}$ is square-integrable and subsequent multiplication, integration, and use of the orthogonality relationships show that the index $G(\beta_1^T Y, \beta_2^T Y)$ equals

$$\sum_{i=1}^{\infty}(2\pi)\frac{a_{i0}^2}{4} + \sum_{i=0}^{\infty}\sum_{k=1}^{\infty}\pi(a_{ik}^2 + b_{ik}^2) - (2\pi)\left(\frac{2}{4\pi}\right)\left(\frac{a_{00}}{2}\right) + (2\pi)\left(\frac{1}{16\pi^2}\right) \quad . \quad [1.22]$$

Maximizing [1.22] is equivalent to maximizing the Fourier index defined as

$$G_F(\beta_1^T Y, \beta_2^T Y) \equiv \pi G(\beta_1^T Y, \beta_2^T Y) - \frac{1}{8} \quad .$$

The definitions of the coefficients in G_F yield

$$\begin{aligned} G_F(\beta_1^T Y, \beta_2^T Y) =& \frac{1}{2}\sum_{i=0}^{\infty} E_p^2\left[l_i(R)\right] + \sum_{i=0}^{\infty}\sum_{k=1}^{\infty}\Big(E_p^2\left[l_i(R)\cos(k\Theta)\right] + E_p^2\left[l_i(R)\sin(k\Theta)\right]\Big) \\ & - \frac{1}{2}E_p\left[l_0(R)\right] \quad . \end{aligned} \quad [1.23]$$

This index closely resembles the form of G_L in Friedman (1987). The extra $i = 0$ term in the first sum and the last subtracted term appear since the weight function is the exponential instead of Lebesgue measure as it is for Legendre polynomials. In application, each sum is truncated at the same fixed value and

the expected values are approximated by the sample moments taken over the data points. For example, $E_p^2\,[l_i(R)\cos(k\Theta)]$ is approximated by

$$\left[\frac{1}{n}\sum_{j=1}^{n} l_i(r_j)\cos(k\theta_j)\right]^2$$

where r_j and θ_j are the radius squared and angle for the projected j^{th} observation.

The Fourier index is rotationally invariant. Suppose the projected points are spun by an angle of τ. The radius squared R is unaffected by the shift so the first sum and final term in [1.23] do not change. The sine and cosine of the new angle are

$$\cos(\Theta+\tau) = \sin\Theta\cos\tau + \cos\Theta\sin\tau$$
$$\sin(\Theta+\tau) = \cos\Theta\cos\tau - \sin\Theta\sin\tau \quad .$$

Each component in the second term of [1.23] is

$$\begin{aligned} E_p^2[l_i(R)\cos(k(\Theta+\tau))] + E_p^2[l_i(R)\sin(k(\Theta+\tau))] = \\ \cos^2(k\tau)\,E_p^2[l_i(R)\sin(k\Theta)] + \sin^2(k\tau)E_p^2[l_i(R)\cos(k\Theta)] \\ +2\sin(k\tau)\,\cos(k\tau)E_p[l_i(R)\sin(k\Theta)\cos(k\Theta)] \\ +\cos^2(k\tau)\,E_p^2[l_i(R)\cos(k\Theta)] + \sin^2(k\tau)E_p^2[l_i(R)\sin(k\Theta)] \\ -2\sin(k\tau)\,\cos(k\tau)E_p[l_i(R)\sin(k\Theta)\cos(k\Theta)] \\ = E_p^2[l_i(R)\cos(k\Theta)] + E_p^2[l_i(R)\sin(k\Theta)] \end{aligned} \qquad [1.24]$$

and the index value is not affected by the rotation. The truncated version of the index also has this property as it is true for each component. Moreover, replacing the expected value by the sample mean does not affect [1.24], so the sample truncated version of the index is rotationally invariant.

Hall (1989) proposes a one dimensional Hermite function index. The Hermite weight function of $e^{-\frac{1}{2}u^2}$ helps bound his index for heavytailed projection densities. In addition, he addresses the question of how many terms to include in the truncated version of the index. Similarly, the asymptotic behavior of G_L

is being investigated at present. The Fourier index consists of the bounded sine and cosine Fourier contribution, and of the Laguerre function portion which is weighted by $e^{-\frac{1}{2}u}$. The class of densities for which this index is finite must be determined as well as the number of terms needed in the sample version.

The Fourier G_F and Legendre G_L indices are based on comparing the density of the projection with the null hypothesis density. Jones and Sibson (1987) define a rotationally invariant index based on comparing cumulants. Their index tends to equate structure with outliers while the density indices G_F and G_L tend to find clusters. A possible future rotationally invariant index could be based on Radon transforms (Donoho and Johnstone 1989).

1.4.3 Projection Axes Restriction

The algorithm searches through the possible planes with the weighted objective function [1.9] as a criterion. Every plane has a single projection index value associated with it since G_F is rotationally invariant. Ideally, each plane would have a single possible representation; the most interpretable one as measured by S_v. Unfortunately, the optimal representation of a given plane cannot be solved analytically. However, as the weight on simplicity (λ) is increased, the algorithm tends to represent each plane most simply.

In order to help the algorithm find the most interpretable representation of a plane, the constraints on the linear combinations (β_1, β_2) must be changed. Recall that in the original algorithm, the correlation constraint [1.6] is imposed on the linear combinations (β_1, β_2) which define the solution plane. This constraint translates into an orthogonality constraint for the linear combinations (α_1, α_2) which define the solution plane in the sphered data space. However, simplicity is measured for the two unsphered combinations (β_1, β_2) and is maximized when the two vectors are $(\pm e_k, \pm e_l), k \neq l$. These maximizing combinations are orthogonal in the original data space and correspond to the variable k and variable

l axes. Unless two variables are uncorrelated, the maximum simplicity cannot be achieved by a pair of uncorrelated combinations.

To ensure that the algorithm can find a maximum, given any plane defined by the pair (β_1, β_2) which satisfies the correlation constraint, the interpretability of the plane is calculated after the linear combinations have been translated to orthogonality. Unfortunately, the optimal translation is not known so it is done in the following manner. Without loss of generality, β_1 is fixed and β_2 is spun in the plane until the two vectors are orthogonal. The spinning is done by projecting β_2 onto β_1 and taking the remainder. That is, β_2 is decomposed into the sum of a component which is parallel to β_1 and a component which is orthogonal to β_1. The new β_2, which is called β_2^*, is the latter component. Mathematically,

$$\beta_2^* \equiv \beta_2 - \left(\frac{\beta_1^T \beta_2}{\beta_1^T \beta_1}\right)\beta_1 \quad . \qquad [1.25]$$

Whether $S_v(\beta_1, \beta_2^*)$ is always greater than or equal to $S_v(\beta_1, \beta_2)$ is not clear. However, as noted above, the maximum value of the index can be achieved given this translation.

As an added bonus, the two combinations (β_1, β_2^*) are orthogonal in the original variable space, which is the usual graphic reference frame. This situation is in contrast to the original exploratory projection pursuit algorithm which graphs with respect to the covariance metric as noted in Section 1.1.4. In effect, a further subjective simplification in the solution has been made as the visual representation is more interpretable.

The final solution for any particular λ value is reported as the normed, translated set of vectors

$$\left(\frac{\beta_1}{\sqrt{\beta_1^T \beta_1}}, \frac{\beta_2^*}{\sqrt{\beta_2^{*T} \beta_2^*}}\right) \quad . \qquad [1.26]$$

Throughout the rest of this thesis, the orthogonal translation is assumed and (β_1, β_2) is written for (β_1, β_2^*). As a result, whenever $S_v(\beta_1, \beta_2)$ is referred to, the actual value is $S_v(\beta_1, \beta_2^*)$ and [1.18] becomes

$$S_v(\beta_1, \beta_2) = \frac{1}{2p}[(p-1)S_1(\beta_1) + (p-1)S_1(\beta_2^*) + 2] - \sum_{i=1}^{p} \frac{\beta_{1i}^2}{\beta_1^T \beta_1} \frac{(\beta_{2i}^*)^2}{\beta_2^{*T} \beta_2^*} \quad . \quad [1.27]$$

1.4.4 Comparison With Factor Analysis

Given that factor analysis is used to motivate the interpretability index, comparison with this method is warranted. The two dimensional projection $X = (X_1, X_2)^T$ is defined as

$$X \equiv BY$$

where B is the 2 X p linear combination matrix

$$B \equiv \begin{pmatrix} \beta_1^T \\ \beta_2^T \end{pmatrix}$$

and the observed variables are $Y = (Y_1, Y_2, \ldots, Y_p)^T$. This definition is similar to the one for principal components except that in the latter case, usually all p principal components are found so that the dimension of B is p X p. In addition, as was remarked in Section 1.1.1, principal components maximize a different projection index.

The first attempt to simplify B, due to Friedman (1987) and discussed in Section 1.1.4, involved rigidly spinning the projected points (X_1, X_2) in the solution plane. The new set of points is $QX = QBY$ where Q is a two dimensional orthogonal matrix as in [1.19]. Since the rotation is rigid, it maintains the correlation constraint. Thus simplification through spinning in the plane is achieved by multiplying the linear combination matrix B by Q.

The analogous p variable factor analysis model is

$$Y = \Omega f + \epsilon \quad .$$

In this model, there are assumed to be two unknown underlying factors $f = (f_1, f_2)^T$ and ϵ is a p X 1 error vector. The factor loading matrix Ω is p X 2 and is found by seeking to explain the covariance structure of the variables $(Y_1, Y_2, \ldots, Y_p)$ given certain distributional assumptions. Due to the non-uniqueness of the model, the factors can be orthogonally rotated producing

$$Q \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \quad ,$$

without changing the explanatory power of the model. A rotation is made in order to simplify the factor-loading matrix to ΩQ^T.

Taking the transpose of the new factor-loading matrix produces $Q\Omega^T$, a 2 X p matrix, which is comparable in dimensionality to the new exploratory projection pursuit linear combination matrix QB. Thus, spinning the linear combinations to a more interpretable solution is analogous to simplifying the factor-loading matrix in a two factor model. A transpose is taken since the factor analysis linear combinations are of the underlying factors, while the exploratory data analysis combinations are of the observed variables. This comparison is analagous to that between factor analysis and principal components analysis.

Interpretable exploratory projection pursuit involves rocking the solution plane. In this case, the linear combinations (β_1, β_2) are moved in R^p subject only to the correlation constraint. They are then transposed to orthogonality via [1.25] to further increase interpretability. The more interpretable coefficients may not be linear combinations of the original ones. Such a move is not allowed in the factor-analysis setting. The interpretable method may be thought of as a looser, less restrictive version of factor analysis rotation.

1.4.5 The Optimization Procedure

The interpretable exploratory projection pursuit objective function is

$$(1-\lambda)\frac{G_F(\beta_1^T Y, \beta_2^T Y)}{\max G_F} + \lambda S_v(\beta_1, \beta_2) \qquad [1.28]$$

where $S_v(\beta_1, \beta_2)$ is calculated after the translation [1.25]. The computer algorithm employed is similar to that of the original exploratory projection pursuit algorithm described in Section 1.1.2. The algorithm is outlined below and then comments are made on the specific steps.

0. Sphere the data [1.2].
1. Conduct a coarse search to find a starting point to solve the original problem [1.1] (or [1.28] with $\lambda = 0$).
2. Use an accurate derivative based optimization procedure to find the most structured plane.
3. Spin the solution vectors in the optimal plane to the most interpretable representation. Call this solution P_0.
4. Decide on a sequence $(\lambda_1, \lambda_2, \ldots, \lambda_I)$ of interpretability parameter values.
5. Use a derivative based optimization procedure to solve [1.28] with $\lambda = \lambda_i$ and starting plane P_{i-1}. Call the new solution plane P_i.
6. If $i = I$, **EXIT**. Otherwise, set $i = i + 1$ and **GOTO** 5.

The search for the best plane is performed in the sphered data space, as discussed in Section 1.1.2. However, an important note is that the interpretability of the plane must always be calculated in terms of the original variables. The β_i combinations, not the α_i combinations, are the ones the statistician sees. In fact, she is unaware of the sphering, which is just a computational shortcut behind the scenes.

The modification does require one important difference in the sphering. In Friedman (1987), the suggestion is to consider only the first q sphered variables

Z where $q < p$ and a considerable amount of the variance is explained. The dropping of the unimportant sphered variables is the same as the dropping of unimportant components in principal components analysis and reduces the computational work involved. In Step 5, the interpretability gradients are calculated for (β_1, β_2) and then translated via the inverse of [1.5],

$$\begin{aligned} \alpha_1 &= D^{\frac{1}{2}} U^T \beta_1 \\ \alpha_2 &= D^{\frac{1}{2}} U^T \beta_2 \quad , \end{aligned} \qquad [1.29]$$

to the sphered space. If the gradient components are nonzero only in the $p - q$ dropped dimensions, they become zero after translation. The derivative-based optimization procedure assumes it is at the maximum and stops, even though the maximum has not been reached. Thus, no reduction in dimension during sphering should be made.

The coarse search in Step 1 is done to ensure that the algorithm starts in the vicinity of a large local maximum. The procedure which Friedman (1987) employs is based on the axes in the sphered data space. He finds the most structured pair of axes and then takes large steps through the sphered space. Since the interpretability measure S_v is calculated for the original variable space, a feasible alternative might be to coarse step through the original rather than sphered data space. For example, the starting point could be the most structured pair of original axes. This pair of combinations is in fact the simplest possible. On the other hand, stepping evenly through the unsphered data space might not cover the data adequately as the points could be concentrated in some subspace due to covariance structure. Sphering solves this problem. In all data examples tried so far, the starting point did not have an effect on the final solution.

In Steps 2 and 5, the accurate optimization procedure used is the program NPSOL (Gill et al. 1986). This package solves nonlinear constrained optimization problems. The search direction at any step is the solution to a quadratic

programming problem. The package is employed to solve for the sphered combinations (α_1, α_2) subject to the length and orthogonality constraints [1.4]. The gradients for the projection index G_F are straightforward. The interpretability index S_v derivatives are more difficult as they involve translations from the uncorrelated to orthogonal combinations [1.25] and from the sphered to unsphered space [1.5]. These gradients are given in Appendix A.

The package NPSOL is extremely powerful. At present, work continues to design a steepest descent algorithm which maintains the constraints. However, given the complicated translations between the sphered and unsphered spaces, this problem is a difficult one.

Step 3 can be performed in two ways. The initial pair of vectors can be discretely spun in the plane to the most interpretable representation or Steps 5 and 6 can be run with λ equal to a very small value, say 0.01. This slight weight on simplicity does not overpower the desire for structure. The plane is not permitted to rock but spinning is allowed. The result is the most interpretable representation of the most structured plane. As noted previously, this spinning is similar to Friedman's (1987) simplification except that a two vector varimax interpretability index is used instead of a single vector one.

The initial value of the projection index G_F is used as max G in the denominator of the first term in [1.28]. However, the algorithm may be caught in a local maximum and as the weight on simplicity is increased, the procedure may move to a larger maximum. Thus, the contribution of the projection index term may at some time be greater than one. This is an unexpected benefit of the interpretable projection pursuit approach, both structure and interpretability have been increased.

In the examples tried, the algorithm is not very sensitive to the λ sequence choice as long as the values are not too far apart. For example, the sequence $(0.0, 0.1, \ldots, 1.0)$ produces the same solutions as $(0.0, 0.05, 0.1, \ldots, 1.0)$ but the sequence $(0.0, 0.25, 0.5, 1.0)$ does not.

Throughout the loop in Steps 5 and 6, the previous solution at λ_{i-1} is used as the starting point for the application of the algorithm with λ_i. This approach is in the spirit of rocking the solution away from the original solution. Examples have shown that the objective is fairly smooth and a large gain in simplicity is made initially in turn for a small loss in structure.

As remarked in Section 1.1.4 when the example was considered, the data Y is usually standardized before analysis. The reported combinations are [1.26]. Thus the coefficients represent the relative importance of the variables in each combination. The next section consists of the analysis of an easy example followed by a return to the automobile Example.

1.5 Examples

In this section, two examples of interpretable exploratory projection pursuit are examined. The first is an example of the one dimensional algorithm, while the second is the two dimensional algorithm applied to the automobile data analyzed in Section 1.1.3. Several implementation issues are discussed at the end of the section.

1.5.1 An Easy Example

The simulated data in this example consists of 200 (n) points in two (p) dimensions. The horizontal and vertical coordinates are independent and normally distributed with means of zero and variances of one and nine respectively. The data is spun about the origin through an angle of thirty degrees. Three is subtracted from each coordinate of the first fifty points and three is added to each coordinate of the remaining points. The data appear in Fig. 1.5.

Since the data is only in two dimensions and can be viewed completely in a scatterplot, using interpretable exploratory projection pursuit in this instance is quite contrived. However, this exercise is useful in helping understand the way the procedure works and its outcome.

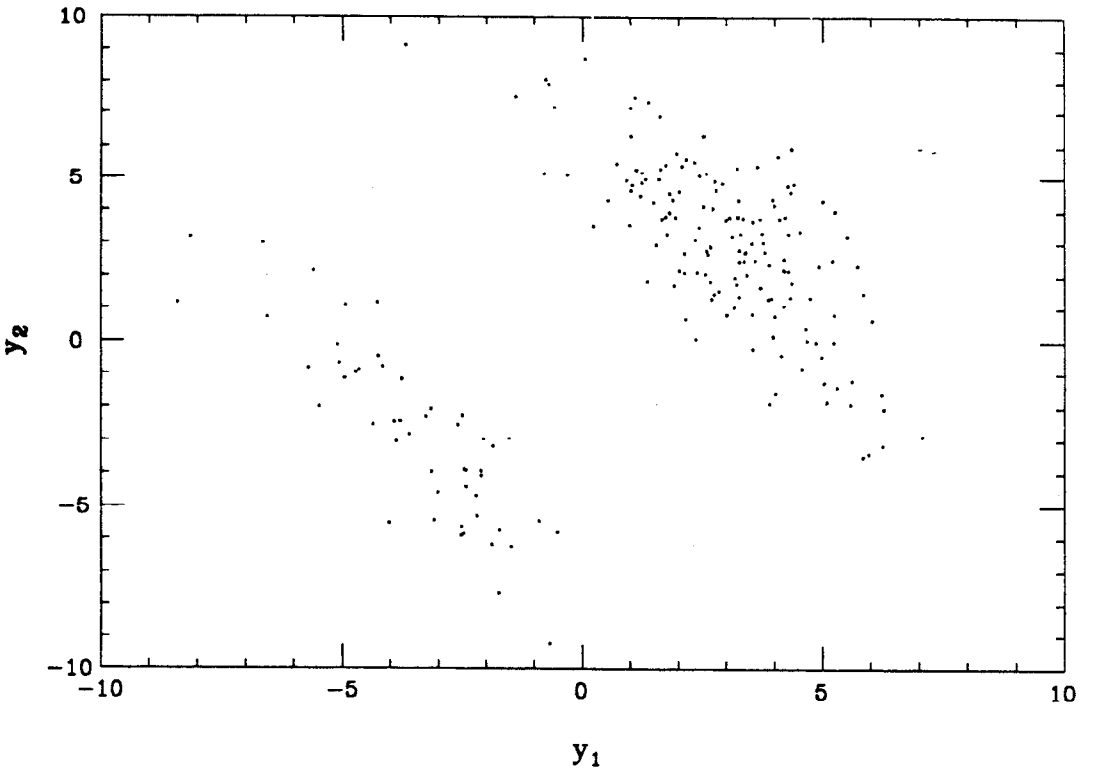


Fig. 1.5 Simulated data with $n = 200$ and $p = 2$.

