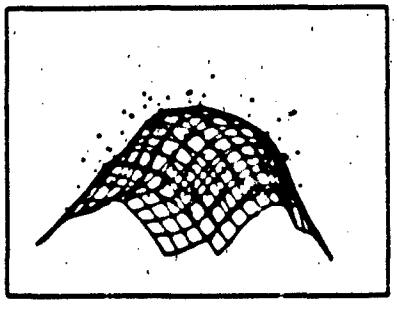
PRINCIPAL CURVES AND SURFACES

Trevor Hastie

Technical Report No. 11
November 1984

Laboratory for Computational Statistics



Department of Statistics
Stanford University

LCS 11 AD-A148837	
4. TITLE (and Subtitle) PRINCIPAL CURVES AND SURFACES	5. TYPE OF REPORT & PERIOD COVERED TECHNICAL 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Trevor Hastie	8. CONTRACT OR GRANT NUMBER(s) N00014-83-K-0472
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics and Computational Group Stanford Linear Accelerator Center Stanford University, Stanford, CA 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Office of Naval Research Department of the Navy Arlington, VA 22217	12. REPORT DATE November 1984 13. NUMBER OF PAGES 103
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Navy position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Principal components, non-linear, smooth, errors in variables, orthogonal regression

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Principal curves are smooth one dimensional curves that pass through the *middle* of a p dimensional data set. They minimise 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

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

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 generalisations of these procedures. We present some real data examples that illustrate these applications.

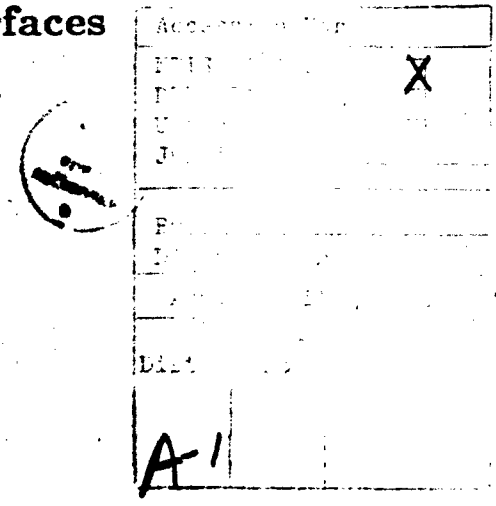
Principal Curves (or surfaces) have a theoretical definition for distributions: they are the *Self Consistent* curves. A curve is self consistent if each point on the curve is the conditional mean of the points that project there. The main theorem proves that principal curves are critical values of the 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 generalise the usual duality between conditional expectation and distance minimisation. 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 generalisations of principal components in the literature; the usual generalisations transform the space, whereas we transform the model. There are also strong ties with multidimensional scaling.

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 minimise 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 generalise the usual duality between conditional expectation and distance minimisation. 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 generalisations of principal components in the literature; the usual generalisations 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.

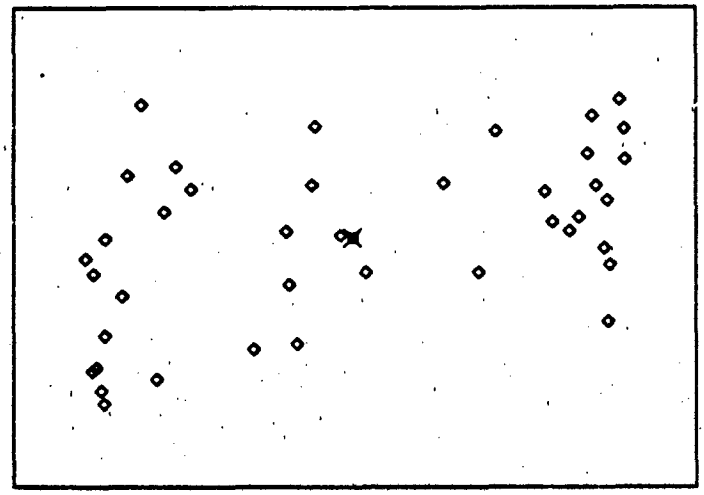


Figure 1.1 A bivariate data set represented by a scatterplot.

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.

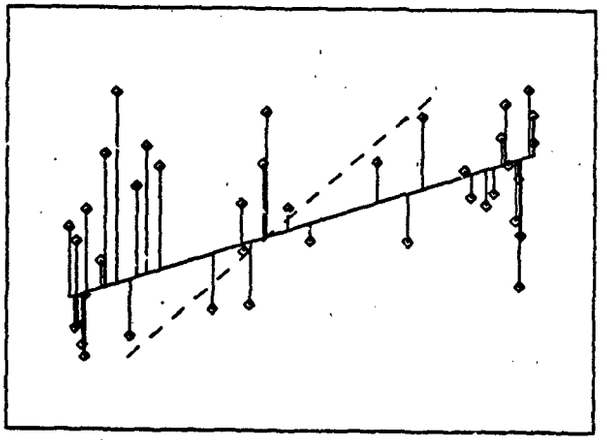


Figure 1.2a The linear regression line minimises the sum of squared errors in the response variable.

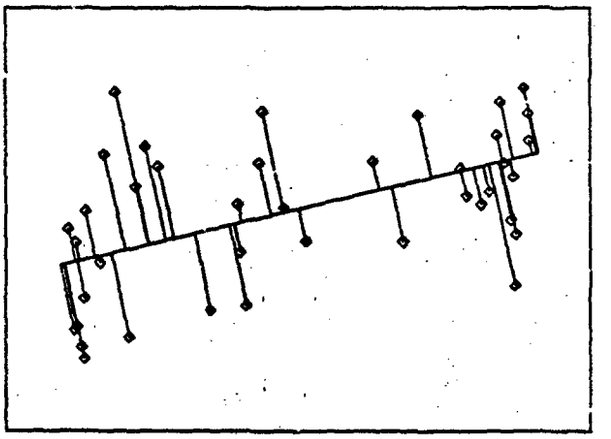


Figure 1.2b The principal component line minimises the sum of squared errors in all the variables.

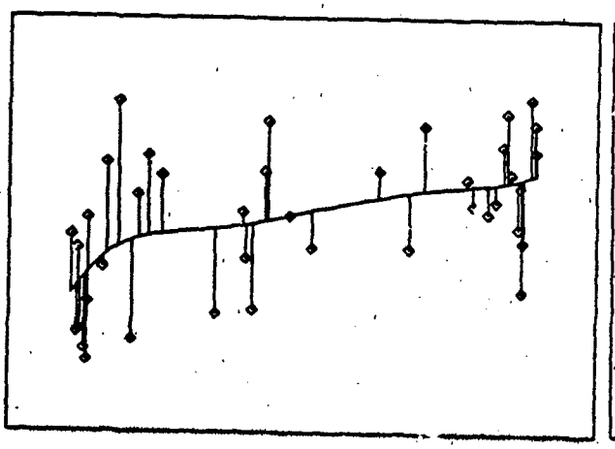


Figure 1.2c The smooth regression curve minimises the sum of squared errors in the response variable, subject to smoothness constraints.

