SLAC REPORT-276
STAN-LCS-11
UC-32
(M)

PRINCIPAL CURVES AND SURFACES*

TREVOR HASTIE

Stanford Linear Accelerator Center
Stanford University
Stanford, California 94305

November 1984

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 A04; Microfiche A01.

*Ph.D Dissertation

Principal Curves and Surfaces

Trevor Hastie

Department of Statistics
Stanford University
and
Computation Group
Stanford Linear Accelerator Center

Abstract

Principal curves are smooth one dimensional curves that pass through the *middle* of a p dimensional data set. They minimize the distance from the points, and provide a non-linear summary of the data. The curves are non-parametric and their shape is suggested by the data. Similarly, principal surfaces are two dimensional surfaces that pass through the middle of the data. The curves and surfaces are found using an iterative procedure which starts with a linear summary such as the usual principal component line or plane. Each successive iteration is a *smooth* or local average of the p dimensional points, where *local* is based on the projections of the points onto the curve or surface of the previous iteration.

A number of linear techniques, such as factor analysis and errors in variables regression, end up using the principal components as their estimates (after a suitable scaling of the co-ordinates). Principal curves and surfaces can be viewed as the estimates of non-linear generalizations of these procedures. We present some real data examples that illustrate these applications.

Principal Curves (or surfaces) have a theoretical definition for distributions: they are the *Self Consistent* curves. A curve is self consistent if each point on the curve is the conditional mean of the points that project there. The main theorem proves that principal curves are critical values of the expected squared distance between the points and the curve. Linear principal components have this property as well; in fact, we prove that if a principal curve is straight, then it is a principal component. These results generalize the usual duality between conditional expectation and distance minimization. We also examine two sources of bias in the procedures, which have the satisfactory property of partially cancelling each other.

We compare the principal curve and surface procedures to other generalizations of principal components in the literature; the usual generalizations transform the space, whereas we transform the model. There are also strong ties with multidimensional scaling.

* Work supported by the Department of Energy under contracts DE-AC03-76SF00515 and DE-AT03-81-ER10843, and by the Office of Naval Research under contract ONR N00014-81-K-0340 and ONR N0014-83-K-0472, and by the U.S. Army Research Office under contract DAAG29-82-K-0056.

Contents

Chapter 1

Introduction

Consider a data set consisting of n observations on two variables, x and y. We can represent the n points in a scatterplot, as in figure 1.1. It is natural to try and summarize the joint behaviour exhibited by the points in the scatterplot. The form of summary we chose depends on the goal of our analysis. A trivial summary is the mean vector which simply locates the center of the cloud but conveys no information about the joint behaviour of the two variables.

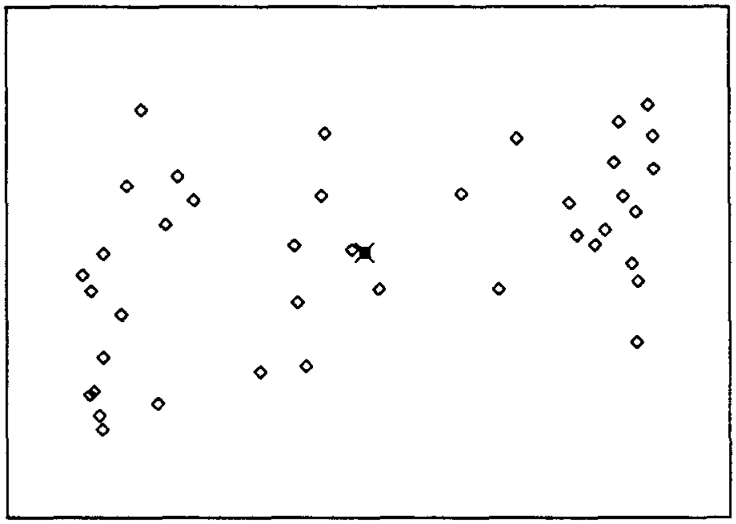

Figure 1.1 A bivariate data set represented by a scatterplot.

It is often sensible to treat one of the variables as a response variable, and the other as an explanatory variable. The aim of the analysis is then to seek a rule for predicting the response (or average response) using the value of the explanatory variable. Standard linear regression produces a linear prediction rule. The expectation of y is modeled as a linear function of x and is estimated by least squares. This procedure is equivalent to finding the line that minimizes the sum of vertical squared errors, as depicted in figure 1.2a.

When looking at such a regression line, it is natural to think of it as a summary of the data. However, in constructing this summary we concerned ourselves only with errors in the response variable. In many situations we don't have a preferred variable that we wish to label response, but would still like to summarize the joint behaviour of x and y. The dashed line in figure 1.2a shows what happens if we used x as the response. So simply assigning the role of response to one of the variables could lead to a poor summary. An obvious alternative is to summarize the data by a straight line that treats the two variables symmetrically. The first principal component line in figure 1.2b does just this — it is found by minimizing the orthogonal errors.