The algorithm is run on the data using the varimax interpretability index for a single vector S_1 and the Legendre index G_L. Rotational invariance and axes restriction modifications are irrelevant in this situation since a one dimensional solution is sought. The values of the simplicity parameter λ for which the solutions are found are $(0, 0.1, 0.2, \ldots, 1.0)$.

The most structured line through the data should be about thirty degrees or the first principal component of the data. When projected onto this line, the observations are split into two groups. The most interpretable lines in the entire space, R^2 in this case, are the horizontal and vertical axes. The algorithm should move the solution line toward the more structured of these two axes. From Fig. 1.5, the horizontal axis exhibits the most structure when the data is projected onto it. If the data is projected onto the vertical axis, the two clusters merge into one. In the horizontal projection, the two groups only overlap slightly.

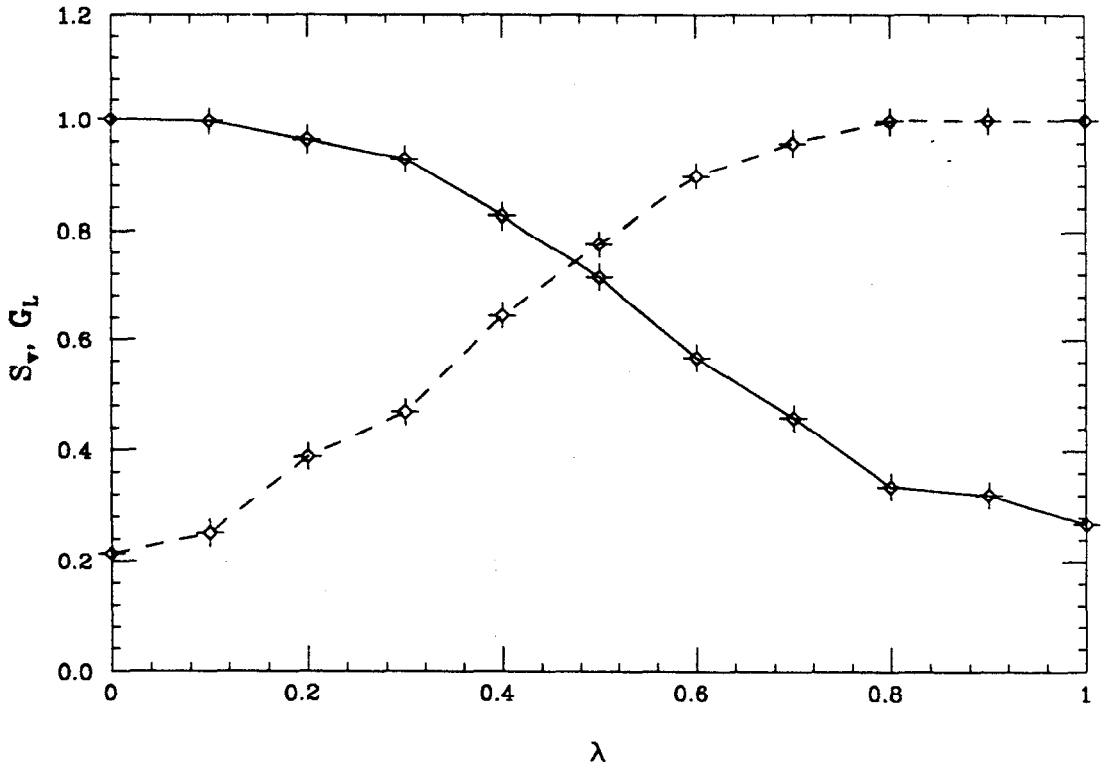


Fig. 1.6 Projection and interpretability indices versus λ for the simulated data. The projection index values are normalized by the $\lambda = 0$ value and are joined by the solid line. The simplicity index values are joined by the dashed line.

Fig. 1.6 shows the values of the projection and interpretability indices versus the values of λ. The projection index begins at 1.0 as it is normed and moves downward to about 0.3 as the weight on simplicity increases. The simplicity index begins at about 0.2 and increases to 1.0.

In addition to this graph, the statistician should view the various projections. Fig. 1.7 shows the projection histograms, values of the indices, and the linear combinations for four chosen values of λ. The histogram binwidths are calculated as twice the interquartile range divided by $n^{\frac{1}{3}}$ and the histograms are normed to have area one.

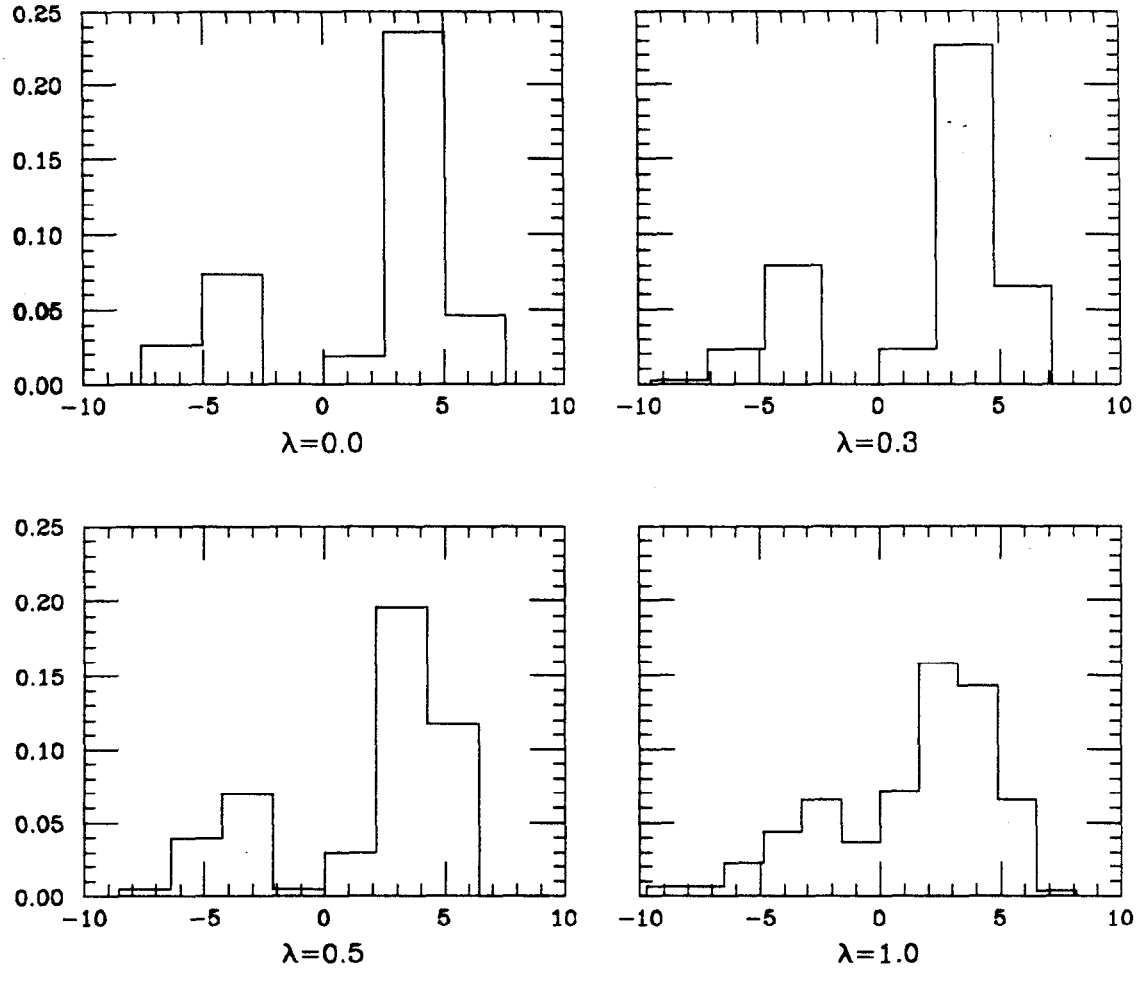


Fig. 1.7 Projected simulated data histograms for various values of λ. The values of the indices and combinations are

$$\lambda = 0.0,\ G_L = 1.00,\ S_1 = 0.21,\ \beta = (0.85, 0.52)^T$$
$$\lambda = 0.3,\ G_L = 0.93,\ S_1 = 0.47,\ \beta = (0.92, 0.47)^T$$
$$\lambda = 0.5,\ G_L = 0.78,\ S_1 = 0.71,\ \beta = (0.97, 0.24)^T$$
$$\lambda = 1.0,\ G_L = 0.27,\ S_1 = 1.00,\ \beta = (1.00, 0.00)^T$$

As predicted, the algorithm moves from an angle of about thirty degrees to the horizontal axis. As the weight on simplicity is increased, the loss of structure is evident in the merging of the two groups. In the final histogram, the two groups are overlapping. Also important to note is the comparison between the

first and second histograms at $\lambda = 0$ and $\lambda = 0.3$ respectively. A loss of 7% in the value of the projection index is traded for a gain in simplicity of 0.26. However, the two histograms are virtually indistinguishable to the eye.

1.5.2 The Automobile Example

The automobile data discussed in Section 1.1.3 is re-examined using two dimensional interpretable exploratory projection pursuit. Recall that this data consists of 392 (n) observations of ten (p) variables.

The interpretable exploratory projection pursuit analysis uses the Fourier G_F projection index instead of the Legendre G_L index due to the desire for rotational invariance as discussed in Section 1.4.1. Naturally, this choice results in a different starting point for the algorithm, the $\lambda = 0$ or P_0 plane in the notation of Section 1.4.5. Comparison of these two solutions demonstrates the lack of rotational invariance in the Legendre projection. Fig. 1.8 shows the P_0 projection using the Fourier index. The dashed axes are (β_1, β_2), which are orthogonal in the original variable space and are the simplest representation of the plane. The corresponding P_0 solution for the Legendre index is shown in Fig. 1.9. This plot is the same as Fig. 1.1 but with the same limits as Fig. 1.8 for comparison purposes.

In Fig. 1.9, the dashed axes are (β_1, β_2), which are orthogonal in the covariance metric. Rigidly rotating the combinations maintains the correlation constraint [1.6]. However, the Legendre index changes as shown in Table 1.1. With imagination, spinning these axes through different angles produces lower index values. For example, the new marginals after a rotation of $\frac{\pi}{4}$ are less clustered and therefore G_L is reduced.

The value of the Fourier index G_F for the projection in Fig. 1.8 is 0.21 while the value for the projection in Fig. 1.9 is 0.19. The Legendre index G_L value for the second is 0.35 as previously reported. For the projection in Fig. 1.8, the Legendre index varies from 0.19 to 0.21, depending on the orientation of the axes.

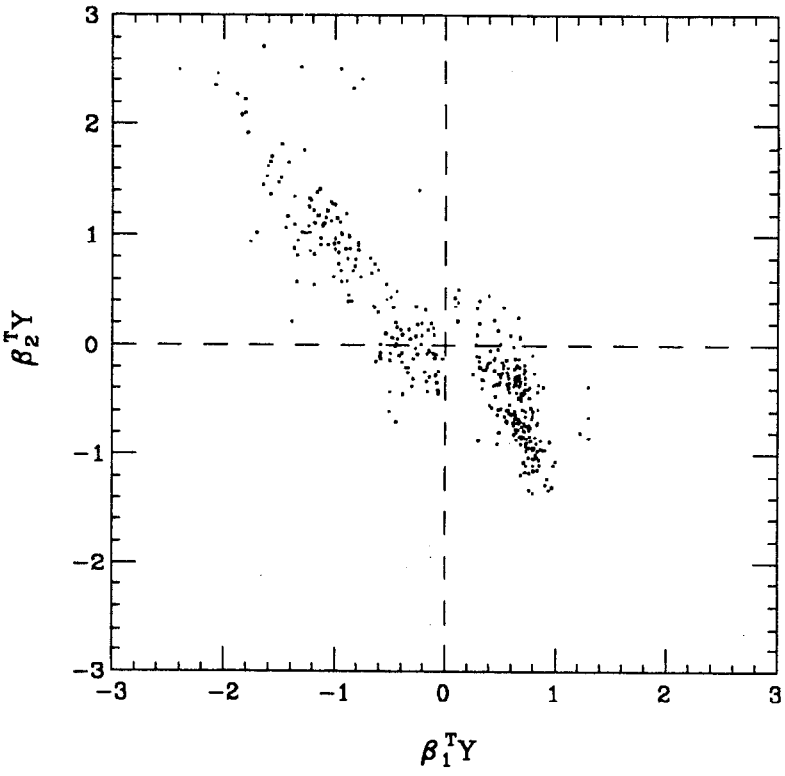


Fig. 1.8 Most structured projection scatterplot of the automobile data according to the Fourier index. The dashed axes are the solution combinations which define the plane.

Both find projections which exhibit clustering into two groups. If the Legendre index combinations are translated to an orthogonal pair via [1.5], they become

$$\beta_1 = (-0.21, -0.03, -0.91,\ 0.16,\ \ 0.30, -0.05, -0.01,\ \ 0.03,\ \ 0.00, -0.02)^T$$

$$\beta_2 = (-0.80, -0.14,\ \ 0.27,\ \ 0.49, -0.02,\ \ 0.03, -0.16, -0.02,\ \ 0.02, -0.01)^T$$

and have interpretability measure $S_v(\beta_1, \beta_2) = 0.50$. Of course, this may not be the simplest representation of the plane though the orthogonalizing transformation may help.

The Fourier index combinations originally are

$$\beta_1 = (-0.06, -0.22, -0.79, -0.44,\ \ 0.34, -0.05,\ \ 0.02, -0.14, -0.05,\ \ 0.04)^T$$

$$\beta_2 = (\ \ 0.00,\ \ 0.11, -0.53,\ \ 0.82, -0.03,\ \ 0.05, -0.01,\ \ 0.16,\ \ 0.04, -0.06)^T$$

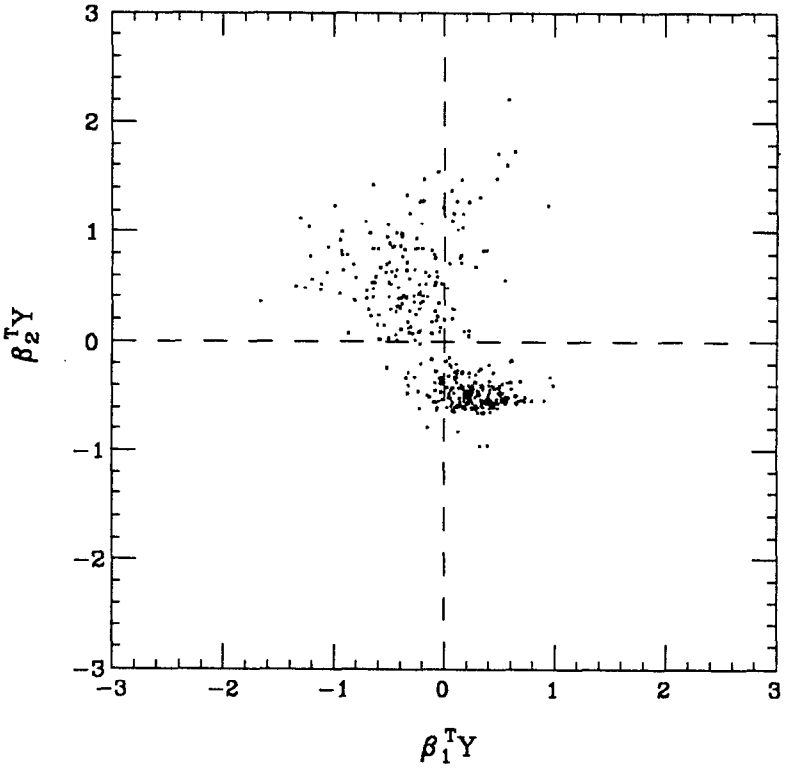


Fig. 1.9 Most structured projection scatterplot of the automobile data according to the Legendre index. The dashed axes are the solution combinations which define the plane maximize G_L and are orthogonal in the covariance metric.

and have interpretability measure 0.17. These axes are spun to the simplest representation as discussed in Section 1.4.5 and orthogonalized. The resulting axes are shown in Fig. 1.8 and the combinations are

$$\beta_1 = (-0.03, -0.12, -0.95, -0.01, \quad 0.29, -0.02, \quad 0.00, -0.03, -0.02, \quad 0.00)^T$$
$$\beta_2 = (-0.02, \quad 0.16, -0.09, \quad 0.94, -0.19, \quad 0.09, -0.01, \quad 0.18, \quad 0.07, -0.06)^T$$

with interpretability measure 0.80.

Fig. 1.10 shows the values of the interpretability and Fourier indices for simplicity parameter values $\lambda = (0.0, 0.1, \ldots, 1.0)$, analogous to Fig. 1.6 for the simulated data. This example demonstrates that the projection index may increase as the interpretability index does, possibly because the algorithm gets bumped

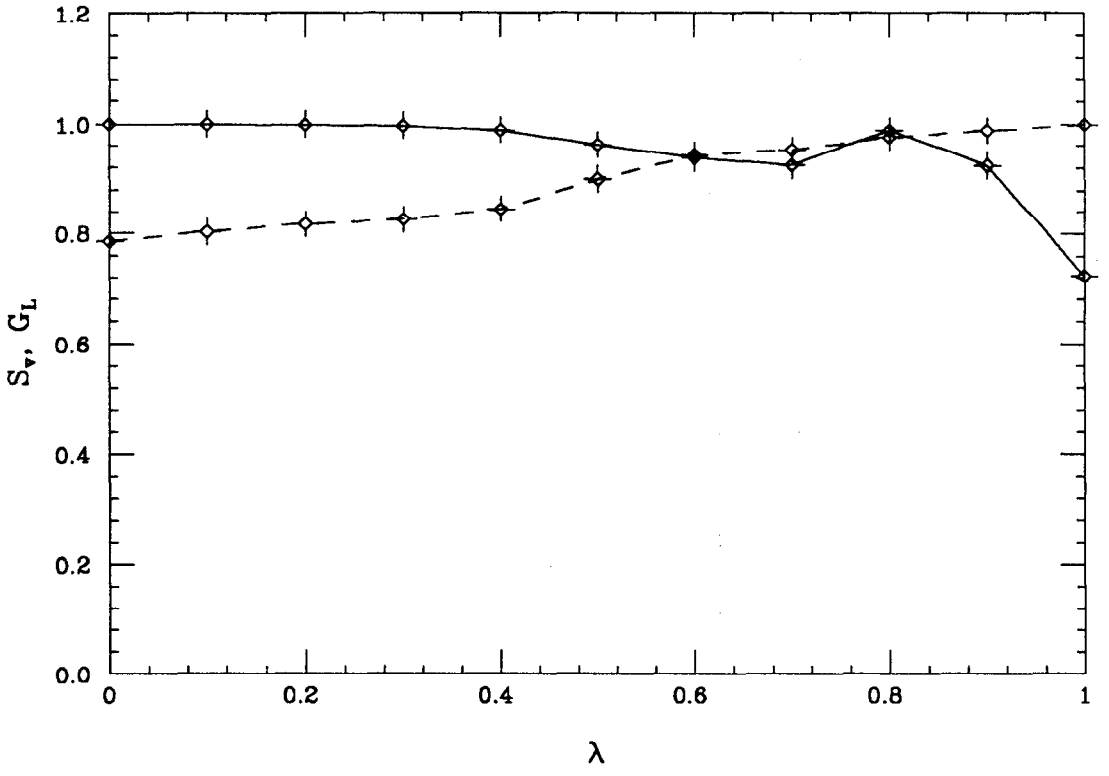


Fig. 1.10 Projection and interpretability indices versus λ for the automobile data. The projection index values are normalized by the $\lambda = 0$ value and are joined by the solid line. The simplicity index values are joined by the dashed line.

out of a projection local maximum as it moves toward a simpler solution.