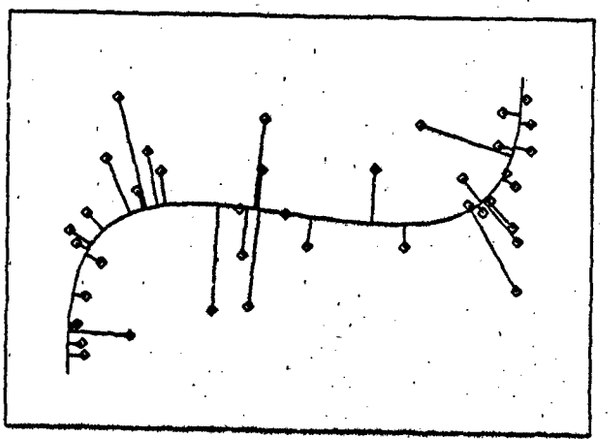


Figure 1.2d The principal curve minimizes the sum of squared errors in all the variables, subject to smoothness constraints.

- 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.

Chapter 2

Background and Motivation

Consider a data matrix X with n rows and p columns. The matrix consists of n points or vectors with p coordinates. In many situations the matrix will have arisen as n observations of a vector random variable.

2.1. Linear Principal Components.

The first (linear) principal component is the normalized linear combination of the p variables with the largest sample variance. It is convenient to think of X as a cloud of n points in p-space. The principal component is then the length of the projection of the n points onto a direction vector. The vector is chosen so that the variance of the projected points along it is largest. Any line parallel to this vector will have the same property. To tie it down we insist that it passes through the mean vector. This line then has the appealing property of being the line in p-space that is closest to the data. Closest is in terms of average squared euclidian distance. We think of the projection as being the best linear one dimensional summary of the data X. Of course this linear summary might be totally inadequate locally but it attempts to provide a reasonable global summary.

The theory and practical issues involved in linear principal components analysis are well known (Barnett 1981, Gnanadesikan 1977), and the technique is originally due to Spearman (1904), and then later developed by Hotelling (1933). We can find the the second component, orthogonal to the first, that has the next highest variance. The plane spanned by the two vectors and including the mean vector is the plane closest to the data. In general we can find the $m < p$ dimensional hyperplane that contains the most variance, and is closest to the data.

The solution to the problem is obtained by computing the singular value decomposition or basic structure of X, (centered with respect to the sample mean vector), or equivalently the eigen decomposition of the sample covariance matrix (Golub and Reinsch 1970, Greenacre 1984). Without any loss in generality we assume from now on that X is centered. If this is not the case, we can center X, perform the analysis, and uncenter the results by

adding back the mean vector.

In particular, the first principal component direction vector a is the largest normalized eigenvector of S, the sample covariance matrix. The principal component itself is Xa, an n vector with elements $\lambda_i = x_i'a$ where x_i' is the ith row of X and λ_i is the one dimensional summary variable for the ith obeservation. The coordinates in p-space of the projection of the ith obeservation on a are given by *

$$a\lambda_i = aa'x_i \tag{2.1}$$

There is no underlying model in the above. We merely regard the first component as a good summary of the original variables if it accounts for a large fraction of the total variance.

2.2. A linear model formulation.

In this section we describe a linear model formulation for the p variables. This formulation includes many familiar models such as linear regression and factor analysis. We end up showing in 2.2.2 that the estimation of the systematic component of some of these models is once again the principal component procedure.

2.2.1. Outline of the linear model.

Consider a model for the obeserved data

$$x_i = u_i + e_i \tag{2.2}$$

where u_i is an unobeservable systematic component and e_i an unobservable random component (We only get to see their sum). We usually impose some linear structure on u_i, namely

$$u_i = u_0 + A\lambda_i \tag{2.3}$$

where u_0 is constant location vector, A is a $p \times m$ matrix and λ_i is an m-vector. For the procedures considered u_0 is always estimated by the sample mean vector $\bar{x}$; without loss of generality we will simply assume that X has been centered and ignore the term u_0. We also

* If X is not centered we center it by forming $\tilde{X} = X - 1\bar{x}$. Then the principal component is $\lambda = \tilde{X}a$ and the estimate in p space for the projection of the ith obeservation onto the principal component line $\bar{x} + a\gamma$ is $\bar{x} + a\lambda_i = \bar{x} + aa'(x_i - \bar{x})$

assume that e_i are mutually independent and identically distributed random vectors with mean $\mathbf{0}$ and covariance matrix Ψ and are independent of the λ_i.

If the λ_i are considered to be random as well, the model is referred to as the linear structural model, or more commonly as the factor analysis model. If the λ_i are fixed it is referred as the linear functional model. The model (2.3) includes some familiar models as special cases:

- Let A be $p \times (p-1)$ with rank $(p-1)$. We can write A as

$$\begin{pmatrix} a' \\ I \end{pmatrix}$$

where a is a $(p-1)$ vector and I is $(p-1) \times (p-1)$ since we can post-multiply A by an arbitrary non-singular $(p-1) \times (p-1)$ matrix and pre-multiply λ_i by its inverse. Thus we can write the model (2.3) as

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} a' \\ I \end{pmatrix} \lambda_i + \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \tag{2.4}$$

where $E(e_i) = \mathbf{0}$ and assume $Cov(e_i) = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$. If $\sigma_2^2 = \sigma_3^2 = \ldots = \sigma_p^2 = 0$ then we have the usual linear regression model with response x_{1i} and regressor variables x_{2i}.

- If the variances are not zero we have the errors in variables regression model. The idea is to find a $(p-1)$ dimensional hyperplane in p-space that approximates the data well. The model takes care of errors in all the variables, whereas the usual linear regression model considers errors only in the response variable. This is a form of linear functional analysis.

- When the λ_i are random we have the usual factor analysis model, which includes the random effects Anova. This is also referred to as the linear structural model.

- If all the variances are zero and the λ_i are random and A is $p \times p$ the model represents the principal component change of basis. In this situation it is clear that the λ_{ji} are each functions of the x_i.

For a full treatment of the above models see Anderson (1982).

2.2.2. Estimation

We return for simplicity to the case where $m = 1$. Thus

$$x_i = a\lambda_i + e_i \tag{2.5}$$

The systematic components $a\lambda_i$ are points in p-space confined to the line defined by a multiple λ_i of the vector a. We need to estimate λ_i for each observation, and the direction vector.