Linear regression has been generalized to include nonlinear functions of x. This has been achieved using predefined parametric functions, and more recently non-parametric scatterplot smoothers such as kernel smoothers, (Gasser and Muller 1979), nearest neighbor smoothers, (Cleveland 1979, Friedman and Stuetzle 1981), and spline smoothers (Reinsch 1967). In general scatterplot smoothers produce a smooth curve that attempts to minimize the vertical errors as depicted in figure 1.2c. The non-parametric versions listed above allow the data to dictate the form of the non-linear dependency.

In this dissertation we consider similar generalizations for the symmetric situation. Instead of summarizing the data with a straight line, we use a smooth curve; in finding the curve we treat the two variables symmetrically. Such curves will pass through the *middle* of the data in a smooth way, without restricting *smooth* to mean linear, or for that matter without implying that the *middle* of the data is a straight line. This situation is depicted in figure 1.2d. The figure suggests that such curves minimize the orthogonal distances to the points. It turns out that for a suitable definition of *middle* this is indeed the case. We name them *Principal Curves*. If, however, the data cloud is ellipsoidal in shape then one could well imagine that a straight line passes through the middle of the cloud. In this case we expect our principal curve to be straight as well.

The principal component line plays roles other than that of a data summary:

- In *errors in variables* regression the explanatory variables are observed with error (as well as the response). This can occur in practice when both variables are measurements of some underlying variables, and there is error in the measurements. It also occurs in observational studies where neither variable is fixed by design. If the aim of the analysis is prediction of y or regression and if the x variable is never observed *without* error, then the best we can do is condition on the observed x's and perform the standard regression analysis (Madansky 1959, Kendall and Stuart 1961, Lindley 1947). If, however, we do expect to observe x without error then we can model the expectation of y as a linear function of the systematic component of x. After suitably scaling the variables, this model is estimated by the principal component line.

- Often we want to replace a number of highly correlated variables by a single variable, such as a normalized linear combination of the original set. The first principal component is the normalized linear combination with the largest variance.

- In factor analysis we model the *systematic* component of the data as linear combinations of a small subset of new unobservable variables called factors. In many cases the models are estimated using the linear principal components summary. Variations of this model have appeared in many different forms in the literature. These include linear functional and structural models, errors in variables and total least squares. (Anderson 1982, Golub and van Loan 1979).

In the same spirit we propose using principal curves as the estimates of the systematic components in non-linear versions of the models mentioned above. This broadens the scope and use of such curves considerably. This dissertation deals with the definition, description and estimation of such principal curves, which are more generally one dimensional curves in p-space. When we have three or more variables we can carry the generalizations further. We can think of modeling the data with a 2 or more dimensional surface in p space. Let us first consider only three variables and a 2-surface, and deal with each of the four situations in figure 1.2in turn.

- If one of the variables is a response variable, then the usual linear regression model estimates the conditional expectation of y given $x = (x_1, x_2)$ by the least squares plane. This is a planar response surface which is once again obtained by minimizing the squared errors in y. These errors are the vertical distances between y and the point on the plane vertically above or below y.

- Often a linear response surface does not adequately model the conditional expectation. We then turn to nonlinear two dimensional response surfaces which are smooth surfaces that minimize the vertical errors. They are estimated by surface smoothers that are direct extensions of the scatterplot smoothers for curve estimation.

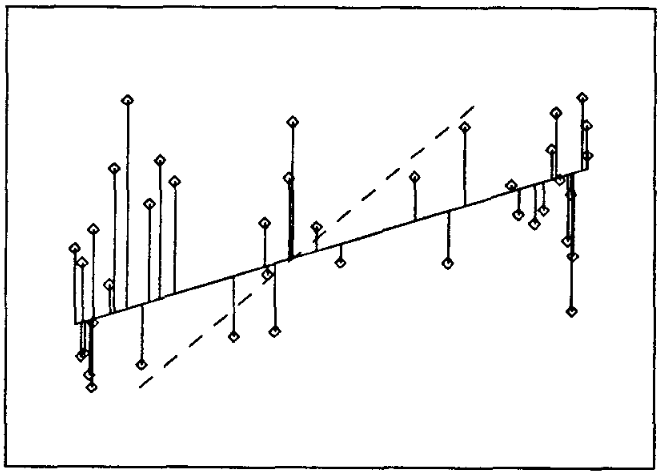

Figure 1.2a The linear regression line minimizes the sum of squared errors in the response variable.