Given this plot, the statistician may then choose to view several of the solution planes for specific λ values. Six solutions are shown in Fig. 1.11. The actual values of the combinations are given in Table 1.2. If coefficients which are less than 0.05 in absolute value are replaced by —, Table 1.3 results. In some sense, this action of discarding 'small' coefficients is contrary to the continuous nature of the interpretability measure. However, the second table shows the rate at which the coefficients decrease. Due to the squaring in the index S_v, the coefficients move quickly to one but slowly to zero. This convergence problem suggests investigating the use of a power less than two in the index in the future as discussed in Section 3.4.3. The table also demonstrates the global simplicity

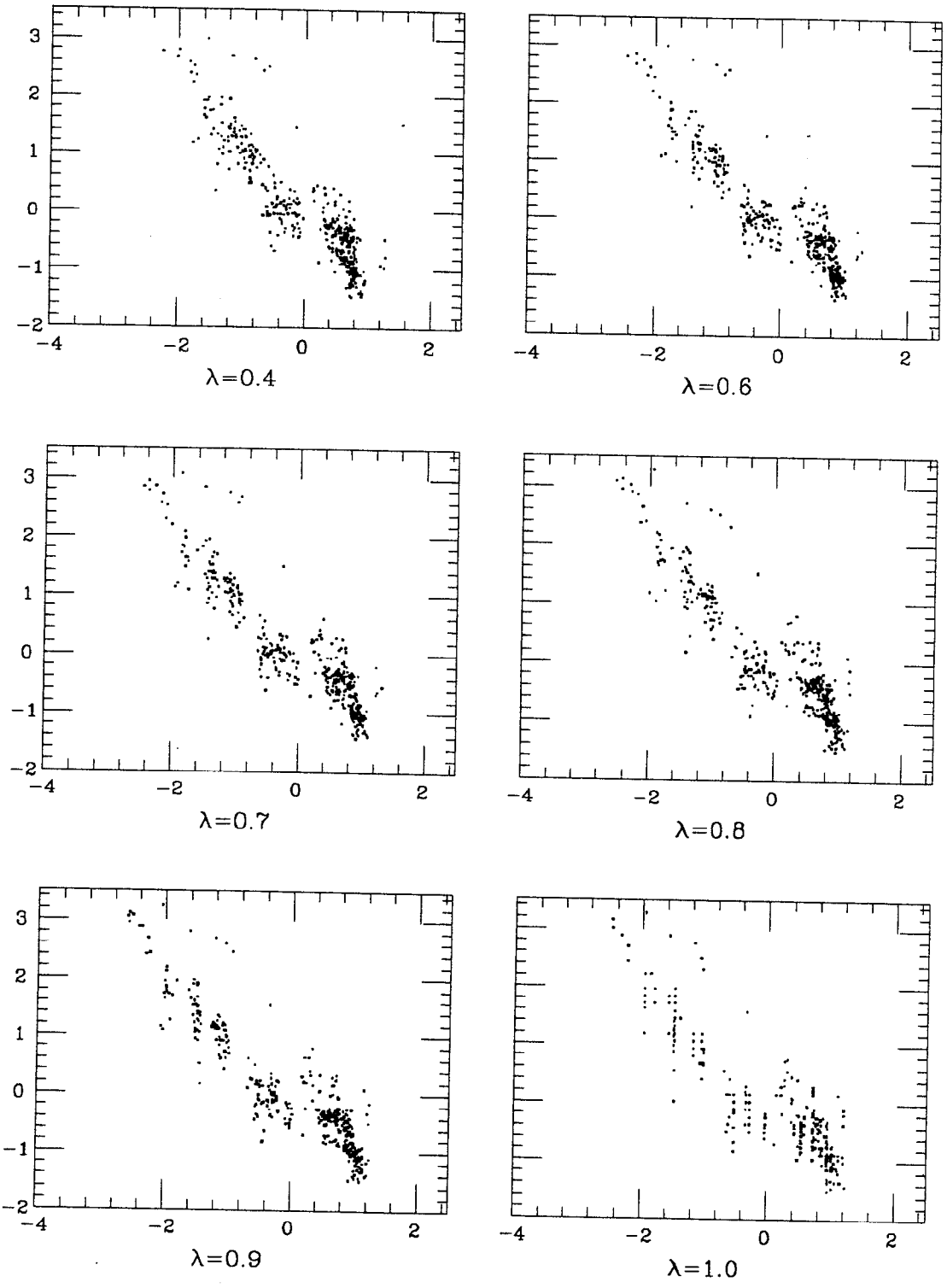


Fig. 1.11 Projected automobile data scatterplots for various values of λ. The values of the indices can be seen in Fig. 1.10 and the combinations are given in Table 1.2

λ	β_1, β_2									
0.0	-0.03	-0.12	-0.95	-0.01	0.29	-0.02	0.00	-0.03	-0.02	0.00
	-0.02	0.16	-0.09	0.94	-0.19	0.09	-0.01	0.18	0.07	-0.06
0.1	-0.02	-0.11	-0.95	0.04	0.27	-0.01	0.00	-0.02	-0.02	0.00
	-0.03	0.17	-0.04	0.94	-0.20	0.09	-0.01	0.18	0.07	-0.06
0.2	-0.01	-0.10	-0.96	0.05	0.25	-0.01	0.01	-0.02	-0.02	0.00
	-0.04	0.17	-0.02	0.94	-0.19	0.08	-0.02	0.18	0.07	-0.05
0.3	-0.01	-0.10	-0.96	0.06	0.24	-0.01	0.01	-0.02	-0.02	0.00
	-0.05	0.17	-0.01	0.94	-0.18	0.08	-0.02	0.18	0.07	-0.05
0.4	0.00	-0.09	-0.97	0.06	0.22	-0.01	0.02	-0.02	-0.02	0.00
	-0.06	0.15	0.00	0.95	-0.15	0.08	-0.03	0.19	0.07	-0.06
0.5	0.00	-0.08	-0.98	0.04	0.16	-0.02	0.03	-0.02	-0.02	-0.01
	-0.03	0.12	0.01	0.97	-0.11	0.12	-0.03	0.14	0.06	-0.04
0.6	0.01	-0.05	-0.99	0.01	0.12	-0.01	0.04	0.00	-0.01	-0.01
	-0.03	0.06	0.00	0.98	-0.07	0.11	-0.03	0.11	0.05	-0.03
0.7	0.01	-0.05	-0.99	0.02	0.09	-0.01	0.05	-0.01	-0.02	-0.01
	-0.02	0.06	0.01	0.98	-0.04	0.10	-0.04	0.11	0.05	-0.03
0.8	0.02	-0.04	-0.99	0.01	0.09	0.04	0.02	0.02	-0.01	0.00
	-0.02	-0.02	0.01	0.99	-0.02	-0.02	0.03	0.01	0.04	-0.09
0.9	0.01	-0.03	-1.00	0.00	0.03	0.03	0.02	0.01	0.00	-0.01
	0.00	-0.01	0.00	1.00	0.00	0.01	0.03	0.02	0.03	-0.07
1.0	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 1.2 Linear combinations for the automobile data. The linear combinations are given for the range of λ values. The first row for each λ value is β_1^T and the second is β_2^T.

λ	β_1, β_2									
0.0	—	-0.12	-0.95	—	0.29	—	—	—	—	—
	—	0.16	-0.09	0.94	-0.19	0.09	—	0.18	0.07	-0.06
0.1	—	-0.11	-0.95	—	0.27	—	—	—	—	—
	—	0.17	—	0.94	-0.20	0.09	—	0.18	0.07	-0.06
0.2	—	-0.10	-0.96	0.05	0.25	—	—	—	—	—
	—	0.17	—	0.94	-0.19	0.08	—	0.18	0.07	-0.05
0.3	—	-0.10	-0.96	0.06	0.24	—	—	—	—	—
	-0.05	0.17	—	0.94	-0.18	0.08	—	0.18	0.07	-0.05
0.4	—	-0.09	-0.97	0.06	0.22	—	—	—	—	—
	-0.06	0.15	—	0.95	-0.15	0.08	—	0.19	0.07	-0.06
0.5	—	-0.08	-0.98	—	0.16	—	—	—	—	—
	—	0.12	—	0.97	-0.11	0.12	—	0.14	0.06	—
0.6	—	-0.05	-0.99	—	0.12	—	—	—	—	—
	—	0.06	—	0.98	-0.07	0.11	—	0.11	0.05	—
0.7	—	-0.05	-0.99	—	0.09	—	0.05	—	—	—
	—	0.06	—	0.98	—	0.10	—	0.11	0.05	—
0.8	—	—	-0.99	—	0.09	—	—	—	—	—
	—	—	—	0.99	—	—	—	—	—	-0.09
0.9	—	—	-1.00	—	—	—	—	—	—	—
	—	—	—	1.00	—	—	—	—	—	-0.07
1.0	—	—	-1.00	—	—	—	—	—	—	—
	—	—	—	1.00	—	—	—	—	—	—

Table 1.3 Abbreviated linear combinations for the automobile data. The linear combinations are given for the range of λ values as in Table 1.2. A — replaces any coefficient less than 0.05 in absolute value.

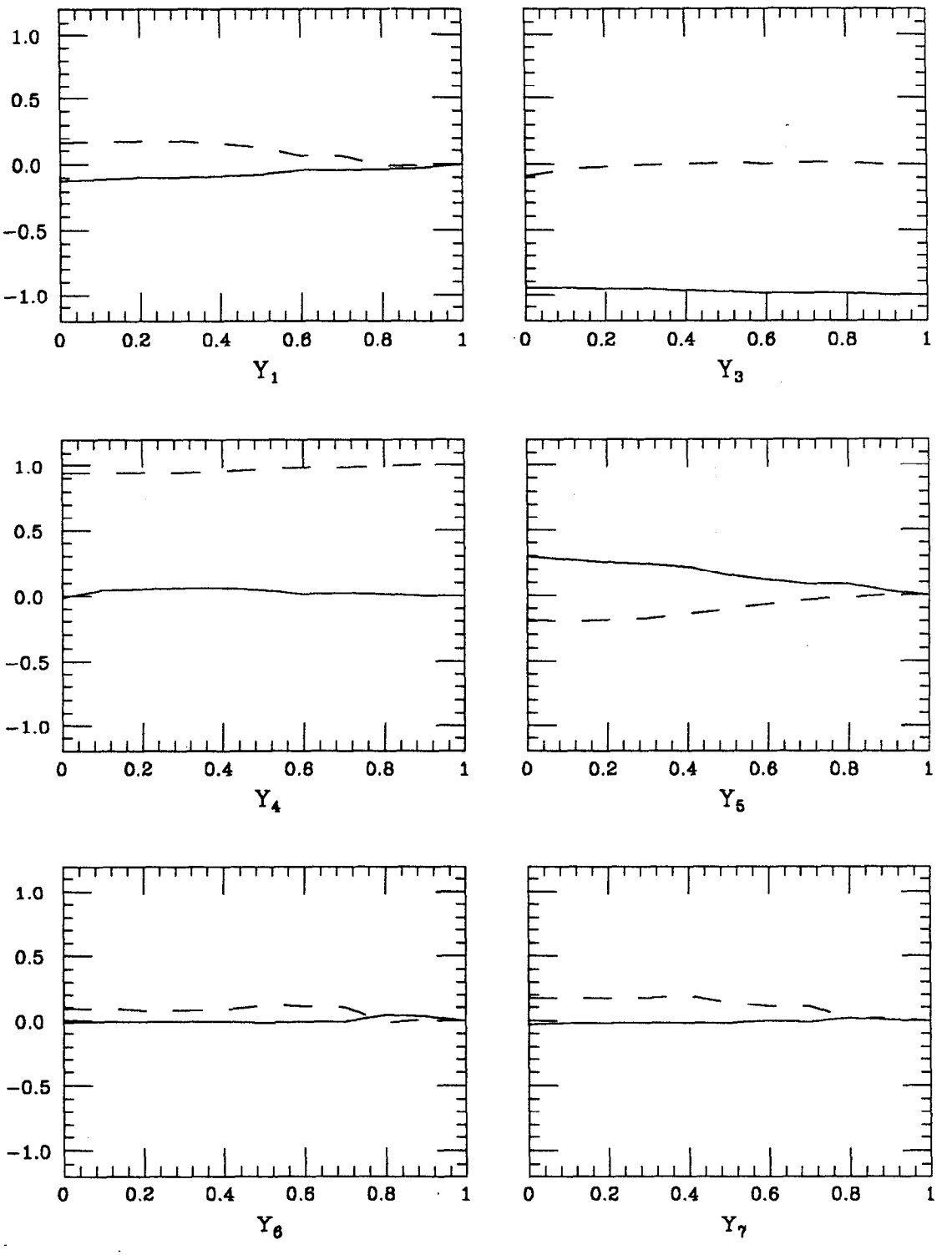


Fig. 1.12 Parameter trace plots for the automobile data. The values of the parameters are given for those variables whose coefficients are large enough. The parameter values (β_{1i} solid, β_{2i} dashed) are plotted versus λ.

of the two combinations. If one variable is in a combination, it usually is absent in the other.

Another useful diagnostic tool fashioned after ridge regression trace graphs is shown in Fig. 1.12. For variables whose coefficients are ≥ 0.10 for any value of λ, the values of the coefficients in each combination are plotted versus λ.

One of the most interesting solution planes is the one for $\lambda = 0.8$. The combinations involve four variables which are split into two pairs, one in each combination. The first combination involves the negative of the third variable engine size and the fifth variable automobile weight. The second combination involves the engine power and the negative of the gaussianized Japanese flag. Fig. 1.13 shows the solution plane with the type of car (Japanese or Non-Japanese) delineated by the plotting symbol.

The statistician must decide when the tradeoff between simplicity and accuracy should be halted. In this example, the $\lambda = 0.8$ model is a good stopping place, especially since the projection index actually increased from the previous plane. However, the decision may be difficult. Given that this is an exploratory technique, no external measurement such as prediction error can be used to judge a plane's usefulness. Rather, the decision rests with the statistician.

No doubt this data has a wealth of planes which exhibit structure. In fact, two are evident in Figs. 1.8 and 1.9. The object is to find all such views. After a structured plane is found, its structure should be removed without disturbing the structure of other interesting planes. Friedman (1987) presents a method for structure removal. He recursively applies a transformation to the structured plane which normalizes it yet leaves other orthogonal planes undisturbed. The interpretable exploratory projection pursuit algorithm can employ the same procedure to find several interesting planes. Once the λ value and hence the particular plane have been chosen, the structure is removed and the next plane is found and simplified as desired.

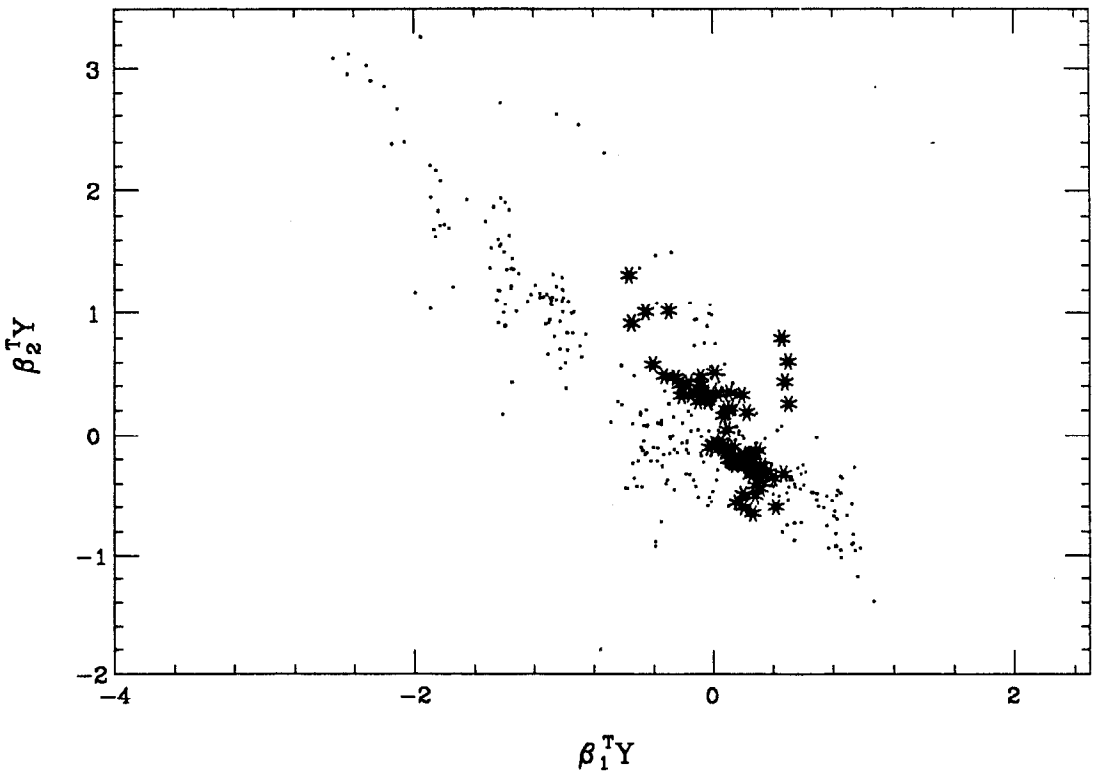


Fig. 1.13 Country of origin projection scatterplot of the automobile data for $\lambda = 0.8$. American and European cars are plotted as points, Japanese cars as asterisks.

The interpretability of a collection of solution planes could be considered. For example, two planes might be simple in relation to each other if they are orthogonal. However, since the structure removal procedure does not affect structure in orthogonal planes, in practice solution sets of planes tend to be orthogonal anyway.

Originally, exploratory projection pursuit was interactive, as mentioned in Section 1.1.1. After choosing three variables, the statistician used a joystick to rotate the point cloud in real time. A fourth dimension could be shown in color. The statistician could then pick out structure by eye. However, this task was time-consuming and only allowed combinations of four variables at best. The solution was to automate the structure identification by defining an index

which measures structure mathematically and to use a computer optimizer. The statistician loses some of her individual choice in structure identification but she can define her own index if desired.