We now state some results which can be found in Anderson (1982).

If either

- the e_i are jointly Normal with a scalar covariance cI, where c is possibly unknown, and if λ_i are random or fixed, and we estimate by maximum likelihood

or

- as above but we drop the Normal assumption and estimate by least squares,

then the estimate of λ_i is once again the first principal component and that of a the principal component direction vector. In both cases the quantity we wish to minimize is

$$RSS(\lambda, a) = \sum_{i=1}^{n} \|x_i - a\lambda_i\|^2. \tag{2.6}$$

It is easy to see that for any a the appropriate value for λ_i is obtained by projecting the point x_i onto a. Thus equation (2.6) reduces to

$$\begin{aligned} RSS(a) &= \sum_{i=1}^{n} \|x_i - aa'x_i\|^2 \\ &= \mathrm{tr}\, XX' - a'X'Xa \end{aligned} \tag{2.7}$$

The normalized solution to (2.7) is the largest eigenvector of $X'X$.

If the error covariance Ψ is general but known, we can transform the problem to the previous case. This is the same as using the Mahalanobis distance defined in terms of Ψ. In particular when Ψ is diagonal the procedure amounts to finding the line that minimizes the weighted distance to the points and is depicted in figure (2.1) below.

If the error covariance is unknown and not scalar then we require replicate observations in order to estimate it.

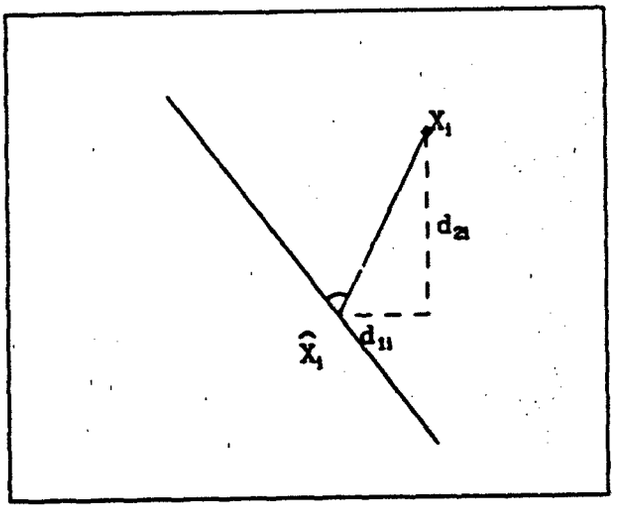


Figure 2.1 If $\Psi = \text{diag}(\sigma_1^2, \sigma_2^2)$ then we minimise the weighted distance $\sum_i (d_{1i}^2/\sigma_1^2 + d_{2i}^2/\sigma_2^2)$ from the points to the line.

2.2.3. Units of measurement.

It is often a problem in multivariate data analysis that variables have different error variances, even though they are measured in the same units. A worse situation is that often the variables are measured in completely different and incommensurable units. When we use least squares to estimate a lower dimensional summary, we explicitly combine the errors on each variable using the usual sum of components loss function, as in (2.6) . This then gives equal weight to each of the components. The solution is thus not invariant to changes in the scale of any of the variables. This is easily demonstrated by considering a spherical point cloud. If we scale up one of the co-ordinates an arbitrary amount, we can create as much linear structure as we like. In this situation we would really like to weigh the errors in the estimation of our model according to the variance of the measurement errors, which is seldom known. The safest procedure in this situation is to standardize each of the coordinates to have unit variance. This could destroy some of the structure that exists but without further knowledge about the scale of the components this yields a procedure that is invariant to coordinate scale transformations.

If, on the other hand, it is known that the variables are measured in the same units, we should not do any scaling at all. An apparent counter-example occurs if we make

measurements of the same quantities in different situations, with different measurement devices. An example might be taking seismic readings at different sights at the same instances with different recording devices. If the error variances of the two devices are different, we would want to scale the components differently.

To sum up so far, the principal component summary, besides being a convenient data reduction technique, provides us with the estimate of a formal parametric linear model which covers a wide variety of situations. An original example of the one factor model given here is that of Spearman (1904). The x_i are scores on psychological tests and the λ_i is some underlying unobservable general intelligence factor.

The estimation in all the cases amounts to finding a m-dimensional hyperplane in p-space that is closest to the points in some metric.

2.3. A non-linear generalization of the linear model.

The above formulation is often very restrictive in that it assumes that the systematic component in (2.2) is linear, as in (2.3). It is true in some cases that we can approximate a nonlinear surface by its first order linear component. In other cases we do not have sufficient data to estimate any more than a linear component. Apart from these cases, it is more reasonable to assume a model of the form

$$x_i = f(\lambda_i) + e_i \tag{2.8}$$

where λ_i is a m-vector as before and f is a p-vector of functions, each with m arguments. The functions are required to be smooth relative to the errors. This is a natural generalization of the linear model.

This dissertation deals with a generalization of the linear principal components. Instead of finding lines and planes that come close to the data, we find curves and surfaces. Just as the linear principal components are estimates for the variety of linear models listed above, so will our non- linear versions be estimates for models of the form (2.8) . So in addition to having a more general summary of multidimensional data, we provide a means of estimating the systematic component in a large class of models suitably generalized to include non-linearities. We refer to these summaries as principal curves and surfaces.

So far the discussion has concentrated on data sets. We can just as well formulate the above models for p dimensional probability distributions. We would then regard the data set

as a sample from this distribution and the functions derived for the data set will be regarded as estimates of the corresponding functions defined for the distribution. These models then define one and two dimensional surfaces that summarize the p dimensional distribution. The point $f(\lambda)$ on the surface that corresponds to a general point x from the distribution is a p dimensional random variable that can be summarized by a two dimensional random variable λ.

2.4. Other generalizations.

There have been a number of generalizations of the principal component model suggested in the literature.

- "Generalized principal components" usually refers to the adaptation of the linear model in which the coordinates are first transformed, and then the standard principal component analysis is carried out on the transformed coordinates.
- Multidimensional scaling (MDS) finds a low dimensional representation for the high dimensional point cloud, such that the sum of squared interpoint distances are preserved. This constraint has been modified in certain cases to cater only for points that are *close* in the original space.
- Proximity analysis provides parametric representations for data without noise.
- Non-linear factor analysis is a generalization similar to ours, except parametric coordinate functions are used.

We have been deliberately brief in listing these alternatives. Chapter 7 contains a detailed discussion and comparison of each of the above with the principal curve and surface models.