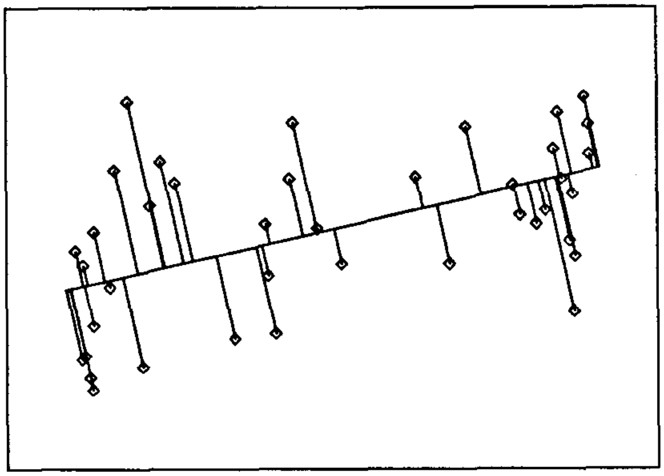

Figure 1.2b The principal component line minimizes the sum of squared errors in all the variables.

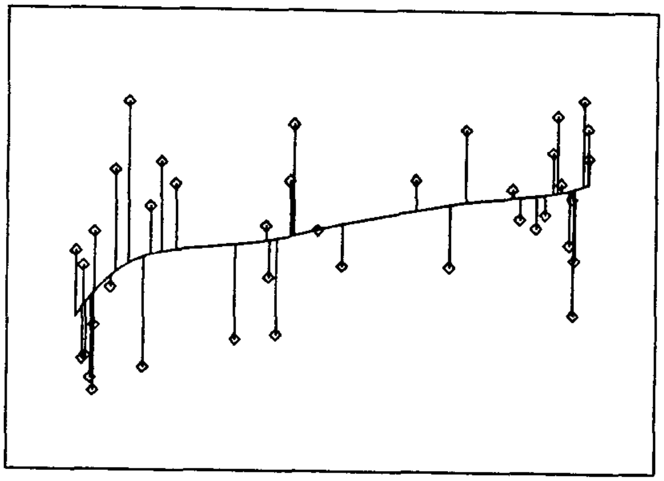

Figure 1.2c The smooth regression curve minimizes the sum of squared errors in the response variable, subject to smoothness con straints.

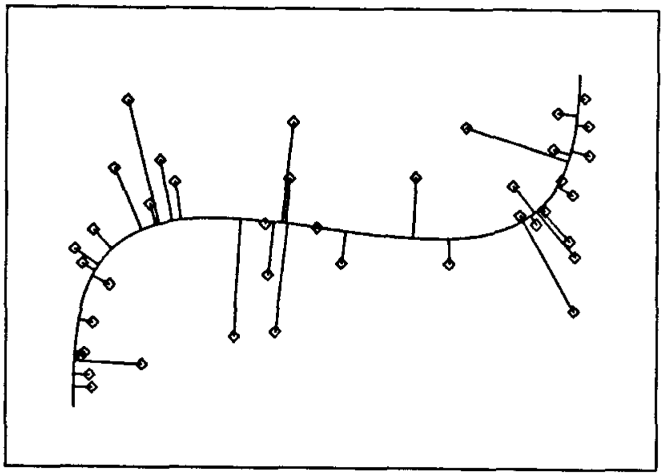

Figure 1.2d The principal curve minimizes the sum of squared errors in all the variables, subject to smoothness constraints.

- If all the variables are to be treated symmetrically the principal component plane passes through the data in such a way that the sum of squared distances from the points to the plane is minimized. This in turn is an estimate for the systematic component in a 2-dimensional linear model for the mean of the three variables.

- Finally, in this symmetric situation, it is often unnatural to assume that the best two dimensional summary is a plane. Principal surfaces are smooth surfaces that pass through the middle of the data cloud; they minimize the sum of squared distances between the points and the surface. They can also be thought of as a an estimate for the two dimensional systematic component for the means of the three variables.

These surfaces are easily generalized to 2-dimensional surfaces in p space, although they are hard to visualize for $p > 3$.

The dissertation is organized as follows:

- In chapter 2 we discuss in more detail the linear principal components model, as well as the linear relationship model hinted at above. They are identical in many cases, and we attempt to tie them together in the situations where this is possible. We then propose the non-linear generalizations.

- In Chapter 3 we define principal curves and surfaces in detail. We motivate an algorithm for estimating such models, and demonstrate the algorithm using simulated data with very definite and *difficult* structure.

- Chapter 4 is theoretical in nature, and proves some of the claims in the previous chapters. The main result in this chapter is a theorem which shows that curves that pass through the *middle* of the data are in fact critical points of a distance function. The principal curve and surface procedures are inherently biased. This chapter concludes with a discussion of the various forms and severity of this bias.

- Chapter 5 deals with the algorithms in detail. There is a brief discussion of scatterplot smoothers, and we show how to deal with the problem of finding the closest point on the curve. The algorithm is explained by means of simple examples, and a method for span selection is given.

- Chapter 6 contains six examples of the use and abilities of the procedures using real and simulated data. Some of the examples introduce special features of the procedures such as inference using the bootstrap, robust options and outlier detection.

- Chapter 7 provides a discussion of related work in the literature, and gives details of some of the more recent ideas. This is followed by some concluding remarks on the work covered in this dissertation.