Interpretable exploratory projection pursuit is an analogous automation of variable selection. The interpretability index is a mathematical measure which, coupled with a numerical optimization routine, takes the place of interactively choosing which three or four variables to view.

Chapter 2

Interpretable Projection Pursuit Regression

This chapter describes interpretable projection pursuit regression. It is similar in organization to the previous chapter though condensed as several issues common to both have been addressed previously. Section 2.1 deals with the original algorithm, reviewing the notation and strategy. In the second section, the modification is considered and the new algorithm is detailed. Due to the differing goals of exploratory projection pursuit and projection pursuit regression, the interpretability index must be changed slightly from that of Chapter 1. The final section consists of an example.

2.1 The Original Projection Pursuit Regression Technique

The original projection pursuit regression technique is presented in Friedman and Stuetzle (1981). Friedman (1984a, 1985) improves several algorithmic features and extends the approach to include classification and multiple response regression in addition to single response regression. In this chapter, only single response regression is considered.

2.1.1 Introduction

The easiest way to understand projection pursuit regression is to consider it as a generalization of ordinary linear regression. Many authors motivate projection pursuit regression in this manner, among them Efron (1988) and McDonald (1982). For ease of notation, suppose the means are removed from each of the predictors $X = (X_1, X_2, \ldots, X_p)^T$. The goal is to model the response Y as a linear function of the centered X. With the usual assumptions and slightly unfamiliar notation, the single response linear regression model may be written

$$Y - E[Y] = \omega^T X + \epsilon \qquad [2.1]$$

where ϵ is the random error term with zero mean. The vector $\omega = (\omega_1, \omega_2, \ldots, \omega_p)^T$ consists of the regression coefficients.

In general, this linear function is estimated by the conditional expectation of Y given particular values of the predictors $x = (x_1, x_2, \ldots, x_p)$. The fitted value of Y is

$$\hat{Y}(x) = E[Y] + \omega^T x \quad . \qquad [2.2]$$

The expected value of Y is estimated by the sample mean. The expected L_2 distance between the true and fitted random variables is

$$L_2(\omega, X, Y) \equiv E[Y - \hat{Y}]^2 \quad .$$

The parameters ω of the model are estimated by

$$\min_{\omega} \; L_2(\omega, X, Y) \quad . \qquad [2.3]$$

In practice, the sample mean over the n data points replaces the population mean.

The model [2.1] may be written

$$Y - E[Y] = \beta(\alpha^T X) + \epsilon \qquad [2.4]$$

where $\beta = \omega^T\omega$ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ is ω normed. The resulting fitted value equation is analogous to [2.2]. The parameters β and α may be estimated as in [2.3] subject to the constraint that $\alpha^T\alpha = 1$.

The rewritten model [2.4] shows that response variable Y depends only on the projection of the predictor variables X onto the direction α. The relationship between the fitted values $\hat{Y}$ and the projection $\alpha^T X$ is a straight line. A natural generalization is to allow this relationship to vary. Projection pursuit regression does just that, allowing the fitted value to be a sum of univariate functions of projections which are smooth but otherwise unrestricted parametrically.

The projection pursuit regression model with m terms is

$$Y - E[Y] = \sum_{j=1}^{m} \beta_j f_j(\alpha_j^T X) + \epsilon \quad . \qquad [2.5]$$

The linear combinations α_j which define the directions of the smooths are restricted to have length one. In addition, the functions f_j are smooth, and have zero mean and unit variance. The parameters β_j capture the variation between the terms. In the usual way, the conditional mean is used to estimate the sum of functions as

$$\hat{Y}(x) = E[Y] + \sum_{j=1}^{m} \beta_j f_j(\alpha_j^T x)$$

with the parameters estimated as in [2.3].

Analogous to exploratory projection pursuit, projection pursuit regression may be more successful than other nonlinear methods by working in a lower dimensional space (Huber 1985). The model [2.5] is useful when the underlying

relationship between the response and predictors is nonlinear, versus ordinary regression [2.4], and when the relationship is smooth, as opposed to other nonlinear methods such as recursive partitioning. Diaconis and Shahshahani (1984) show that any function can be approximated by the model [2.5] for a large enough number of terms m. Substantial work remains to be done with respect to the theoretical properties of the method. In addition, the numerical aspects of the algorithm are difficult as discussed in Section 2.2.3.

2.1.2 The Algorithm

The parameters β_j, α_j and the functions f_j are estimated by minimizing

$$\begin{aligned} \min_{\beta_j,\alpha_j,f_j:j=1,\ldots m} \quad & L_2(\beta,\alpha,f,X,Y) \\ \ni \quad & \alpha_j^T\alpha_j = 1 \\ & E[f_j] = 0 \\ & \text{and } \mathrm{Var}[f_j] = 1 \qquad j = 1,\ldots,m \quad . \end{aligned} \tag{2.6}$$

The criterion [2.6] cannot be minimized simultaneously for all the parameters. However, if certain ones are fixed, the optimal values of others are easily solved for. Friedman (1985) employs such an 'alternating' optimization strategy. His results are discussed in this section as they are pertinent when the modified algorithm is considered in Section 2.2.3. First, he considers a specific term k, $k = 1,\ldots,m$. The problem [2.6] may be written

$$\begin{aligned} \min_{\beta_k,\alpha_k,f_k} \quad & E[R_k - \beta_k f_k(\alpha_k^T X)]^2 \\ & \text{where } R_k \equiv Y - E[Y] - \sum_{j\neq k}^{m} \beta_j f_j(\alpha_j^T X) \quad . \end{aligned} \tag{2.7}$$

For the k^{th} term, the three sets of parameters β_k, α_k, and f_k are estimated in turn while all others are held constant. After all elements of the k^{th} term have been found, the next term is considered. The algorithm cycles through the

terms in the model until the objective in [2.6] does not decrease sufficiently. The alternating strategy is discussed in more detail in Section 2.2.3.

The minimizing β_k is

$$\beta_k = \frac{E[R_k f_k(\alpha_k^T X)]}{E[f_k(\alpha_k^T X)]^2} \quad . \qquad [2.8]$$

The minimizing function f_k for any particular point $\alpha_k^T x$ is

$$f_k(\alpha_k^T x) = \frac{E[\beta_k R_k | \alpha_k^T x]}{\beta_k^2} \quad . \qquad [2.9]$$

This estimate is found using the nonparametric smoother discussed in Friedman (1984b). The resulting curve is standardized to satisfy mean and variance constraints. It is not expressed as a mathematical function but rather as an estimated value for each observation.

The minimizing direction α_k cannot be determined directly and requires an iterative procedure. Minimizing the criterion [2.6] as a function of α_k is a least-squares problem as seen in [2.7]. In applying an iterative search procedure to minimize a function, the goal is to use as much information about the function as possible. Thus, rather than use a method which only employs first derivatives, a method which also uses the actual or approximate Hessian is preferable. Usually the difficulty of applying these methods, specifically determining or accurately estimating the Hessian, outweighs their additional optimization properties. However, if the function to be minimized is of least-squares form, the Hessian simplifies and is easily approximated (Gill et al. 1981). Friedman (1985) capitalizes on this fact and uses the Gauss-Newton procedure to find the optimal α_k.

2.1.3 Model Selection Strategy

Model selection in the original projection pursuit regression consists of choosing the number of terms m in the model. Interpretable projection pursuit regression considers not only the number of terms but also the interpretability of those terms.

Friedman (1985) suggests starting the algorithm with a large number of terms M. The procedure for finding this model is discussed in Section 2.2.3. A model with $M-1$ terms is then determined using the $M-1$ most important terms in the previous model as a starting point. The importance of a term k in a model of m terms is defined as

$$I_k \equiv \frac{|\beta_k|}{|\beta_l|} \qquad [2.10]$$

where $|\beta_l|$ is the maximum absolute parameter. Since the functions f_j are constrained to have variance one, $|\beta_k|$ measures the contribution of the term to the model.

The statistician may then plot the value of the objective in [2.6], which is the model's residual sum of squares, versus the number of terms for each model. In most cases, the plot has an elbow shape. The usual advice is to choose that model closest to the tip of the elbow, where the increase in accuracy due to the additional term is not worth the increased complexity.

2.2 The Interpretable Projection Pursuit Regression Approach

A projection pursuit regression analysis produces a series of models with number of terms $m = 1, \ldots, M$. The models' nonlinear components are the functions f_j. Since these functions are smooths and not functionally expressed, they are visually assessed by the statistician. Each is considered along with its associated combination α_j in order to understand what aspect of the data it represents. The parameters β_j measure the relative importance of the terms.

Each direction α_j must be considered in the context of the original number of variables p. The collection of combinations is not subject to any global restriction such as the correlation constraint in exploratory projection pursuit. As with any regression technique, a variable selection method which causes the same subgroup of variables to appear in all the combinations is desirable for parsimony.

Arguments similar to those in Section 1.2 show that a combinatorial approach to variable selection in projection pursuit regression is not feasible. As before, a weighted penalty balancing the goodness-of-fit criterion [2.6] with an interpretability criterion is employed. The minimization problem becomes

$$\min_{\beta_j,\alpha_j,f_j:j=1,\ldots,m} (1-\lambda)\frac{L_2(\beta,\alpha,f,X,Y)}{\min L_2} - \lambda S(\alpha_1,\alpha_2,\ldots,\alpha_m) \qquad [2.11]$$

for $\lambda \in [0,1]$. The denominator in the first term causes both contributions to be $\in [0,1]$. The simplicity index measures the interpretability of the collection of combinations and is subtracted since the objective is minimized. Interpretability index choices are discussed in the next two subsections.

2.2.1 The Interpretability Index

Consider for the moment that the number of terms m is fixed. This assumption is discussed in Section 2.2.2. The goal is define an interpretability index S for a group of m vectors $(\alpha_1,\alpha_2,\ldots,\alpha_m)$. At first glance, the situation is a generalization of interpretable exploratory projection pursuit where the simplicity of two vectors (β_1,β_2) is measured. However, in the latter case, the vectors define a plane. In relation to each other, the simplest pairs of combinations are ones which, when normed and squared, are orthogonal. This type of orthogonality is named squared orthogonality in Section 2.3.5. Measures of this squared orthogonality and each vector's individual interpretability enter the index S_v [1.18]. Within each combination, variables are selected via the single vector varimax interpretability index S_1 [1.12] by encouraging the vector coefficients to vary.

Different variables are selected for either vector by forcing the combinations to exclude the same variables.

Projection pursuit regression has a different goal. First, each vector should be simple within itself. That is, homogeneity of the coefficients within a vector is penalized. The varimax index S_1 for one vector achieves this. However, in the interest of decreasing the total number of variables in the model, the same small subset of variables should be nonzero in each vector. Summing the simplicities of each direction as in [1.17] does not force the vectors to include the same variables necessarily. On the other hand, if the exploratory projection pursuit varimax measure [1.16] for $q = m$ vectors is used, the vectors would be forced to squared orthogonality and would contain different variables. This old index may be used for projection pursuit regression to achieve this outcome. However, if variable selection is the object, a new index must be developed.

Consider the m X p matrix of combinations

$$\Omega \equiv \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} \quad .$$

Though the directions α_j are constrained to have length one in [2.6], the index is defined for all sets of combinations in general. The goal is that all the combinations are nonzero or load into the same columns so that the same variables are in all combinations. An index which forces this outcome is based on a summary vector γ of the matrix Ω whose components are

$$\gamma_i \equiv \sum_{j=1}^{m} \frac{\alpha_{ji}^2}{\alpha_j^T \alpha_j} \qquad i = 1, \ldots, p \quad . \qquad [2.12]$$

Each component γ_i is positive and contains the sum of each term's relative contribution to variable i, or the total column weight. The components of the new

vector sum to m. The object is to force these components to be as varied as possible. For a single combination, this is achieved via the varimax interpretability index S_v [1.12]. For γ appropriately normed, this measure is

$$S_1(\gamma) = \frac{p}{p-1}\sum_{i=1}^{p}\left(\frac{\gamma_i}{m} - \frac{1}{p}\right)^2 \quad . \qquad [2.13]$$

The index [2.13] forces the weight of the columns of Ω to vary widely. As a function of all the combinations, this index is called S_g and may also be written

$$S_g(\alpha_1, \alpha_2, \ldots, \alpha_m) = \frac{1}{m^2}\left[\sum_{j=1}^{m} S_1(\alpha_j) + \frac{p}{2(p-1)}\sum_{j\neq k}^{m}\sum_{i=1}^{p}\frac{\alpha_{ji}^2}{\alpha_j^T\alpha_j}\frac{\alpha_{ki}^2}{\alpha_k^T\alpha_k}\right] + \frac{1-m}{m(p-1)} \quad .$$

The function S_1 measures the individual interpretability of each combination. This re-expression is in contrast to S_v for interpretable exploratory projection pursuit [1.18], in which the cross-product over the terms is subtracted in order to force the squared orthogonality of the combinations.

The index forces the overall weights of each column to be dissimilar. The dispersion of the each total column weight over the rows of the matrix, or terms of the model, depends on the goodness-of-fit criterion. An example using this index is discussed in Section 2.3.

2.2.2 Attempts to Include the Number of Terms

In the discussion so far, the number of terms m in the model is fixed. The argument that a model with fewer terms is more interpretable than one with more persuasive. However, on further reflection, this conclusion does not appear so certain. Each term involves its combination α_j and the smooth function f_j. The interpretability of the combination can be measured but the function must be visually assessed.

Consider the following example with two variables X_1 and X_2 ($p = 2$). Without knowing the data context, ranking the two fitted models

$$(1)\ m = 2,\ f_1(X_1),\ f_2(X_2)$$

$$(2)\ m = 1,\ f(\frac{X_1 + X_2}{2})$$

in order of interpretability is impossible. The first model involves two functions of one variable each, the simplest combination. The second involves only one function consisting of the most complex combination of variables, the average. Situations easily can be imagined in which clearly one or the other model is more interpretable. For example, if the two variables are distinct measurements of widely varying traits (apples vs. oranges), then a combination of the two would be difficult to understand and Model (1) would be preferable. However, if the two were aggregate measures of similar characteristics (reading and spelling scores) which could be combined easily into a single variable (verbal ability), the second model might be easier to interpret. However, considering ways to include the number of terms in the interpretability measure of a model is enlightening even if present attempts are unsuccessful.

As with interpretable exploratory projection pursuit, interpretable projection pursuit regression can be envisioned as an interactive process in which the statistician is equipped with an interpretability dial. As the dial is turned, the weight (λ) in [2.11] is increased and the model becomes more simple. If the number of terms to begin with is m and this parameter is included in the measure of simplicity, the algorithm should drop terms as it simplifies the model. In the following discussion, three methods for including m in the index [2.13] are considered. Factors required to put the index values $\in [0, 1]$ are ignored. Unfortunately, none of the resulting indices works in practice for the reasons given below.

Since the interpretability of a model decreases with the number of terms, the first attempt is to multiply the interpretability index [2.13] by $\frac{1}{m}$, obtaining

$$S_a(\alpha_1, \alpha_2, \ldots, \alpha_m) \equiv \frac{1}{m}\left[\frac{p}{p-1}\sum_{i=1}^{p}\left(\frac{\gamma_i}{m} - \frac{1}{p}\right)^2\right] \quad .$$

The resulting index S_a decreases as the number of terms increases. However, this measure does not work in practice as each term's contribution is reweighted when the number of terms changes. Instead, the index should be such that submodels of the current model measured contribute the same to the index, regardless of the size of the complete model.

Each model can be thought of as a point $\in R^p$. As terms are added, the number of points increases. Consider a distance interpretability index as in Section 2.3.4. The total of distance from a set of points to a particular point increases as the number of points does. Since the object is that all the terms contain the same variables, a plausible interpretability index is

$$S_b(\alpha_1, \alpha_2, \ldots, \alpha_m) \equiv -\min_{\nu_l \in V} \sum_{j=1}^{m} \|\nu_{\alpha_j} - \nu_l\|^2 \quad .$$

As in Chapter 1, the set V is composed of simple vectors [1.14] and ν_{α_j} is the squared and normed version of the α_j term [1.13]. This index is similar to the one dimensional index [1.15]. The minimum is not taken of each individual term j but of the sum of distances in order to ensure that all the terms simplify to the same interpretable vector ν_l. If the set V consists of the e_l's ($l = 1, \ldots, p$), S_b reduces to

$$-\sum_{i=1}^{p}\sum_{j=1}^{m}\frac{\alpha_{ji}^2}{\alpha_j^T\alpha_j} - 2\gamma_k + m$$

where γ_k is the largest component as defined in [2.12]. Unfortunately, even with the removal of the minimization, the derivatives of the given index are discontinuous.

Averaging all the possible distances produces a third index

$$S_c(\alpha_1, \alpha_2, \ldots, \alpha_m) \equiv -\sum_{l=1}^{p}\sum_{j=1}^{m} \|\nu_{\alpha_j} - \nu_l\|^r$$

which is continuous. The power r must be chosen so that each term simplifies toward one of the interpretable vectors ν_l. As a result, r must be less than one so that the closest ν_l overpowers the others and pulls the term toward it. Unfortunately, though continuous as opposed to S_b, the index S_c does not force all the terms to collapse toward the same simple vector.

In conclusion, none of the three attempts at incorporating the number of terms into the index works. A method for simultaneously and smoothly measuring both the number of terms and the interpretability of the terms themselves has not been found yet. Without such a method, projection pursuit regression model selection cannot be reduced to a one parameter situation in which the interpretability parameter λ completely controls the tradeoff between interpretability and accuracy. The outlook for an analogy with model selection in linear regression is poor. Instead, interpretable projection pursuit regression is a procedure to explore the model space, rather than a strict model selection process. How the statistician should compare these models is discussed in Section 2.3. The optimization procedure is described in the next subsection.

2.2.3 The Optimization Procedure

The strategy is to begin with a large number of models M. For each submodel with number of terms $m = 1, \ldots, M$, the following algorithm is employed to find a sequence of interpretable models for different values of the interpretability parameter λ. Steps 1 and 2 find the original projection pursuit regression m term model which minimizes [2.6]. Steps 3 and 4 find the sequence of models which minimize [2.11] for various λ values. Throughout the description, updating a parameter means noting the new value and basing subsequent dependent

calculations on it. Moving a model means permanently changing its parameter values.

1. The objective function is [2.6]. Use the stagewise modeling procedure outlined in Friedman and Stuetzle (1981) to build the M term model.

2. Use a backwards stepwise approach to fit down to the m term model.

   For $i = 1, \ldots, M - m$
   begin

   Rank the terms from most (term 1) to least important (term $M - i + 1$) as measured by [2.10]. Discard the least important term.