Chapter 3

The Principal Curve and Surface models

In this chapter we define the principal curve and surface models, first for a p dimensional probability distribution, and then for a p dimensional finite data set. In order to achieve some continuity in the presentation, we motivate and then simply state results and theorems in this chapter, and prove them in chapter 4.

3.1. The principal curves of a probability distribution.

We first give a brief introduction to one dimensional surfaces or curves, and then define the *principal curves* of smooth probability distributions in p space.

3.1.1. One dimensional curves.

A one dimensional curve f is a vector of functions of a single variable, which we denote by λ. These functions are called the coordinate functions, and λ provides an ordering along the curve. If the coordinate functions are smooth, then f will be a smooth curve. We can clearly make any monotone transformation to λ, say $m(\lambda)$, and by modifying the coordinate functions appropriately the curve remains unchanged. The parametrization, however, is different. There is a natural parametrization for curves in terms of the arc-length. The arc-length of a curve f from λ_0 to λ_1 is given by

$$l = \int_{\lambda_0}^{\lambda_1} \|f'(z)\|\, dz.$$

If $\|f'(z)\| \equiv 1$ then $l = \lambda_1 - \lambda_0$. This is a rather desirable situation, since if all the coordinate variables are in the same units of measurement, then λ is also in those units. The vector $f'(\lambda)$ is tangent to the curve at λ and is sometimes called the *velocity vector* at λ. A curve with $\|f'\| \equiv 1$ is called a unit speed parametrized curve. We can always reparametrize any smooth curve to make it unit speed. If v is a unit vector, then $f(\lambda) = v_0 + \lambda v$ is a unit speed *straight* curve.

The vector $f''(\lambda)$ is called the acceleration of the curve at λ, and for a unit speed curve, it is easy to check that it is orthogonal to the tangent vector. In this case $f''/\|f''\|$

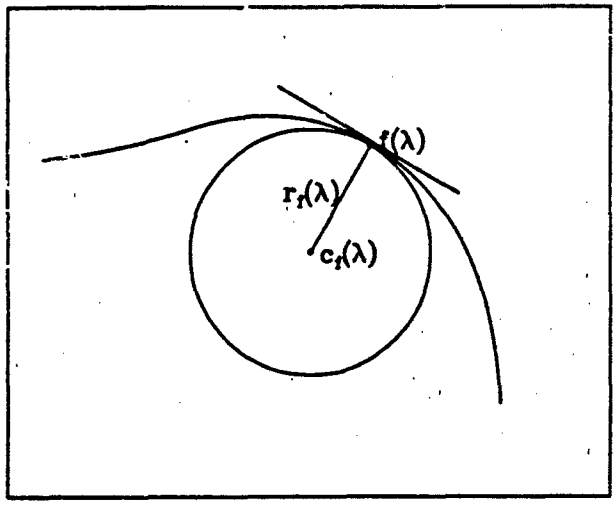


Figure (3.1) The radius of curvature is the radius of the circle tangent to the curve with the same acceleration as the curve.

is called the *principal normal* of the curve at λ. Since the acceleration measures the rate and direction in which the tangent vector turns, it is not surprising that the curvature of a parametrized curve is defined in terms of it. The easiest way to think of curvature is in terms of a circle. We fit a circle tangent to the curve at a particular point and lying in the plane spanned by the velocity vector and the principal normal. The circle is constructed to have the same acceleration as the curve, and the radius of curvature of the curve at that point is defined as the radius of the circle. It is easy to check that for a unit speed curve we get

$$r_f(\lambda) \stackrel{\text{def.}}{=} \text{radius of curvature of } f \text{ at } \lambda$$
$$= 1/\|f''(\lambda)\|$$

The *center of curvature* of the curve at λ is denoted by $c_f(\lambda)$ and is the center of this circle.

3.1.2. Definition of principal curves.

We now define what we mean by a curve that passes through the *middle* of the data — what we call a *principal curve*. Figure 3.2 represents such a curve. At any particular location on the curve, we collect all the points in p space that have that location as their closest point on the curve. Loosely speaking, we collect all the points that *project* there. Then the location on the curve is the average of these points. Any curve that has this property

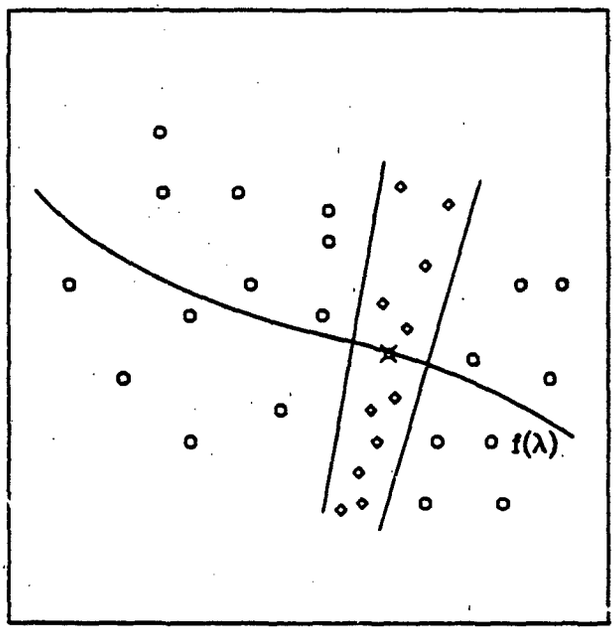


Figure (3.2) Each point on a principal curve is the average of the points that project there.

is called a principal curve. One might say that principal curves are their own conditional expectation. We will prove later these curves are critical points of a distance function, as are the principal components.

In the figure we have actually shown the points that project into a neighborhood on the curve. We do this because usually for finite data sets at most one data point projects at any particular spot on the curve. Notice that the points lie in a segment with center at the center of curvature of the arc in question. We will discuss this phenomenon in more detail in the section on bias in chapter 4.

We can formalize the above definition. Suppose X is a random vector in p-space, with continuous probability density $h(x)$. Let $\mathcal{G}$ be the class of differentiable 1-dimensional curves in $\mathbb{R}^p$, parametrized by λ. In addition we do not allow curves that form closed loops, so they may not intersect themselves or be tangent to themselves. Suppose $\lambda \in \Lambda_f$ for each f in $\mathcal{G}$. For $f \in \mathcal{G}$ and $x \in \mathbb{R}^p$, we define the projection index $\lambda_f : \mathbb{R}^p \mapsto \Lambda_f$ by

$$\lambda_f(x) = \max_{\lambda}\{\lambda : \|x - f(\lambda)\| = \inf_{\mu} \|x - f(\mu)\|\}. \tag{3.1}$$

The projection index $\lambda_f(x)$ of x is the value of λ for which $f(\lambda)$ is closest to x. There might be a number of such points (suppose f is a circle and x is at the center), so we pick the largest such value of λ. We will show in chapter 4 that $\lambda_f(x)$ is a measureable mapping from $\mathbf{R}^p$ to $\mathbf{R}^1$, and thus $\lambda_f(X)$ is a random variable.