   Use an alternating procedure to minimize the remaining $M - i$ terms.

   a. **For** $k = 1, \ldots, M - i$
   begin

   Update the term's parameters β_k, α_k and curve f_k assuming the other terms are fixed. Choosing from among several steplengths, do a single Gauss-Newton step to find the new direction α_k. Complete the iteration by updating β_k and f_k using [2.8] and [2.9] respectively. Only one Gauss-Newton step is taken for each iteration due to the expense of a step. Continue iterating until the objective stops decreasing sufficiently.

   end

   b. If the objective decreased on the last complete loop through the terms (*a*), move the model. If the objective decreased sufficiently, perform another pass (**GOTO** *a.*). Otherwise, the optimization of the $M - i$ term model is complete.

   end

3. Let $\lambda_0 = 0$ and call the m term model resulting from Steps 1 and 2 the λ_0 model. Choose a sequence of interpretability parameters $(\lambda_1, \lambda_2, \ldots, \lambda_I)$. Let $i = 1$.

4. The objective function is [2.11]. Set $\lambda = \lambda_i$ and solve for the m term model using a forecasting alternating procedure and the λ_{i-1} model as the starting point.

   Reorder the terms in reverse order of importance [2.10] from least (term 1) to most important (term m).

   Make a move in the best direction possible.

   a. **For** $k = 1, \ldots, m$
   begin

   Choosing from among several steplengths, update the α_k resulting from the best step in the steepest descent direction. Complete the iteration by updating β_k and f_k using [2.8] and [2.9] respectively. Only one steepest descent direction step is taken due to the expense of a step. Always perform at least one iteration and then continue iterating until the objective stops decreasing sufficiently.

   end

   b. If only one loop through the terms (*a*) has been completed, move the model regardless of whether the objective decreases or not and perform another pass (**GOTO** *a*). If more than one loop has been completed and the objective decreased, move the model. If more than one loop has been completed and the objective decreased sufficiently, perform another pass (**GOTO** *a*). Otherwise, the optimization of the m term model with interpretability parameter λ_i is complete.

   c. If $i = I$, **EXIT**. Otherwise, let $i = i + 1$ and **GOTO** 4.

The forecasting alternating procedure in Step 4 differs from the alternating procedure in Step 2. In the latter case, determining the optimal direction α_k for a specific term reduced to a least-squares problem as noted in Section 2.1.2. While this problem could not be solved analytically and had to be iterated, the least-squares form lent itself to a special procedure (Gauss-Newton) which utilized the estimated Hessian. However, with the addition of the interpretability term the objective [2.11] no longer has this form. Thus, a cruder optimization method must be used. Steepest descent is employed and the step direction is the negative of the gradient of the objective. Unfortunately, this method is not as accurate as Gauss-Newton. The gradients of the objective are given in Appendix A.

The Step 4 approach is also different in that an attempt is made to forecast the effect a simplification of one term will have on the others. In the original algorithm as described in Step 2, each term is considered separately. A term is not updated unless the objective function decreases as a result. This approach ignores the interaction between the terms, sacrificing accuracy in the solution for ease in calculation. However, scenarios exist in which this approximation does not produce the global minimum. For example, suppose a change in term one does not produce a decrease in the objective. However, possibly this change in term one and the resulting shift in term two produces a decrease. Such a combination of moves is not considered in the original algorithm.

Using this 'look one step ahead' approach in interpretable projection pursuit regression does not work because of the strong interaction between the terms in the simplicity index S_g and the equality among term contributions to the index. The increase in interpretability for an individual term must be very large before it changes on its own. However, once it changes, the other terms quickly follow suit, like a row of dominoes. The result is that if Step 2 approach is used, the algorithm produces either very complex or very simple models but none in between.

Thus, a compromise is reached based on empirical evidence. The first change is due to the fact that all terms contribute equally to the interpretability measure, irrespective of their relative importance [2.10]. Thus, when considering a simplification of the model, that term which least affects the model fit should be considered first as it does as well as any other at increasing interpretability. As a result, the terms are looked at in reverse order of importance (Step 4).

The second change is that the algorithm moves the entire set of m terms at once as opposed to one term at a time. A move is not evaluated until it is formed from a sequence of submoves, each of which is a shift of an individual term (Step $4a$). These submoves are made in reverse order of goodness-of-fit importance and are made in the best direction possible, the steepest descent one.

The third change is that the algorithm always moves the model, even if the move appears to be a poor one. In some respects this jiggling of the model is a form of annealing (Lundy 1985). The minimum steplength considered in Step $4a$ is positive yet quite small, so large unwelcome increases in the objective are impossible.

Both the second and third changes are attempts at forecasting the effect of the simplification of one term on subsequent terms. The given algorithm works well in practice as is demonstrated in the next section. Occasionally, the requirement that the algorithm always move means that a clearly worse model results. However, the algorithm quickly adjusts for the next value of λ. This behavior is seen in the following example.

2.3 The Air Pollution Example

The example in this section concerns air pollution. The data is analyzed using additive models in Hastie and Tibshirani (1984) and Buja et al. (1989), and using alternating conditional expectations (ACE) in Breiman and Friedman (1985). It consists of 330 (n) observations of one response variable (Y) and nine (p) independent variables each. The daily observations were recorded in Los Angeles in 1976. The variables are

Y : ozone concentration
X_1 : Vandenburg 500 millibar height
X_2 : windspeed
X_3 : humidity
X_4 : Sandburg Air Force Base temperature
X_5 : inversion base height
X_6 : Daggott pressure gradient
X_7 : inversion base temperature
X_8 : visibility
X_9 : day of the year

As in exploratory projection pursuit, all of the variables are standardized to have mean zero and variance one before projection pursuit regression is applied.

As suggested by Friedman (1984a), the original algorithm (Steps 1 and 2 in Section 2.2.3) is run for a large value of M initially. For this example, M is chosen to be nine (p). The algorithm produces all the submodels with number of terms $m = 1, \ldots, M$ by backstepping from the largest. The inaccuracy of each model is measured as the fraction of variance it cannot explain. As noted in the introduction, inaccuracy in this thesis denotes lack of fit as measured with respect to the data rather than to the population in general. From [2.6], this fraction is defined as

$$U \equiv \frac{L_2(\beta, \alpha, f, X, Y)}{\text{Var}\,(Y)} \quad .$$

The plot of the number of terms and fraction of unexplained variance of each model is shown in Fig. 2.1.

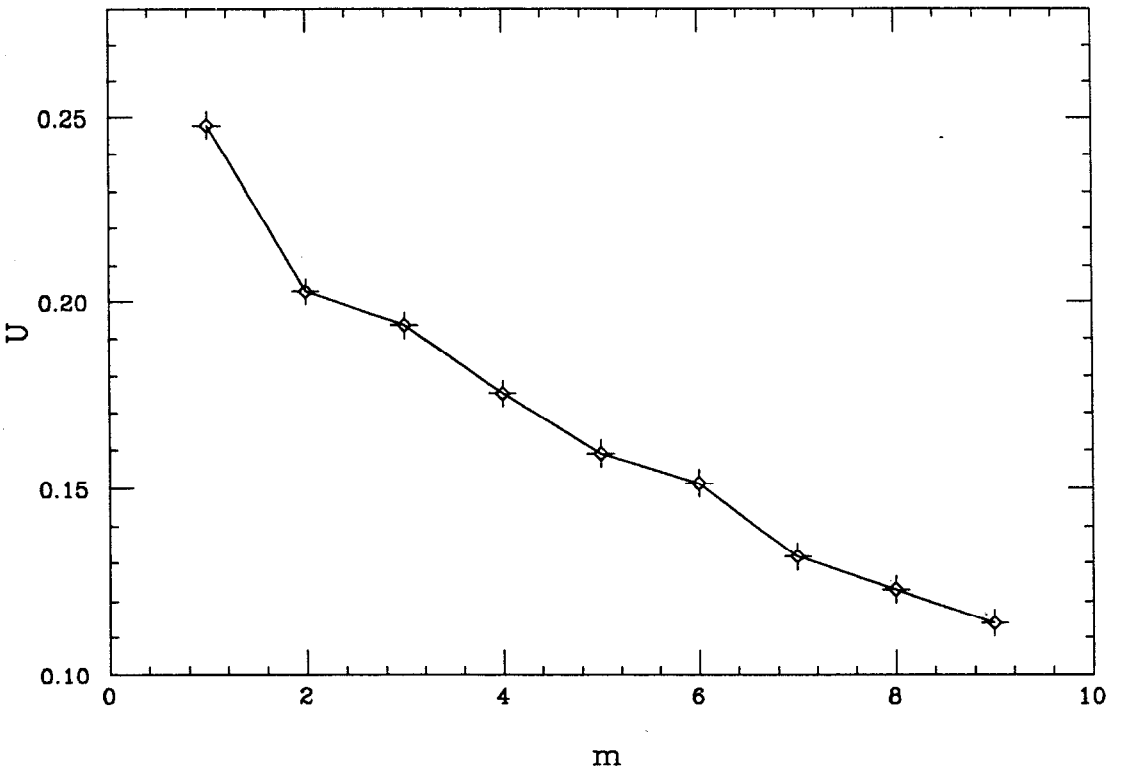


Fig. 2.1 Fraction of unexplained variance U versus number of terms m for the air pollution data. Slight 'elbows' are seen at $m = 2, 5, 7$.

Using the original approach, the statistician chooses the model from this plot by weighing the increase in accuracy against the additional complexity of adding a term. She generally chooses a model at an 'elbow' in the plot, where the marginal increase in accuracy due to the next term levels off. In some situations, such an elbow may not exist or it may not be a good model choice in actuality. Only one model for each number of terms m is found. For a particular m, the model space is one dimensional in U.

Interpretable projection pursuit regression expands the model space for a particular number of terms m to two dimensions by adding an interpretability measure. The starting point for each simplicity search for a given m is the model shown in Fig. 2.1. Then for a sequence of λ values which signify an increasing weight on simplicity, the algorithm cuts a path through the model plane [U X S_g].

Lubinsky and Pregibon (1988) discuss searching through a description or model space in general. Their premise is that a formalization of this action provides the structure by which such a search can be automated. Their descriptive space characterization, which is based on Mallows' (1983) work, is more comprehensive than the two dimensional U and S_g summary given above. They agree that two important description dimensions are accuracy and parsimony. In their work, the latter concept is an extension of the usual number of parameters measure and is in the spirit of interpretability as defined in this thesis. It includes both 'the conciseness of the description and its usefulness in conveying information.'

Initially for this and other examples, the interpretability parameter λ sequence is $(0.0, 0.1, \ldots, 1.0)$. However, the usual result is a path through the model space which consists of a few clumps of models separated by large breaks in the path. Even the forecasting nature of the algorithm described in Section 2.2.3 cannot completely eliminate these large hops between model groups. In order to produce a smoother path, the statistician is advised to run the algorithm with additional values of λ specifically chosen to produce a more continuous curve. For example, if on the first pass the path has a large hole between the $\lambda = 0.3$ and $\lambda = 0.4$ models, the algorithm should be run with additional λ values of $(0.33, 0.36, 0.39)$. Using this strategy, twenty models are produced for each value of $m = 1, \ldots, 9$. The actual λ values used are not shown in the following figures as their values are not important. The interpretability parameter is solely a guide for the algorithm through the model space.

Various diagnostic plots can be made of the collection of models which are distinguished by their m, U and S_g values. Chambers et al. (1983) provide several possibilities. Given that the number of terms variable m is discrete, a partitioning approach is used. Partitioning plots are shown in Figs. 2.2 and 2.3. Each point in a plot represents a model with the given number of terms, interpretability S_g and inaccuracy U. Ideally, for the best comparison, these plots should be

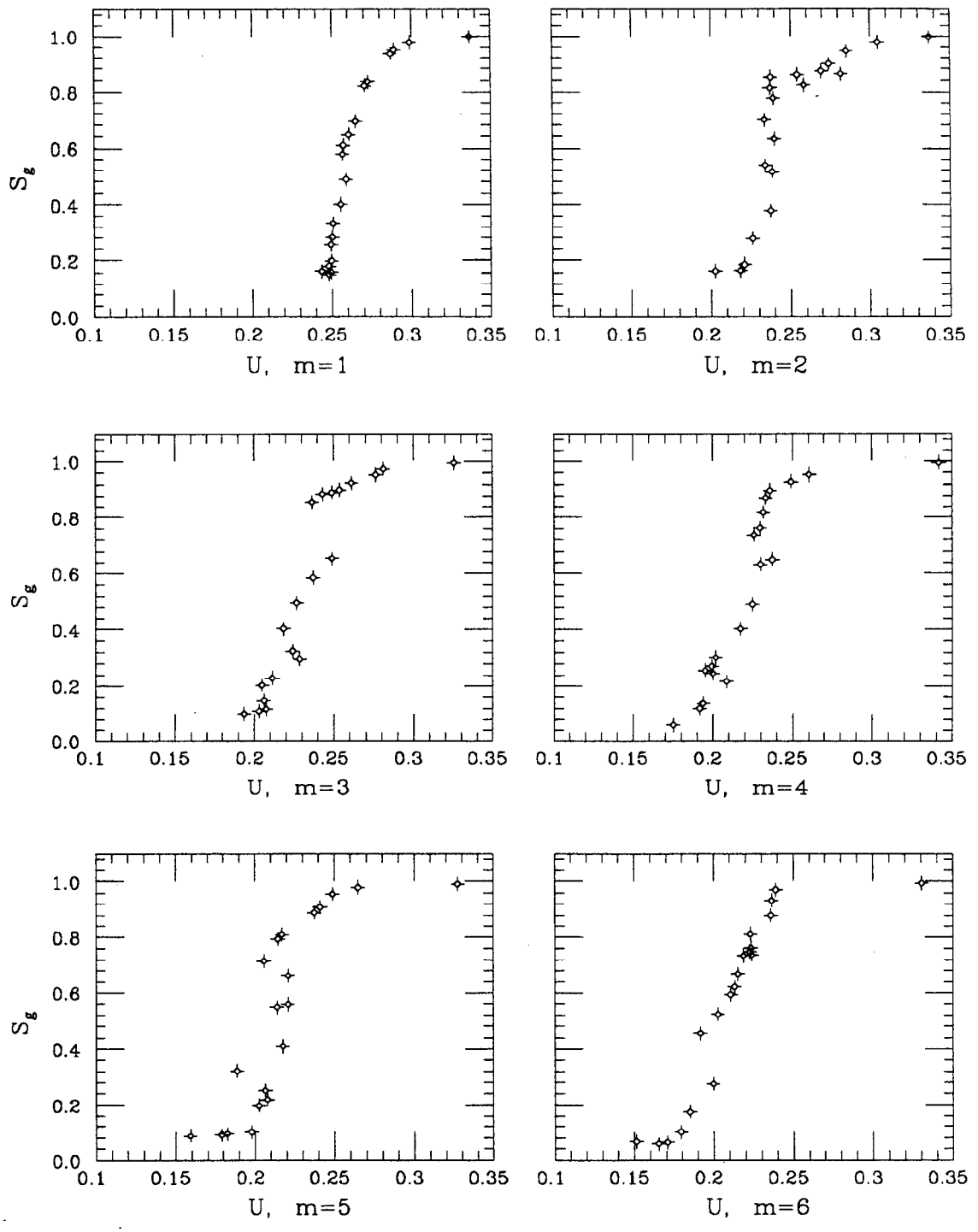


Fig. 2.2 Model paths for the air pollution data for models with number of terms $m = 1, \ldots, 6$. Each point indicates the interpretability S_g and fraction of unexplained variance U for a model with the given number of terms.

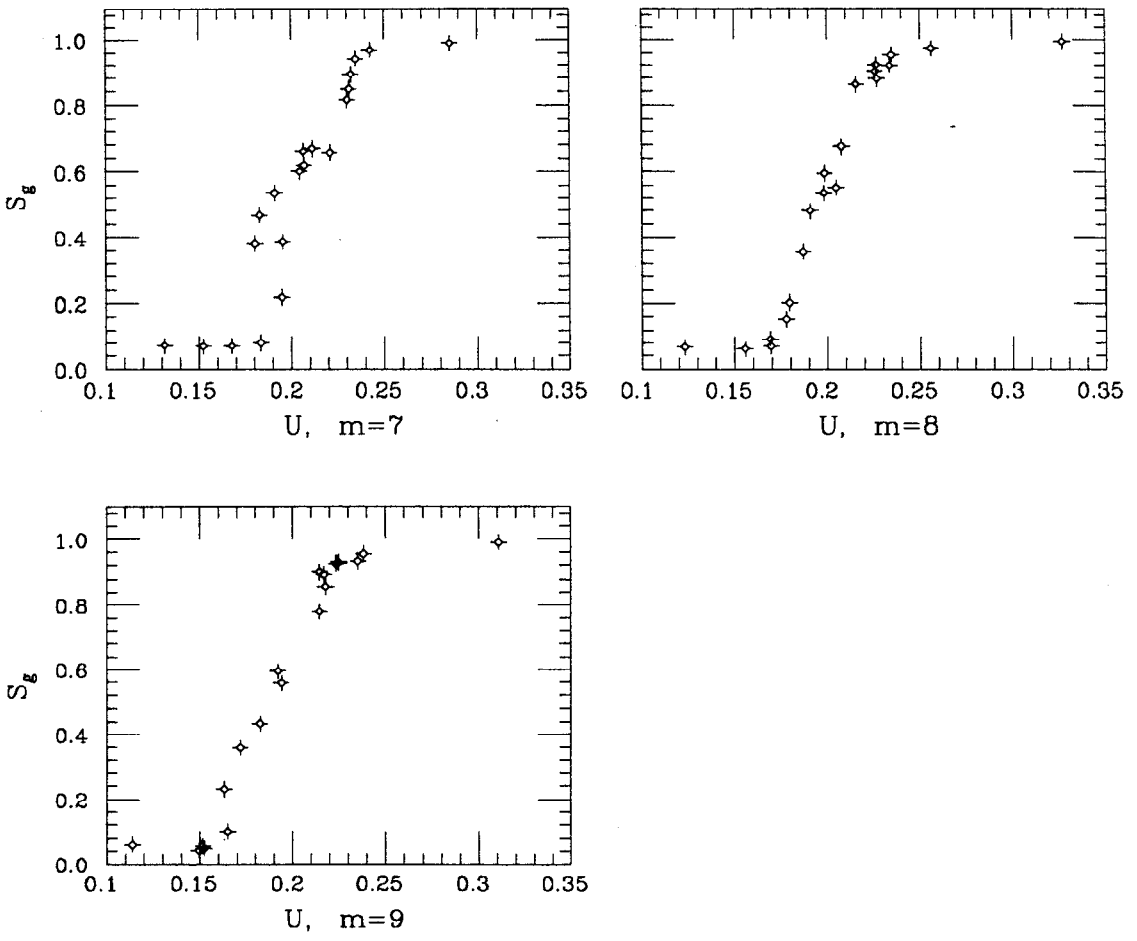


Fig. 2.3 Model paths for the air pollution data for models with number of terms $m = 7, 8, 9$. Each point indicates the interpretability S_g and fraction of unexplained variance U for a model with the given number of terms.

lined up side by side. However, note that though the interpretability S_g scales superficially appear to be the same for all the graphs, they are not as implicit in each plot is the number of terms m. A symbolic scatterplot in which all models are graphed in one plot of unexplained variance U versus interpretability S_g with a particular graphing symbol for the number of terms, also obscures the fact that the simplicity scales are dependent on the number of terms in the model.

As interpretability increases, so does the inaccuracy of a model. For most values of m, the path through the model is 'elbow-shaped', indicating that initially a large gain in interpretability is made for a small increase in inaccuracy. The curves shift to the left as m increases, as the additional terms decrease the overall inaccuracy of all possible models. For all values of m, the $\lambda = 1$ models have the same inaccuracy. These models have all directions parallel and equal to an e_j, so in effect they only have one term of one variable.

Due to the forecasting algorithm employed, the path through the model space is not always monotonic. Occasionally, a clearly worse model as evidenced by a non-monotonic move resulting in smaller interpretability and larger inaccuracy, is found. However, usually on the next step, the algorithm readjusts. This type of behavior is evident for $m = 3$, interpretability values $S_g \in [0.2, 0.4]$. The intermediate poor move may be needed to force the algorithm out of a local minimum. The algorithm does not find the global minimum for a particular value of λ, rather it helps describe the models which are possible for a given number of terms m. In contrast, linear regression variable selection methods find the model which minimizes inaccuracy for a fixed interpretability value, which is usually the number of parameters.

The draftsman's display in Fig. 2.4 shows how the number of terms m and the inaccuracy U and simplicity S_g vary. Again, note that if a model with fewer terms is considered simpler than one with more, plotting all the models versus the same interpretability scale is misleading. This set of plots is useful in determining the models from which to choose if certain requirements must be met, such as $U \leq 0.20$, $S_g \geq 0.50$, or some combination thereof. More formally, the statistician can define equivalence classes of models from the draftsman and partitioning plots consisting of sets of models with different number of terms which satisfy certain inaccuracy and interpretability criteria.

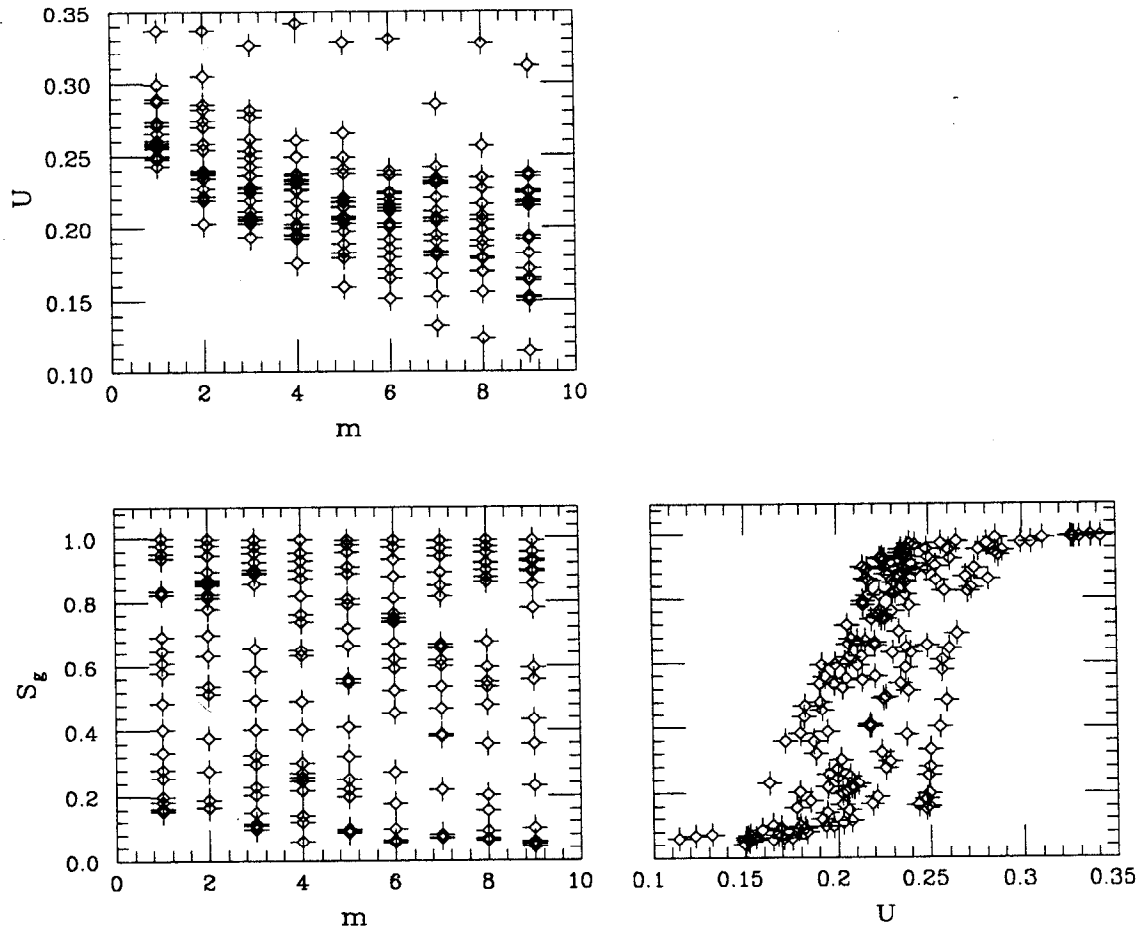


Fig. 2.4 Draftsman's display for the air pollution data All possible pairwise scatterplots of number of terms m, fraction of unexplained variance U and interpretability S_g are shown.

The statistician must now choose a model. The first term explains the bulk of the variance, approximately 75% ($U = 0.25$). For models with $m \leq 6$, models with even moderate interpretability ($S_g \geq 0.40$), cannot be achieved without inaccuracy crossing the 0.20 threshold ($U \geq 0.20$) as seen in Fig. 2.2. If a model which explains more than 80% of the variance is required, a seven term model with $S_g = 0.60$ is possible. However, its advantages over the original two term model (Fig. 2.1) which explains slightly less is debatable. If close to 0.25

inaccuracy U is acceptable, simple two or three terms models are possible. For example, a three term model exists with $S_g = 0.40$ and $U = 0.21$. Alternatively, a two term model exists with $S_g = 0.85$, $U = 0.23$ and combinations

$$\beta_1 = 0.21, \quad \alpha_1 = (\ 0.0,\ 0.0,\ 0.0,\ 1.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0)^T$$
$$\beta_2 = 1.10, \quad \alpha_2 = (\ 0.0,\ 0.0,\ 0.3,\ 0.9,\ 0.0,\ 0.0,\ 0.1,\ 0.0, -0.2)^T \quad .$$

As discussed in Sections 1.5.2 and 3.4.3, the convergence of the coefficients is slow and an interpretability index with a power less than two may be warranted.

The fourth variable, Sandburg Air Force Base temperature, is the most influential as is cited in other analyses of this data (Hastie and Tibshirani 1984, Breiman and Friedman 1985). The last variable day of the year also has an effect. The projection pursuit regression model is different in form to the additive and alternating conditional expectation models as it includes smooths of several variables rather than of one variable. Thus, the three models cannot be compared functionally. However, the three inaccuracy measures are comparable.

As remarked in Section 2.2.2, the inclusion of the number of terms in the gauging of a model's interpretability is subjective and depends on the context of the data. In addition, the statistician views the functions f_j when choosing among the models. Though each of the curves is a smooth and therefore does not have a functional description, one model may be more understandable and explainable than another. For example, a function may have a clear quadratic form. Similar to the weighing of the model's number of terms m, this qualitative assessment is a further subjective notion of interpretability which is not automated by the index S_g.

In contrast to exploratory projection pursuit, projection pursuit regression is a modeling procedure. As such, an objective measure of predictive error may be applied to choose between models. Methods such as cross-validation may be used to produce an unbiased estimate of the prediction accuracy in the models. A good strategy is to choose a small number of models based on the above procedure and

then distinguish between them using a resampling method. Unfortunately, due to a lack of identifiability which results from the fact that the number of terms cannot be included in the interpretability index, the cross-validation procedure is not a one parameter (λ) minimization problem. A subjective measure of how the model's complexity increases with an additional term must be made. Thus, interpretable projection pursuit regression is an exploratory rather than a strict modeling technique.

Chapter 3

Connections and Conclusions

A comparison of the accuracy and interpretability tradeoff approach described in the previous two chapters with other model selection techniques is warranted and interesting. In this chapter, connections between the proposed method and established ideas are considered. This discussion is preliminary and topics of future work, including other data analysis methods to which this method could be extended, are identified. The last section includes an example which demonstrates the generality of the trading accuracy for interpretability approach.

3.1 Interpretable Linear Regression

Since other model selection procedures have not yet been proposed for projection pursuit, the interpretable modification is considered for linear regression in this section. This setting provides various other model space search methods whose properties are known for comparison.

First, the notation for the linear regression problem is described. Rather than use random variable notation as in Chapter 2, matrix notation is used. The problem is stated as a minimization of the squared distance between the observed and fitted values rather than an expected value minimization. The vector Y consists of the response values for the n observations $(y_1, y_2, \ldots, y_n)$, and the n X p matrix X has entries x_{ij}, the value of the j^{th} predictor for the i^{th}

observation. If an intercept term is required, a column of ones is included. The error vector is ϵ and the model may be written as

$$Y = X\beta + \epsilon \quad .$$

The parameters $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ are estimated by minimizing the squared distance between the n fitted and actual observations. The problem in matrix form is

$$\min_{\beta} \ (Y - X\beta)^T(Y - X\beta) \quad .$$

The least squares estimates solve the normal equations

$$\hat{\beta}_{LS} \equiv (X^T X)^{-1} X^T Y \quad . \qquad [3.1]$$

The modification for values of the interpretability parameter $\lambda \in [0, 1]$, is

$$\min_{\beta} \ (1 - \lambda) \frac{(Y - X\beta)^T(Y - X\beta)}{(Y - X\hat{\beta}_{LS})^T(Y - X\hat{\beta}_{LS})} \ - \ \lambda S(\beta) \qquad [3.2]$$

where the interpretability index S is the single vector varimax index S_1 defined in [1.12] for example. The denominator of the first term is the minimum squared distance possible, which would be at the ordinary linear regression solution [3.1]. Recall that the combination of the predictors $X\beta$ that has maximum correlation with the response is the ordinary linear regression solution. Thus, if the correlation is used instead of squared distance as a measure of the model's fit, the problem becomes a maximization and the simplicity term should be added rather than subtracted.

As the interpretability parameter λ increases, the fitted vector $\hat{Y} = X\hat{\beta}$ moves away from the ordinary least squares fit $\hat{Y}_{LS}$ in the space spanned by the

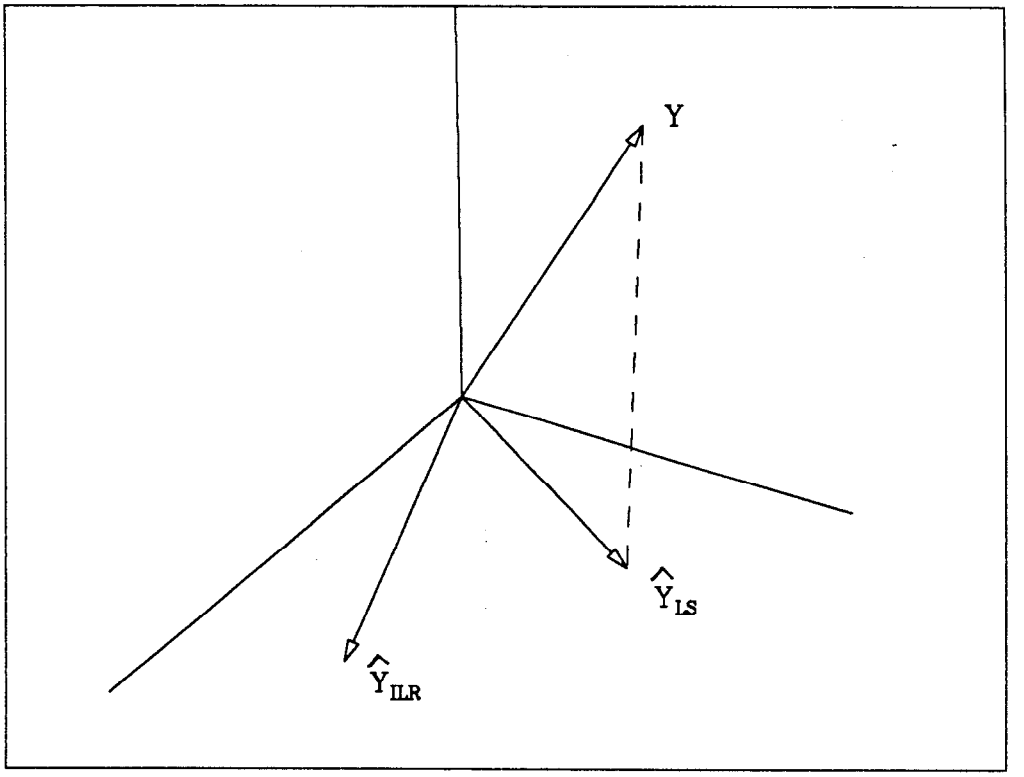


Fig. 3.1 Interpretable linear regression. As the interpretability parameter λ increases, the fit moves away from the least squares fit $\hat{Y}_{LS}$ to the interpretable fit $\hat{Y}_{ILR}$ in the space spanned by the p predictors.

p predictors as shown in Fig. 3.1. The interpretable fits $\hat{Y}_{ILR}$ are not necessarily the same length as the least squares one.

As a variable's coefficient decreases toward zero , the fit moves into the space spanned by a subset of $p-1$ predictors. The most interpretable $p-1$ variables may not be the best $p-1$ in a least squares sense. Even if the variable subset is the same, the interpretable search may not guide the statistician to the best fitting least squares model. The interpretability index S attempts to pull the coefficients β_i apart since a diverse group is considered more simple, whereas the least squares method considers only squared distance when choosing a model.

The definition of the interpretability index S as a smooth function means that it can be used as a 'cognostic' (Tukey 1983) to guide the automatic search for a more interpretable model. However, these smooth and differentiable properties are the root of the reason the equation [3.2] cannot be solved explicitly. The problem is that the interpretability index contains the squared and normed coefficients which make the objective a nonquadratic function. A linear solution similar to [3.1] is not possible.

In comparison, Mallows' (1973) C_p and Akaike's (1974) AIC variable selection criteria involve a count function as an interpretability index as defined in [1.11]. As [3.2] does not admit a linear solution, the more general Mallows' C_L criterion which may be used for any linear estimator cannot be determined for the interpretable estimate. For both the C_p and AIC criteria, the complexity of a model is equal to the number of variables in the model, say k. With appropriate norming, the associated interpretability index is

$$S_M(\beta) \equiv 1 - \frac{k}{p} \quad . \qquad [3.3]$$

The Mallows variable selection technique uses an unbiased estimate of the model prediction error to choose the best model. The resulting search through the model space includes only one model for a given number of variables. Alternatively, the interpretability search procedure is a model estimation procedure which guides the statistician in a nonlinear manner through the model space. Another result of this search is variable selection as variables are discarded for interpretability. Using this terminology, the discreteness of the criterion [3.3] means that the C_p and AIC techniques are solely variable, rather than model, selection procedures.

As is discussed in Section 1.3.4, an alternate interpretability index S_d can be defined as the negative distance from a set of simple points. The natural question is whether the modification [3.2] can be thought of as a type of shrinkage.

3.2 Comparison With Ridge Regression

In Section 1.3.4, a distance interpretability index S_d is defined in [1.15] which involves the distances to a set of simple points $V = \{\nu_1, \ldots, \nu_J\}$. Before restricting its values to be $\in [0, 1]$, this simplicity measure of the coefficient vector β is the negative of the minimum distance to V or

$$- \min_{j=1,\ldots,J} \sum_{i=1}^{p} \left(\frac{\beta_i^2}{\beta^T \beta} - \nu_{ji} \right)^2 \quad .$$

The coefficient vector β is squared and normed so that the relative mass, not the absolute size or sign, of the elements matters. Though these actions lead to a nonlinear solution, vectors of any length may be compared equally and properties such as Schur-convexity are possible. Suppose for the moment that the response Y and predictors X are standardized to have mean zero and variance one. This standardization ensures that one variable does not overwhelm the others in the coefficient vector though the length of the coefficient vector itself is not necessarily one. This action partially removes the reason for normalizing the coefficient vector in the interpretability index.

In order to remove the need for squaring, all possible sign combinations must be considered. The simple set vectors ν_j are no longer the squares of vectors on the unit R^p sphere but just vectors on it. For example, they could be $\pm e_j, j = 1, \ldots, p$. The distance to each simple point is written in matrix form as

$$(\beta - \nu_j)^T (\beta - \nu_j) \qquad j = 1, \ldots, p \quad .$$

If the optimization problem is reparameterized with a new interpretability parameter κ, [3.2] becomes

$$\min_{\beta} \; (Y - X\beta)^T (Y - X\beta) + \kappa \min_{j=1,\ldots,p} \; (\beta - \nu_j)^T (\beta - \nu_j) \qquad \kappa > 0 \quad .$$

The second minimum may be placed outside of the first since the first term does not involve ν_j resulting in

$$\min_{j=1,\ldots,p} \left[\min_{\beta} (Y - X\beta)^T(Y - X\beta) + \kappa(\beta - \nu_j)^T(\beta - \nu_j)\right] \quad . \qquad [3.4]$$

The solution vector is

$$\hat{\beta} \equiv (X^TX + \kappa I)^{-1}(X^TY + \kappa\nu_l) \qquad [3.5]$$

where I is the p X p identity matrix and l is the index which minimizes the bracketed portion of [3.4].

This estimated vector is similar to that of ridge regression (Thisted 1976, Draper and Smith 1981) with ridge parameter κ. The ridge estimate is

$$\hat{\beta}_R \equiv (X^TX + \kappa I)^{-1}X^TY \quad . \qquad [3.6]$$

Ridge regression may be examined from either a frequentist or Bayesian viewpoint. The method is advised in situations where the matrix X^TX is unstable, which can occur when the variables are collinear. In addition, it does better than linear regression in terms of mean square error as it allows biased estimates.

If a Bayesian analysis is used, the distributional assumption is made that the error term ϵ has independent components each with mean zero and variance σ^2. Given this assumption, the prior belief

$$\beta \sim N(0, \frac{\sigma^2}{p}I)$$

produces the Bayes rule [3.6].

The ridge estimate $\hat{\beta}_R$ [3.6] shrinks away from the least squares solution $\hat{\beta}_{LS}$ [3.1] toward the origin as the ridge parameter κ increases. The interpretability

estimate $\hat{\beta}$ [3.5] shrinks away from the least squares solution toward the simple point ν_l as the interpretability parameter κ increases. The index l may change during the procedure however.

If the varimax index [1.12] is used instead of the distance index [1.15], the result is a solution similar to [3.5] except that the shrinkage is away from the set of 2^p points $(\pm\frac{1}{\sqrt{p}}, \pm\frac{1}{\sqrt{p}}, \ldots \pm\frac{1}{\sqrt{p}})$. Shrinkage toward points is the usual ridge regression terminology, so the distance index is used in the explanation above.

As Draper and Smith (1981) point out, ridge regression places a restriction on the size of the coefficients β, whether or not that restriction is called a prior. The interpretability approach [3.2] does not require any restriction as it norms and squares the coefficients before examining them. As noted, however, this produces mathematical problems. Clearly, the approach can be viewed as placing a prior on the coefficients and this Bayesian viewpoint is discussed in the next section.