Definition

The *Principal Curves* of h are those members of $\mathcal{G}$ which are *self consistent*. A curve $f \in \mathcal{G}$ is self consistent if

$$\mathbf{E}(X \mid \lambda_f(X) = \lambda) = f(\lambda) \;\; \forall \lambda \in \Lambda_f$$

We call the class of principal curves $\mathcal{F}(h)$.

3.1.3. Existence of principal curves.

An immediate question might be whether such curves exist or not, and for what kinds of distributions. It is easy to check that for ellipsoidal distributions, the principal components are in fact principal curves. For a spherically symmetric distribution, any line through the mean vector is a principal curve.

What about data generated from a model as in equation 2.8, where λ_i is 1 dimensional? Is f a principal curve for this distribution? The answer in general is no. Before we even try to answer it, we have to enquire about the distribution of λ_i and e_i. Suppose that the data is well behaved in that the distribution of e_i has tight enough support, so that no points can fall beyond the centers of curvature of f. This guarantees that each point has a unique closest point to the curve. We show in the next chapter that even under these ideal conditions (spherically symmetric errors, slowly changing curvature) the average of points that project at a particular point on the curve from which they are generated lies *outside* the circle of curvature at that point on the curve. This means that the principal curve will be different from the generating curve. So in this situation an unbiased estimate of the principal curve will be a biased estimate of the functional model. This bias, however, is small and decreases to zero as the variance of the errors gets small relative to the radius of curvature.

3.1.4. The distance property of principal curves.

The principal components are critical points of the squared distance from the points to their projections on straight curves (lines). Is there any analogous property for principal curves?

It turns out that there is. Let $d(x, f)$ denote the usual euclidian distance from a point x to its projection on the curve f:

$$d(x, f) \stackrel{\text{def}}{=} \left\| x - f(\lambda_f(x)) \right\| \tag{3.2}$$

and define the function $D^2 : \mathcal{G} \rightarrow \mathbb{R}^1$ by

$$D^2(f) \stackrel{\text{def}}{=} \mathbf{E} d^2(X, f).$$

We show that if we restrict the curves to be straight lines, then the principal components are the only critical values of $D^2(f)$. Critical value here is in the variational sense: if f and g are straight lines and we form $f_\epsilon = f + \epsilon g$, then we define f to be a critical value of D^2 iff

$$dD^2(f_\epsilon)/d\epsilon|_{\epsilon=0} = 0.$$

This means that they are minima, maxima or saddle points of this distance function. If we restrict f and g to be members of the subset of $\mathcal{G}$ of curves defined on a compact A, then principal curves have this property as well. In this case f_ϵ describes a class of curves about f that shrink in as ϵ gets small. The corresponding result is: $dD^2(f_\epsilon)/d\epsilon|_{\epsilon=0} = 0$ iff f is a principal curve of h. This is a key property and is an essential link to all the previous models and motivation in chapter 2. This property is similar to that enjoyed by conditional expectations or projections; the residual distance is minimized. Figure (3.3) illustrates the idea, and in fact is almost a proof in one direction.

Suppose k is not a principal curve. Then the curve defined by $f(\lambda) = \mathbf{E}(X \mid \lambda_k(X) = \lambda)$ certainly gets closer to the points in any of the neighborhoods than the original curve. This is the property of conditional expectation. Now the points in any neighborhood defined by λ_k might end up in different neighborhoods when projected onto f, but this reduces the distances even further. This shows that k cannot be a critical value of the distance function.

An immediate consequence of these two results is that if a principal curve is a straight line, then it is a principal component. Another result is that principal components are self consistent if we replace conditional expectations by linear projections.

3.1.4.1 A smooth subset of principal curves.

We have defined principal curves in a rather general fashion without any smoothness restrictions. The distance theorem tells us that if we have a principal curve, we will not find any curves nearby with the same expected distance. We have a mental image of what we

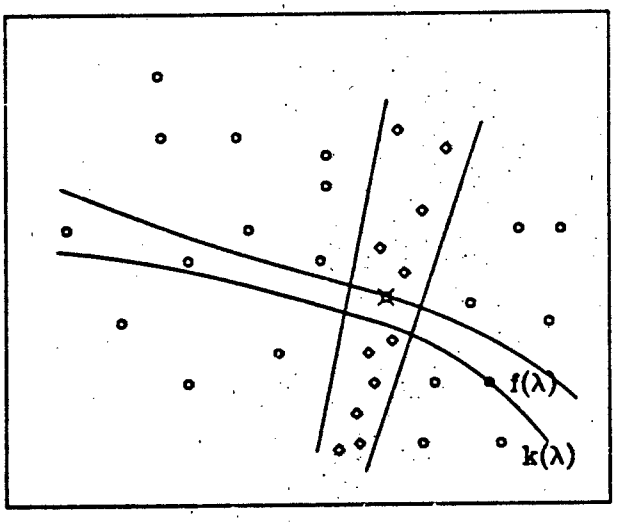


Figure 3.3 The conditional expectation curve gets at least as close to the points as the original curve.

would like the curves to look like. They should pass through the data smoothly enough so that each data point has an unambiguous closest point on the curve. This smoothness will be dictated by the density h. It turns out that we can neatly summarize this requirement. Consider the subset $\mathcal{F}_c(h) \subset \mathcal{F}(h)$ of principal curves of h, where $f \in \mathcal{F}_c(h)$ iff $f \in \mathcal{F}(h)$ and $\lambda_f(x)$ is continuous in x for all points x in the support of h. In words this says that if two points x and y are close together, then their points of projection on the curve are close together. This has a number of implications, some of which are obvious, which we will list now and prove later.

- There is only one closest point on the principal curve for each x in the support of h.
- The curve is globally well behaved. This means that the curve cannot bend back and come too close to itself since that will lead to ambiguities in projection. (If we want to deal with closed curves, such as a circle, a technical modification in the definition of λ is required).
- There are no points at or beyond the centers of curvature of the curve. This says that the curve is smooth relative to the variance of the data about the curve. This has intuitive appeal. If the data is very noisy, we cannot hope to recover more than a very smooth curve (nearly a straight line) from it.