3.3 Interpretability as a Prior

The least squares solution is also the maximum likelihood solution given certain conditions. The necessary assumptions are that the elements of the error term ϵ are independent and identically distributed $N(0, \sigma^2)$. Then the squared distance is the negative of the log likelihood or

$$-\log L(\beta; Y) = (Y - X\beta)^T(Y - X\beta) \quad .$$

The maximum likelihood estimate minimizes this expression. Given a prior distribution, the posterior distribution is equal to the likelihood multiplied by the prior. Thus, as Good and Gaskins (1971) point out, minimizing the reparameterized [3.2]

$$(Y - X\beta)^T(Y - X\beta) - \kappa S(\beta)$$

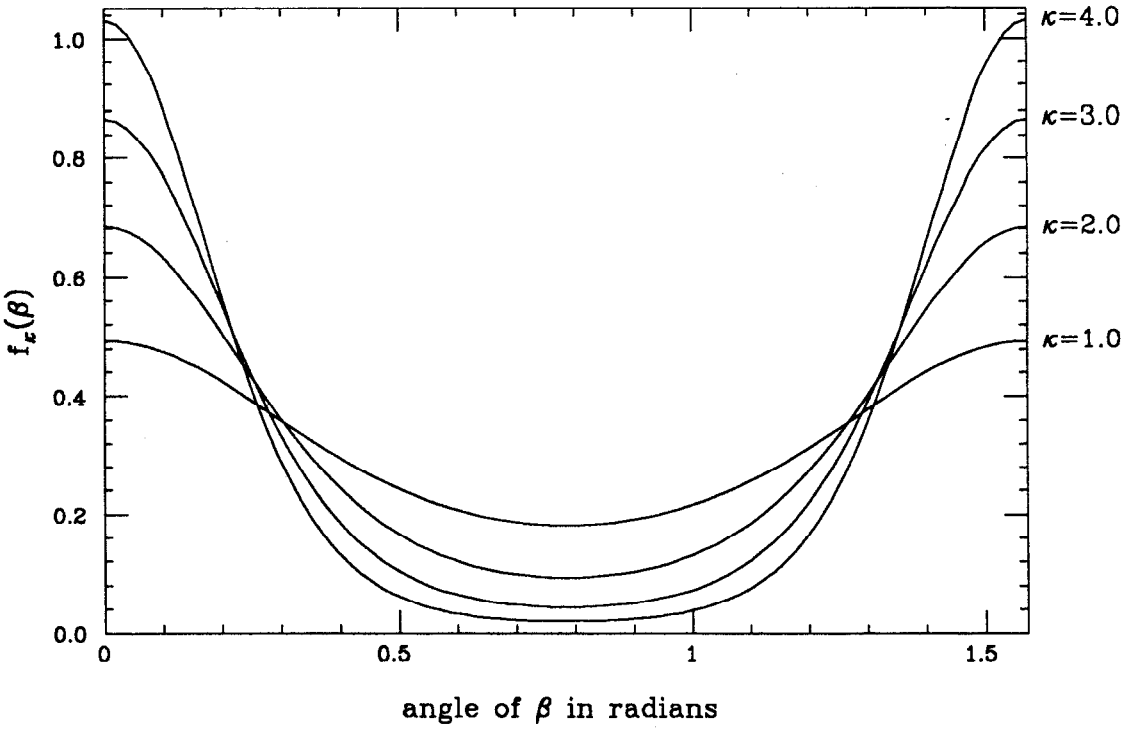


Fig. 3.2 Interpretability prior density for $p = 2$. The prior $f_\kappa(\beta)$ is plotted versus the angle of the coefficient vector in radians for various values of κ.

is equivalent to minimizing the negative log likelihood minus the log prior. To do so is to put a prior density on β where the prior is proportional to

$$\exp(\kappa S(\beta)) \qquad \kappa > 0 \quad .$$

For the varimax interpretability index, the prior distribution of the coefficients for a given interpretability parameter value $\kappa > 0$ is

$$f_\kappa(\beta) \equiv C_\kappa \exp\left[\kappa \frac{p}{p-1} \sum_{i=1}^{p} \left(\frac{\beta_i^2}{\beta^T\beta} - \frac{1}{p}\right)^2\right]$$

where C_κ is the normalizing constant. The coefficients are dependent. The density belongs to the general exponential family defined by Watson (1983).

The prior for the $p = 2$ case is plotted in Fig. 3.2. The constant C_κ is calculated using numerical integration. Since the exponential is a monotonic function, the basic shape of the curve resembles the index S_1 as in Fig. 1.2. As κ increases, the prior becomes more steep as the weight on interpretability increases. The coefficients are pushed toward an angle of zero ($\beta = (1, 0)$) or an angle of $\frac{\pi}{2}$ ($\beta = (0, 1)$). Similar results would be seen for $p = 3$.

3.4 Future Work

The previous sections may serve as a foundation for the comparison of the interpretable method with other variable selection techniques. Further work connecting this approach with others is proposed below. In addition, ideas for both extending the method to other data analysis methods and improving the algorithm are given.

3.4.1 Further Connections

Schwarz (1978) and C. J. Stone (1981, 1982) suggest other model selection techniques based on Bayesian assumptions. M. Stone (1979) asymptotically compares the Schwarz and Akaike (1974) criteria, noting that the comparison is affected by the type of analysis used. In the Bayesian framework, the interpretable technique could be compared with these others utilizing asymptotic analysis. Another technique which may provide an interesting comparison is that of Rissanen (1987), who applies the minimum description length principle from coding theory to measure the complexity of a model. In addition, the work by Copas (1983) on shrinkage in stepwise regression may provide ways to extend the ridge regression discussion of Section 3.2. Interestingly enough, all the model selection rules noted involve the number of parameters, k in [3.3], rather than a smooth measure of complexity.

In Chapter 2, the varimax and entropy indices have similar intuitive and computational appeal and both have historical motivation. Interpretability as

measured by these indices could be compared with simplicity as defined using philosophical terminology in Good (1968), Sober (1975) and Rosenkrantz (1977). Though the framing of the interpretable approach as the placing of a prior on the coefficients attempts to do this, a less mathematical and more philosophical discussion may prove beneficial.

3.4.2 Extensions

The interpretable approach changes the usual combinatorial model space search into a numerical one. It can be applied to any data analysis method whose resulting description or model involves linear combinations. If the index is extended to measure the simplicity of other description types such as functions, the tradeoff of interpretability and accuracy might prove useful for even more complicated descriptions.

At present, however, its main improvement is that it provides a model selection procedure for methods that produce linear combinations but for which feasible variable selection approaches do not exist. For linear regression, all subsets regression is possible for two reasons. First, the least squares solution for any subset of predictors is known analytically to be [3.1]. Second, branch and bound search methods can be employed which smartly search the model space, eliminating areas so that all possible contenders need not be considered.

Though linear regression variable selection procedures exist, interpretable linear regression may help in collinear situations. The least squares solution is unstable and in fact, ridge regression is usually suggested. At present, examples seem to indicate that interpretable linear regression clearly chooses the variables to include from a group of collinear ones and produces a stable solution.

A general class of models for which the interpretable method may prove useful is generalized linear models (McCullagh and Nelder 1983), of which logistic regression is an example. Whenever a generalized linear model is considered, a separate optimization must be done. At present, stepwise methods are used

to choose models. These methods could be compared with the interpretable approach on the basis of prediction error.

3.4.3 Algorithmic Improvements

As described in Chapter 1, the interpretable exploratory projection pursuit algorithm would benefit from further improvements. First, the rotationally invariant Fourier projection index G_F needs further testing and comparison with others. Other rotationally invariant projection indices are possible as suggested in Section 1.4.2.

Second, present work involves designing a procedure to solve the constrained optimization instead of using the general routine as outlined in Section 1.4.5. The improvement should decrease the computational time involved. Analogously, the projection pursuit regression forecasting procedure described in Section 2.2.3 needs further investigation.

As mentioned in Chapters 1 and 2, the convergence of the coefficients to zero as the weight on interpretability increases is slow. This property results because of the squaring used in the varimax index and the result that the slope goes to zero as the combinations moves to a maximizing e_i (Figs. 1.2 and 1.3). Solutions might be to use a lower power or a piecewise function which is the varimax index except for a range close to the e_i's where it is linear. The sections of the latter index could be matched to have the same derivatives at their joined points.

3.5 A General Framework

The main result of this thesis is the computer implementation of a modification of projection pursuit which makes the results more understandable. In addition, the approach provides a general framework for tackling similar problems. Beyond this specific application, the search for interpretability at the expense of accuracy is made often in statistics, sometimes implicitly. The identification and

formalization of this action is useful since choices which previously were subjective become objective. In the next subsection, the rounding of the binwidth for a histogram is examined. This common example shows that the consequences of a simplifying action in terms of accuracy loss can be approximated. The definition of interpretability must be broadened to deal with a much more elusive set of outcomes than linear combinations. Measuring the interpretability increase is difficult.

3.5.1 The Histogram Example

Consider the problem of drawing a histogram of n observations $x_1, x_2, \ldots, x_n$. Two quantities must be determined. The first is the left endpoint of the first bin x_0, which usually is chosen so that no observations fall on the boundaries of a bin. The second is the binwidth h, which generally is calculated according to a rule of thumb and then rounded so that the resulting intervals are simple, usually multiples of powers of ten. Based on the mathematical approach used to determine the rules widely employed, the loss in accuracy incurred by rounding the binwidth can be calculated.

The purpose of a histogram is to estimate the shape of the true underlying density. Scott (1979) determines a binwidth rule which asymptotically minimizes the integrated mean square error of the histogram density from the true density. Diaconis and Freedman (1981) use the same criterion to lead to a slightly different rule and further theoretical results. A certain approximation step which Scott employs is useful in approximating the further accuracy lost if the binwidth is rounded. The discussion below follows his approach.

The integrated mean square error of the estimated histogram density $\hat{f}$ from the true density f is

$$\begin{aligned} IMSE &\equiv \int_{-\infty}^{\infty} E\Big[\hat{f}(x) - f(x)\Big]^2 dx \\ &= \frac{1}{nh} + \frac{1}{12}h^2 \int_{-\infty}^{\infty} f'(x)^2 dx + O\Big(\frac{1}{n} + h^3\Big) \end{aligned} \qquad [3.7]$$

where h is the histogram binwidth. Minimizing the first two terms of [3.7] produces the estimate

$$\hat{h} \equiv \Big(\frac{6}{\int f'(x)^2 dx}\Big)^{\frac{1}{3}} n^{-\frac{1}{3}} \quad . \qquad [3.8]$$

Scott also shows that if the binwidth is multiplied by a factor $c > 0$, the increase in $IMSE$ is

$$IMSE(c\hat{h}) = \frac{c^3 + 2}{3c} IMSE(\hat{h}) \quad . \qquad [3.9]$$

Via Monte Carlo studies, Scott shows [3.8] and [3.9] to be good approximations for normal data.

In reality, [3.8] is useless since the underlying density f and therefore its derivative f' are unknown. Scott, and Diaconis and Freedman use the normal density as a reference distribution. Scott suggests the approximation

$$\hat{h}_S \equiv 3.49 s n^{-\frac{1}{3}}$$

where s is the estimated standard deviation of the data. Diaconis and Freedman suggest the similar approximation

$$\hat{h}_D \equiv 2IQn^{-\frac{1}{3}}$$

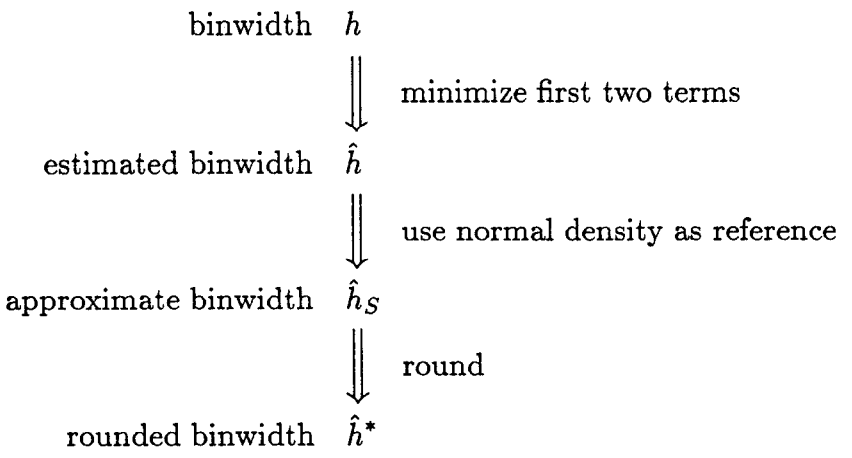
where IQ is the interquartile range of the data. Monte Carlo studies have shown these approximations to be robust.

Given either approximation $\hat{h}_S$ or $\hat{h}_D$, the statistician may elect to further approximate the binwidth by rounding. The benefits of such simplification are discussed in a moment. At present, consider rounding the estimate $\hat{h}_S$. The new estimate $\hat{h}^*$ is $\in (\hat{h}_S - u, \hat{h}_S + u)$ where u is some rounding unit. For example, u would be $\frac{1}{2}$ if the binwidth is rounded to an integer. The new binwidth may also be written as

$$\hat{h}^* \equiv (1 + e)\hat{h}_S$$

where e is a positive or negative factor depending on whether the old estimate is rounded up or down. This multiplying factor must be ≥ 1 as a negative binwidth is impossible.

The estimation procedure may be drawn schematically:



If [3.9] is used as an approximation for the resulting loss in $IMSE$ due to rounding, the relationship between the $IMSE$ of $\hat{h}_S$ and $\hat{h}^*$ is written

$$IMSE(\hat{h}^*) \approx \frac{(1+e)^3+2}{3(1+e)} IMSE(\hat{h}_S) \quad .$$

The percent change in $IMSE$ can be plotted as a function of the multiplying factor e as shown in Fig. 3.3.

This exercise demonstrates that rounding a binwidth up or down results in different repercussions in terms of accuracy. The increase in interpretability is difficult to measure explicitly. The histogram is easier to draw and describe since the class boundaries are simpler. Certainly the class delineations are easier to remember. In fact, Ehrenberg (1979) shows that two digits other than zero are all that can be retained in short-term memory. In addition, the rounding removes confusion that might result in explaining the histogram or comparing it to another. Finally, the accuracy in the actual observations x_i may prompt

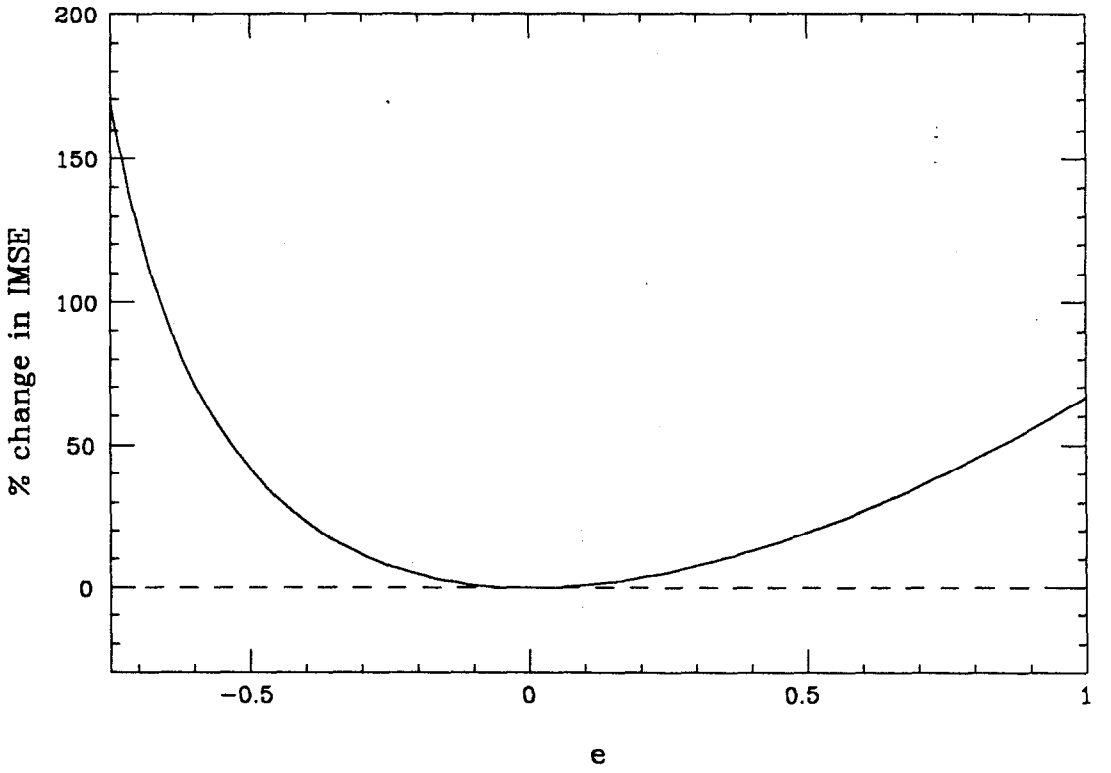


Fig. 9.9 Percent change in $IMSE$ versus multiplying fraction e in the binwidth example. The rounded binwidth $\hat{h}^* \equiv (1+e)\hat{h}_S$, where $\hat{h}_S$ is the estimated binwidth due to Scott (1979).

rounding. Extra digits beyond the number of significant ones in the data leads to a false sense of accuracy.

How to measure the interpretability of a histogram directly is difficult to determine. Diaconis (1987) suggests conducting experiments to quantify the interpretability gain. A typical experiment might be to divide a statistics class into two matched groups and to present to either group respectively a unrounded histogram and its simplified version. Measurements of interpretability could be made on the basis of the correctness of answers to questions such as 'How would adding the following observations change the shape of the histogram?' or 'Where is the lower quartile?'. As interpretability is increased, some information may be lost. For example, the question 'Where is the mode?' may become unanswerable.

This loss of information, a more general measure of inaccuracy than $IMSE$, could be measured along with interpretability. In addition, expert opinion could be included by asking data analysts their reaction to rounding.

3.5.2 Conclusion

Computers have changed statistics substantially and irreversibly. On the one hand, the resulting flexibility has cultivated previously undreamed of abilities and applications. On the other, the sheer number and complexity of possibilities can be both bewildering and unmanageable. The aim of this thesis is to use the principle of parsimony to monitor the sacrifice of an acceptable amount of flexibility in return for more interpretable results in a particular computer-intensive technique, projection pursuit. In this manner, the statistician retains her new modes of information translation without losing the ability to achieve her basic goal of clearly understanding and communicating those results to others.

With these computer tools has come freedom. In the initial stages of an analysis, these abilities provide the power with which to follow the basic tenets of exploratory data analysis (Tukey 1977), to let the data drive the analysis rather than subjecting it to preconceived assumptions which may not be true. In later stages, the wealth of models or descriptions which can be fit and evaluated has expanded enormously. Confirmation of these models can be readily answered using the bootstrap (Efron 1979) or other resampling procedures. As Tukey noted, the statistician no longer answers 'What can be confirmed?' but rather 'What can be done?'.

Computing power is extending and eliminating previous mathematical, computational and confirmational boundaries. Even unconscious restrictions on the statistician's imagination are alleviated (McDonald and Pedersen 1985). The results of such an analysis can be complex, hard to understand, and even more important, hard to explain. Though this progress is exciting, in a sense a Pandora's box has been opened. Just as grappling with the theoretical demons of

new methods such as projection pursuit (Huber 1985) is a necessary and difficult task, so too is considering the parsimonious aspects.

In order to understand and communicate the results of a statistical analysis effectively, a controlled use of these new methods is helpful. Fortunately, the very computing power which has produced these novel techniques provides a means to balance the search for an accurate and truthful description of the data with an equally important desire for simplicity. Interpretable projection pursuit strikes such a balance.

Appendix A

Gradients

A.1 Interpretable Exploratory Projection Pursuit Gradients

In this section, the gradients for the interpretable exploratory projection pursuit objective function

$$F(\beta_1^T Y, \beta_2^T Y) \equiv (1-\lambda)\,\frac{G_F(\beta_1^T Y, \beta_2^T Y)}{\max G_F} \;+\; \lambda\, S_v(\beta_1, \beta_2) \qquad [A.1]$$

are calculated. Since the search procedure is conducted in the sphered space, the desired gradients are

$$\frac{\partial F}{\partial \alpha_j} = (\frac{\partial F}{\partial \alpha_{j1}}, \frac{\partial F}{\partial \alpha_{j2}}, \ldots, \frac{\partial F}{\partial \alpha_{jp}})^T \qquad j = 1, 2 \quad .$$

From [A.1], the gradients may be written in vector notation as

$$\frac{\partial F}{\partial \alpha_j} = \frac{(1-\lambda)}{\max G_F}\,\frac{\partial G_F}{\partial \alpha_j} + \lambda \frac{\partial S_v}{\partial \alpha_j} \qquad j = 1, 2 \quad .$$

The Fourier projection index gradients are calculated from [1.23], yielding

$$\begin{aligned}\frac{\partial G_F}{\partial \alpha_j} &= \sum_{i=0}^{\infty} E_p[l_i(R)]E_p[l_i'(R)\frac{\partial R}{\partial \alpha_j}] \\ &\quad + 2\sum_{i=0}^{\infty}\sum_{k=1}^{\infty}\Big[E_p[l_i(R)\cos(k\Theta)](E_p[l_i'(R)\cos(k\Theta)\frac{\partial R}{\partial \alpha_j}] \\ &\qquad - k\sin(k\Theta)E_p[l_i(R)\frac{\partial \Theta}{\partial \alpha_j}]) \\ &\qquad + E_p[l_i(R)\sin(k\Theta)](E_p[l_i'(R)\sin(k\Theta)\frac{\partial R}{\partial \alpha_j}] \\ &\qquad + k\cos(k\Theta)E_p[l_i(R)\frac{\partial \Theta}{\partial \alpha_j}])\Big] \\ &\quad - \frac{1}{2}E_p[l_0'(R)\frac{\partial R}{\partial \alpha_j}] \qquad j = 1,2 \quad .\end{aligned} \tag{A.2}$$

From the definition of the Laguerre functions [1.20],

$$l_i'(R) = \frac{\partial L_i}{\partial R}e^{-\frac{1}{2}R} - \frac{1}{2}e^{-\frac{1}{2}R} \qquad i = 0,1,\ldots \quad ,$$

with recursive equations derived from the definition of the Laguerre polynomials [1.21]

$$\begin{aligned}\frac{\partial L_0}{\partial R} &= 0 \\ \frac{\partial L_1}{\partial R} &= -1 \\ \frac{\partial L_2}{\partial R} &= u - 2 \\ \frac{\partial L_i}{\partial R} &= \Big(\frac{2i-1}{i} - \frac{1}{i}R\Big)\frac{\partial L_{i-1}}{\partial R} - \frac{1}{i}L_{i-1} - \Big(\frac{i-1}{i}\Big)\frac{\partial L_{i-2}}{\partial R} \qquad i = 3,\ldots \quad .\end{aligned}$$

The gradients of the radius squared R and angle Θ are calculated using the definition [1.20]. If $X_1 \equiv \alpha_1^T Z$ and $X_2 \equiv \alpha_2^T Z$, the gradients are

$$\begin{aligned}\frac{\partial R}{\partial \alpha_j} &= (2X_j)Z \qquad j = 1,2 \\ \frac{\partial \Theta}{\partial \alpha_1} &= -\frac{X_2}{X_1^2}\Big(\frac{1}{1+(\frac{X_2}{X_1^2})}\Big)Z \\ \frac{\partial \Theta}{\partial \alpha_2} &= \frac{1}{X_1}\Big(\frac{1}{1+(\frac{X_2}{X_1^2})}\Big)Z \quad .\end{aligned} \tag{A.3}$$

In [A.3], note that Z is a vector $\in R^p$, while X_1 and X_2 are scalars. As with the calculation of the value of the index, the expected values are approximated by sample means over the data.

The gradients of the simplicity index are calculated using the index definition [1.27], the orthogonally translated component β_2^* definition [1.25], and the mapping from the unsphered to the sphered space [1.5]. The gradients may be written in matrix notation as

$$\begin{aligned} \frac{\partial S_v}{\partial \alpha_1} &= \frac{\partial S_v}{\partial \beta_1}\frac{\partial \beta_1}{\partial \alpha_1} + \frac{\partial S_v}{\partial \beta_2^*}\frac{\partial \beta_2^*}{\partial \beta_1}\frac{\partial \beta_1}{\partial \alpha_1} \\ \frac{\partial S_v}{\partial \alpha_2} &= \frac{\partial S_v}{\partial \beta_2^*}\frac{\partial \beta_2^*}{\partial \beta_2}\frac{\partial \beta_2}{\partial \alpha_2} \quad . \end{aligned} \qquad [A.4]$$

In [A.4], note that the partial derivative of one vector with respect to another is a p X p matrix. For example,

$$\left\{\frac{\partial \beta_1}{\partial \alpha_1}\right\}_{rs} = \frac{\partial \beta_{1r}}{\partial \alpha_{1s}} \qquad r,s = 1,\ldots,p \quad .$$

Using [1.5] yields

$$\frac{\partial \beta_{jr}}{\partial \alpha_{js}} = \frac{U_{rs}}{\sqrt{D_s}} \qquad r,s = 1,\ldots,p$$

where U and D are the eigenvector and eigenvalue matrices defined in [1.3]. Using [1.25] yields

$$\begin{aligned} \frac{\partial \beta_{2r}^*}{\partial \beta_{1s}} &= -\beta_{1r}\left(\frac{\beta_{2s}(\beta_1^T\beta_1) - 2\beta_{1s}(\beta_1^T\beta_2)}{(\beta_1^T\beta_1)^2}\right) - \frac{\beta_{1r}}{\beta_1^T\beta_1}\left(\beta_{2s} - 2\beta_{1s}\frac{(\beta_1^T\beta_2)}{(\beta_1^T\beta_1)}\right) \\ \frac{\partial \beta_{2r}^*}{\partial \beta_{2s}} &= -\frac{\beta_{1r}}{\beta_1^T\beta_1}(\beta_{1s}) \qquad r,s = 1,\ldots,p \text{ and } r \neq s \quad . \end{aligned}$$

The diagonal elements are

$$\begin{aligned} \frac{\partial \beta_{2r}^*}{\partial \beta_{1r}} &= -\left(\frac{\beta_1^T\beta_2}{\beta_1^T\beta_1}\right) - \frac{\beta_{1r}}{\beta_1^T\beta_1}\left(\beta_{2r} - 2\beta_{1r}\frac{(\beta_1^T\beta_2)}{(\beta_1^T\beta_1)}\right) \\ \frac{\partial \beta_{2r}^*}{\partial \beta_{2r}} &= 1 - \frac{\beta_{1r}^2}{\beta_1^T\beta_1} \qquad r = 1,\ldots,p \quad . \end{aligned}$$

The two vector varimax simplicity index S_v [1.12] may be written as a combination of the individual simplicities and a cross-term denoted as C yielding

$$S_v(\beta_1,\beta_2) = \frac{1}{2p}[(p-1)S_1(\beta_1) + (p-1)S_1(\beta_2) + 2] - C(\beta_1,\beta_2) \quad .$$

Taking partial derivatives yields

$$\frac{\partial S_v}{\partial \beta_{1r}} = 2\frac{\beta_{1r}}{\beta_1^T\beta_1}\left[\frac{\beta_{2r}^2}{\beta_2^T\beta_2} - \frac{1}{p} - (\frac{p-1}{p})S_1(\beta_1) - \frac{\beta_{1r}^2}{\beta_1^T\beta_1} + C(\beta_1,\beta_2)\right] \qquad r = 1,\ldots,p \quad .$$

The partial with respect to β_{2r}^* is identical with β_2^* components replacing β_1 components.

A.2 Interpretable Projection Pursuit Regression Gradients

In this section, the gradients for the interpretable projection pursuit regression objective function

$$F(\beta,\alpha,f,X,Y) \equiv (1-\lambda)\,\frac{L_2(\beta,\alpha,f,X,Y)}{\min L_2} \;-\; \lambda\, S_v(\alpha_1,\alpha_2,\ldots,\alpha_p) \qquad [A.5]$$

are calculated. The gradients used in the steepest descent search for directions α_j are

$$\frac{\partial F}{\partial \alpha_j} = (\frac{\partial F}{\partial \alpha_{j1}}, \frac{\partial F}{\partial \alpha_{j2}}, \ldots, \frac{\partial F}{\partial \alpha_{jp}})^T \qquad j = 1,\ldots,m \quad .$$

From [A.5], the gradients may be written in vector notation as

$$\frac{\partial F}{\partial \alpha_j} = \frac{(1-\lambda)}{\max G}\,\frac{\partial L_2}{\partial \alpha_j} + \lambda\frac{\partial S_v}{\partial \alpha_j} \qquad j = 1,\ldots,m \quad .$$

Friedman (1985) calculates the L_2 distance gradients from [2.7], yielding

$$\frac{\partial L_2}{\partial \alpha_j} = -2E[R_j - \beta_j f_j(\alpha_j^T X)]\beta_j f_j'(\alpha_j^T X)X \qquad j = 1,\ldots,m \quad .$$

The partials of the curves f_j are estimated using interpolation.

The gradients of the simplicity index are calculated using the index definition [2.13]. Taking partial derivatives yields

$$\frac{\partial S_v}{\partial \alpha_{ji}} = \frac{2p}{m(p-1)} \sum_{k=1}^{p} \left(\frac{\gamma_k}{m} - \frac{1}{p}\right) \frac{\partial \gamma_k}{\partial \alpha_{ji}} \qquad j = 1, \ldots, m \text{ and } i = 1, \ldots, p \quad .$$

The partials of the overall vector γ [2.12] are

$$\frac{\partial \gamma_k}{\partial \alpha_{ji}} = \frac{-2\alpha_{ji}\alpha_{jk}^2}{(\alpha_j^T \alpha_j)^2} \qquad k \neq i$$
$$\frac{\partial \gamma_k}{\partial \alpha_{jk}} = \frac{2\alpha_{jk}(\alpha_j^T \alpha_j - \alpha_{jk}^2)}{(\alpha_j^T \alpha_j)^2}$$

for $j = 1, \ldots, m$ and $i = 1, \ldots, p$.

References

Akaike, H. (1974). "A new look at the statistical model identification," *IEEE Transactions on Automatic Control* **AC-19**, 716-723.

Asimov, D. (1985). "The Grand Tour: A tool for viewing multidimensional data," *SIAM Journal of Scientific and Statistical Computing* **6**, 128-143.

Bellman, R. E. (1961). *Adaptive Control Processes*, Princeton University Press, Princeton.

Breiman, L. and Friedman, J. H. (1985). "Estimating optimal transformations for multiple regression and correlation (with discussion)," *Journal of the American Statistical Association* **80**, 580-619.

Buja, A., Hastie, T. and Tibshirani, R. (1989). "Linear smoothers and additive models (with discussion)," *Annals of Statistics* **17**, 453-555.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Wadsworth, Boston.

Copas, J. B. (1983). "Regression, prediction and shrinkage (with discussion)," *Journal of the Royal Statistical Society, Series B* **45**, 311-354.

Dawes, R. M. (1979). "The robust beauty of improper linear models in decision making," *American Psychologist* **34**, 571-582.

Diaconis, P. (1987). Personal communication.

Diaconis, P. and Freedman, D. (1981). "On the histogram as a density estimator: L_2 theory," *Zeitschrift fuer Wahrscheinlichkeitstheorie und Verwandte Gebiete* **57**, 453-476.

Diaconis, P. and Shahshahani, M. (1984). "On nonlinear functions of linear combinations," *SIAM Journal of Scientific and Statistical Computing* **5**, 175-191.

Donoho, D. L. and Johnstone, I. M. (1989). "Projection-based approximation and a duality with kernel methods," *Annals of Statistics* **17**, 58-106.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis*, Wiley, New York

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, CMBS 38, SIAM-NSF, Philadelphia.

Efron, B. (1988). "Computer-intensive methods in statistical regression," *SIAM Review* **30**, 421-449.

Ehrenberg, A. S. C. (1981). "The problem of numeracy," *Journal of the American Statistical Association* **35**, 67-71.

Friedman, J. H. (1984a). "SMART user's guide," Technical Report LCS001, Department of Statistics, Stanford University.

Friedman, J. H. (1984b). "A variable span smoother," Technical Report LCS005, Department of Statistics, Stanford University.

Friedman, J. H. (1985). "Classification and multiple regression through projection pursuit," Technical Report LCS012, Department of Statistics, Stanford University.

Friedman, J. H. (1987). "Exploratory projection pursuit," *Journal of the American Statistical Association* **82**, 249-266.

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

Friedman, J. H. and Tukey, J. W. (1974). "A projection pursuit algorithm for exploratory data analysis," *IEEE Transactions on Computers* **C-23**, 881-889.

Gill, P., Murray, W. and Wright, M. H. (1981). *Practical Optimization*, Academic Press, London.

Gill, P. E., Murray, W., Saunders, M. A. and Wright, M. A. (1986). "User's guide for NPSOL," Technical Report SOL 86-2, Department of Operations Research, Stanford University.

Good, I. J. (1968). "Corroboration, explanation, evolving probability, simplicity and a sharpened razor," *British Journal for the Philosophy of Science* **19**, 123-143.

Good, I. J. and Gaskins, R. A. (1971). "Nonparametric roughness penalties for probability densities," *Biometrika* **58**, 255-277.

Gorsuch, R. L. (1983). *Factor Analysis*, Lawrence Erlbaum Associates, New Jersey.

Hall, P. (1987). "On polynomial-based projection indices for exploratory projection pursuit," *Annals of Statistics* **17**, 589-605.

Harman, H. H. (1976). *Modern Factor Analysis*, The University of Chicago Press, Chicago.

Hastie, T. and Tibshirani, R. (1984). "Generalized additive models," LSC002, Department of Statistics, Stanford University.

Huber, P. (1985). "Projection pursuit (with discussion)," *Annals of Statistics* **13**, 435-525.

Jones, M. C. (1983). "The projection pursuit algorithm for exploratory data analysis," Ph.D. Dissertation, University of Bath.

Jones, M. C. and Sibson, R. (1987). "What is projection pursuit? (with discussion)," *Journal of the Royal Statistical Society, Series A* **150**, 1-36.

Krzanowski, W. J. (1987). "Selection of variables to preserve multivariate data structure, using principal components," *Applied Statistics* **36**, 22-33.

Lubinsky, D. and Pregibon, D. (1988). "Data analysis as search," *Journal of Econometrics* **38**, 247-268.

Lundy, M. (1985). "Applications of the annealing algorithm to combinatorial problems in statistics," *Biometrika* **72**, 191-198.

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.

Mallows, C. L. (1973). "Some comments on C_p," *Technometrics* **15**, 661-676.

Mallows, C. L. (1983). "Data description," in *Scientific Inference Data Analysis, and Robustness*, eds. G. E. P. Box, T. Leonard and C.-F. Wu, Academic Press, New York, 135-151.

McCabe, G. P. (1984). "Principal variables," *Technometrics* **26**, 137-144.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*, Chapman and Hall, New York.

McDonald, J. A. (1982). "Interactive graphics for data analysis," Ph.D. Dissertation, Department of Statistics, Stanford University.

McDonald, J. A. and Pedersen, J. (1985). "Computing environments for data analysis part I: introduction," *SIAM Journal of Scientific and Statistical Computing* **6**, 1004-1012.

Reinsch, C. H. (1967). "Smoothing by spline functions," *Numerische Mathematik* **10**, 177-183.

Rényi, A. (1961). "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, ed. J. Neyman, 547-561, University of California Press, Berkeley.

Rissanen, J. (1987). "Stochastic complexity (with discussion)" *Journal of the Royal Statistical Society, Series B* **49**, 223-239, 252-265.

Rosenkrantz, R. D. (1977). *Inference, Method and Decision*, Reidel, Boston.

Schwarz, G. (1978). "Estimating the dimension of a model," *Annals of Statistics* **6**, 461-464.

Scott, D. (1979). "On optimal and data-based histograms," *Biometrika* **66**, 605-610.

Silverman, B. W. (1984). "Penalized maximum likelihood estimation," in *Encyclopedia of Statistical Sciences*, eds. S. Kotz and N. L. Johnson, Wiley, New York, 664-667.

Sober, E. (1975). *Simplicity*, Clarendon Press, Oxford.

Stone, C. J. (1981). "Admissible selection of an accurate and parsimonious normal linear regression model," *Annals of Statistics* **9**, 475-485.

Stone, C. J. (1982). "Local asymptotic admissibility of a generalization of Akaike's model selection rule," *Annals of the Institute of Statistical Mathematics* **34**, 123-133.

Stone, M. (1979). "Comments on model selection criteria of Akaike and Schwarz," *Journal of the Royal Statistical Society, Series B* **41**, 276-278.

Sun, J. (1989). "P-values in projection pursuit," Ph.D. Dissertation, Department of Statistics, Stanford University.

Thisted, R. A. (1976). "Ridge regression, minimax estimation and empirical Bayes methods," Ph.D. Dissertation, Department of Statistics, Stanford University.

Thurstone L. L. (1935). *The Vectors of the Mind*, University of Chicago Press, Chicago.

Tukey, J. W. (1961). "Discussion, emphasizing the connection between analysis of variance and spectrum analysis," *Technometrics* **3**, 201-202.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Welsey, Reading, MA.

Tukey, J. W. (1983). "Another look at the future," in *Computer Science and Statistics: Proceedings of the* 14^{th} *Symposium on the Interface*, eds. K. Heiner, R. Sacher and J. Wilkinson, Springer-Verlag, New York, 2-8.

Tukey, P. W. and Tukey, J. W. (1981). "Preparation; prechosen sequences of views," in *Interpreting Multivariate Data*, ed. V. Barnett, Wiley, New York, 189-213.

Watson, G. S. (1983). *Statistics on Spheres*, Wiley, New York.