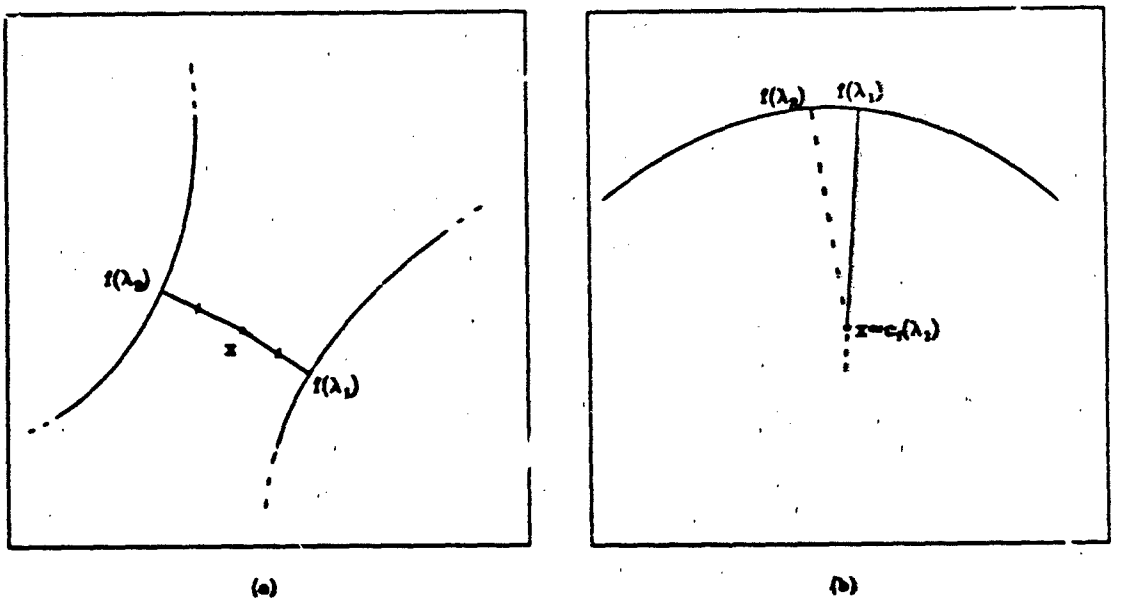


Figure 3.4 The continuity constraint avoids global ambiguities (a) and local ambiguities (b) in projection.

Figure 3.4 illustrates the way in which the continuity constraint avoids global and local ambiguities. Notice that $\mathcal{F}_c(h)$ depends on the density h of X. We say in the support of h, but if the errors have an infinite range, this definition would only allow straight lines. We can make some technical modifications to overcome this hurdle, such as insisting that h has compact support. This rules out any theoretical consideration of curves with gaussian errors, although in practice we always have compact support. Nevertheless, the class $\mathcal{F}_c(h)$ will prove to be useful in understanding some of the properties of principal curves.

3.2. The principal surfaces of a probability distribution.

3.2.1. Two dimensional surfaces.

The level of difficulty increases dramatically as we move from one dimensional surfaces or curves to higher dimensional surfaces. In this work we will only deal with 2-dimensional surfaces in p space. In fact we shall deal only with 2-surfaces that admit a global parametrization. This allows us to define f to be a smooth 2-dimensional globally parametrized surface

if $f : \Lambda \mapsto \mathbf{R}^p$ for $\Lambda \subseteq \mathbf{R}^2$ is a vector of smooth functions:

$$f(\lambda) = \begin{pmatrix} f_1(\lambda) \\ f_2(\lambda) \\ \vdots \\ f_p(\lambda) \end{pmatrix} = \begin{pmatrix} f_1(\lambda_1, \lambda_2) \\ f_2(\lambda_1, \lambda_2) \\ \vdots \\ f_p(\lambda_1, \lambda_2) \end{pmatrix} \tag{3.3}$$

Another way of defining a 2-surface in p space is to have $p-2$ constraints on the p coordinates. An example is the unit sphere in $\mathbf{R}^3$. It can be defined as $\{x : x \in \mathbf{R}^3, \|x\| = 1\}$. There is one constraint. We will call this the *implicit* definition.

Not all 2-surfaces have implicit definitions (möbius band), and similarly not all surfaces have global parametrizations. However, locally an equivalence can be established (Thorpe 1978).

The concept of arc-length generalizes to surface area. However, we cannot always reparametrize the surface so that units of area in the parameter space correspond to units of area in the surface. Once again, local parametrizations do permit this change of units.

Curvature also takes on another dimension. The curvature of a surface at any point might be different depending on which direction we look from. The way this is resolved is to look from all possible directions, and the *first principal curvature* is the curvature corresponding to the direction in which the curvature is greatest. The *second principal curvature* corresponds to the largest curvature in a direction orthogonal to the first. For 2-surfaces there are only two orthogonal directions, so we are done.

3.2.2. Definition of principal surfaces.

Once again let X be a random vector in p-space, with continuous probability density $h(x)$. Let $\mathcal{G}^2$ be the class of differentiable 2-dimensional surfaces in $\mathbf{R}^p$, parametrized by $\lambda \in \Lambda_f$, a 2-dimensional parameter vector.

For $f \in \mathcal{G}^2$ and $x \in \mathbf{R}^p$, we define the projection index $\lambda_f(x)$ by

$$\lambda_f(x) = \max_{\lambda_2} \max_{\lambda_1} \{\lambda : \|x - f(\lambda)\| = \inf_{\mu} \|x - f(\mu)\|\}. \tag{3.4}$$

The projection index defines the closest point on the surface; if there is more than one, it picks the one with the largest first component. If this is still not unique, it then maximizes over the second component. Once again $\lambda_f(x)$ is a measureable mapping from $\mathbb{R}^p$ into $\mathbb{R}^2$, and $\lambda_f(X)$ is a random vector.

Definition

The *Principal Surfaces* of h are those members of $\mathcal{G}^2$ which are self consistent:

$$\mathbf{E}(X \mid \lambda_f(X) = \lambda) = f(\lambda)$$

Figure (5.5) demonstrates the situation.

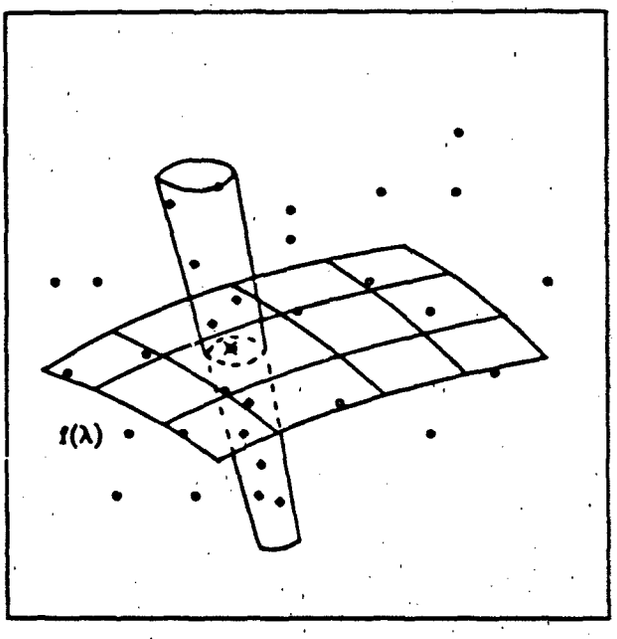


Figure 5.5 Each point on a principal surface is the average of the points that project there.

The plane spanned by the first and second principal components minimizes the distance from the points to their projections onto any plane. Once again let $d(x, f)$ denote the usual euclidian distance from a point x to its projection on the surface f, and $D^2(f) = \mathbf{E}d^2(X, f)$. If the surfaces are restricted to be planes, then the planes spanned by any pair of principal

components are the only critical values of $D^2(f)$. There is a result analogous to the one to be proven for principal curves. If we restrict f to be the members of $\mathcal{G}^2$ defined on connected compact sets in $\mathbf{R}^2$, then the principal surfaces of h are the only critical values of $D^2(f)$.

Let $\mathcal{F}^2(h) \subset G^2$ denote the class of principal 2-surfaces of h. Once again we consider a smooth subset of this class. Form the subset $\mathcal{F}_c^2(h) \subset \mathcal{F}^2(h)$, where $f \in \mathcal{F}_c^2(h)$ iff $f \in \mathcal{F}^2(h)$ and $\lambda_f(x)$ is continuous in x for all points x in the support of h. Surfaces in $\mathcal{F}_c^2(h)$ have the following properties.

- There is only one closest point on the principal surface for each x in the support of h.
- The surface is globally well behaved, in that it cannot fold back upon itself causing ambiguities in projection.
- We saw that for principal curves in $\mathcal{F}_c(h)$, there are no points at or beyond the centers of curvature of the curve. The analogous statement for principal surfaces in $\mathcal{F}_c^2(h)$ is that there are no points at or beyond the centers of normal curvature of any unit speed curve in the surface.

3.3. An algorithm for finding principal curves and surfaces.

We are still in the theoretical situation of finding principal curves or surfaces for a probability distribution. We will refer to curves (1-dimensional surfaces) and 2-dimensional surfaces jointly as surfaces in situations where the distinction is not important.

When seeking principal surfaces or critical values of $D^2(f)$, it is natural to look for a smooth curve that corresponds to a local minimum. Our strategy is to *start* with a smooth curve and then to look around it for a local minimum. Recall that

$$D^2(f) = \mathbf{E} \left\| X - f(\lambda_f(X)) \right\|^2 \tag{3.5}$$

$$= \mathbf{E}_{\lambda_f(X)} \mathbf{E} \left[\left\| X - f(\lambda_f(X)) \right\|^2 | \lambda_f(X) \right]. \tag{3.6}$$

We can write this as a minimization problem in f and λ: find f and λ such that

$$D_1^2(f, \lambda) = \mathbf{E} \left\| X - f(\lambda) \right\|^2 \tag{3.7}$$

is a minimum. Clearly, given any candidate solution f and λ, f and λ_f is at least as good. Two key ideas emerge from this:

- If we knew f as a function of λ, then we could minimize (3.7) by picking $\lambda = \lambda_f(x)$ at each point x in the support of h.

- Suppose, on the other hand, that we had a function $\lambda(x)$. We could rewrite (3.7) as:

$$D_1^2(f,\lambda) = \mathbf{E}_{\lambda(X)} \sum_{j=1}^{p} \mathbf{E}[(X_j - f_j(\lambda(X))^2 \,|\, \lambda(X)] \tag{3.8}$$

  We could minimize D_1^2 by choosing each f_j separately so as to minimize the corresponding term in the sum in (3.8) . This amounts to choosing

$$f_j(\lambda) = \mathbf{E}(X_j \,|\, \lambda(X) = \lambda). \tag{3.9}$$

In this last step we have to check that the new f is differentiable. One can construct many situations where this is not the case by allowing the starting curve to be globally wild. On the other hand, if the starting curve is well behaved, the sets of projection at a particular point in the curve or surface lie in the normal hyperplanes which vary smoothly. Since the density h is smooth we can expect that the conditional expectation in (3.9) will define a smooth function. We give more details in the next chapter. The above preamble motivates the following iterative algorithm.

Principal surface algorithm

initialization: Set $f^{(0)}(\lambda) = A\lambda$ where A is either a column vector (principal curves) and is the direction vector of the first linear principal component of h or A is a $p \times 2$ matrix (principal surfaces) consisting of the first two principal component direction vectors. Set $\lambda^{(0)} = \lambda_{f^{(0)}}$.

repeat: over iteration counter j

1) Set $f^{(j)}(\cdot) = \mathbf{E}(X \,|\, \lambda^{(j-1)}(X) = \cdot)$.

2) Choose $\lambda^{(j)} = \lambda_{f^{(j)}}$.

3) Evaluate $D^{2\,(j)} = D_1^2(f^{(j)}, \lambda^{(j)})$.

until: $D^{2\,(j)}$ fails to decrease.

Although we start with the linear principal component solution, any reasonable starting values can be used.

It is easy to check that the criterion $D^2(j)$ must converge. It is positive and bounded below by 0. Suppose we have $f^{(j-1)}$ and $\lambda^{(j-1)}$. Now $D_1^2(f^{(j)}, \lambda^{(j-1)}) \leq D_1^2(f^{(j-1)}, \lambda^{(j-1)})$ by the properties of conditional expectation. Also $D_1^2(f^{(j)}, \lambda^{(j)}) \leq D_1^2(f^{(j)}, \lambda^{(j-1)})$ since the $\lambda^{(j)}$ are chosen that way. Thus each step of the iteration is a decrease, and the criterion converges. This does not mean that the procedure has converged, since it is conceivable that the algorithm oscillates between two or more curves that are the same expected distance from the points. We have not found an example of this phenomenon.

The definition of principal surfaces is suggestive of the above algorithm. We want a smooth surface that is self consistent. So we start with the plane (line). We then check if it is indeed self consistent by evaluating the conditional expectation. If not we have a surface as a by-product. We then check if this is self consistent, and so on. Once the self consistency condition is met, we have a principal surface. By the theorem quoted above, this surface is a critical point of the distance function.

3.4. Principal curves and surfaces for data sets.

So far we have considered the principal curves and surfaces for a continuous multivariate probability distribution. In reality, we usually have a finite multivariate data set. How do we define the principal curves and surfaces for them? Suppose then that X is a $n \times p$ matrix of n observations on p variables. We regard the data set as a sample from an underlying probability distribution, and use it to estimate the principal curves and surfaces of that distribution. We briefly describe the ideas here and leave the details for chapters 5 and 6.

- The first step in the algorithm uses linear principal components as starting values. We use the sample principal components and their corresponding direction vectors as initial estimates of λ_f and $f^{(0)}$.

- Given functions $\hat{f}^{(j-1)}$ we can find for each x_i in the sample a value $\hat{\lambda}_i^{(j-1)} = \lambda_{\hat{f}^{(j-1)}}(x_i)$. This can be done in a number of ways, using numerical optimization techniques. In practice we have $\hat{f}^{(j-1)}$ evaluated at n values of λ, in fact at $\hat{\lambda}_1^{(j-2)}, \hat{\lambda}_2^{(j-2)}, \cdots, \hat{\lambda}_n^{(j-2)}$. $\hat{f}^{(j-1)}$ is evaluated at other points by interpolation. To illustrate the idea let us consider a curve for which we have $\hat{f}^{(j-1)}$ evaluated at $\hat{\lambda}_i^{(j-2)}$, for $i = 1, \cdots, n$. For each point i in the sample we can project x_i onto the line joining each pair $(\hat{f}^{(j-1)}(\hat{\lambda}_k^{(j-2)}), \hat{f}^{(j-1)}(\hat{\lambda}_{k+1}^{(j-2)}))$. Suppose the distance to the projection is d_{ik}, and if the point projects beyond either endpoint, then d_{ik} is the distance to the closest endpoint. Corresponding to each d_{ik} is a value $\lambda_{ik} \in [\hat{\lambda}_k^{(j-2)}, \hat{\lambda}_{k+1}^{(j-2)}]$. We then let $\hat{\lambda}_i^{(j-1)}$ be the λ_{ik} that

corresponds to the smallest value of d_{ik}. This is an $O(n^2)$ procedure, and as such is rather naive. We use it as an illustration and will describe more efficient algorithms later.

- We have to estimate $f^{(j)}(\lambda) = \mathbf{E}(X \mid \lambda^{(j-1)} = \lambda)$. We restrict ourselves to estimating this quantity at only n values of $\lambda^{(j-1)}$, namely $\hat{\lambda}_1^{(j-1)}, \cdots, \hat{\lambda}_n^{(j-1)}$ which we have already estimated. We require $\mathbf{E}(X \mid \lambda^{(j-1)} = \hat{\lambda}_i^{(j-1)})$. This says that we have to gather all the observations that project onto $\hat{f}^{(j-1)}$ at $\hat{\lambda}_i^{(j-1)}$, and find their mean. Typically we have only one such observation, namely x_i. It is at this stage that we introduce the *scatterplot smoother*, the fundamental building block in the principal curve and surface procedures for finite data sets. We estimate the conditional expectation at $\hat{\lambda}_i^{(j-1)}$ by averaging all the observations x_k in the sample for which $\hat{\lambda}_k^{(j-1)}$ is close to $\hat{\lambda}_i^{(j-1)}$. As long as these observations are close enough and the underlying density is smooth, the bias introduced will be small. On the other hand, the variance of the estimate decreases as we include more observations in the neighborhood. Figure (3.6) demonstrates this local averaging. Once again we have just given the ideas here, and will go into details in later chapters.

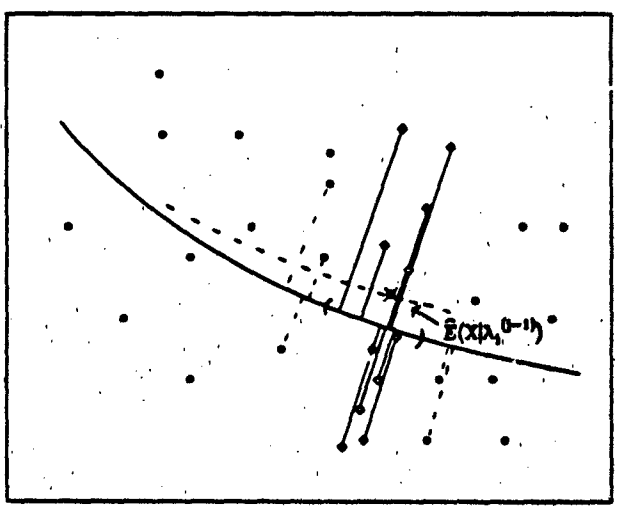


Figure 3.6 We estimate the conditional expectation $\mathbf{E}(X \mid \lambda^{(j-1)} = \hat{\lambda}_i^{(j-1)})$ by averaging the observations x_k for which $\hat{\lambda}_k^{(j-1)}$ is close to $\hat{\lambda}_i^{(j-1)}$.

- One property of scatterplot smoothers in general is that they produce smooth curves and surfaces as output. The larger the neighborhood used for averaging, the smoother the output. Since we are trying to estimate differentiable curves and surfaces, it is convenient that our algorithm, in seeking a conditional expectation estimate, does produce smooth estimates. We will have to worry about how smooth these estimates should be, or rather how big to make the neighborhoods. This becomes a variance versus bias tradeoff, a familiar issue in non-parametric regression.

- Finally, we estimate $D^2(j)$ in the obvious way, by adding up the distances of each point in the sample from the current curve or surface.

3.5. Demonstrations of the procedures.

We look at two examples, one for curves and one for surfaces. They both are generated from an underlying *true* model so that we can easily check that the procedures are doing the correct thing.

3.5.1. The circle in two-space.

The series of plots in figure 3.7 show 100 data points generated from a circle in 2 dimensions with independent Gaussian errors in both coordinates. In fact, the generating functions are

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5\sin(\lambda) \\ 5\cos(\lambda) \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \tag{3.10}$$

where λ is uniformly distributed on $[0, 2\pi]$ and e_1 and e_2 are independent $\mathcal{N}(0,1)$.

The solid curve in each picture is the estimated curve for the iteration as labelled, and the dashed curve is the true function. The starting curve is the first principal component, in figure 3.7b. Figure 3.7a gives the usual scatterplot smooth of x_2 against x_1, which is clearly an inappropriate summary for this constructed data set.

The curve in figure 3.7k does substantially better than the previous iterations. The figure caption gives us a clue why — the span of the smoother is reduced. This means that the size of the neighborhood used for local averaging is smaller. We will see in the next chapter how the bias in the curves depends on this span.

The square root of the average squared orthogonal distance is displayed at each iteration. If the true curve was linear the expected orthogonal distance for any point would be $\sqrt{\mathbf{E}X_1^2} = 1$. We will see in chapter 4 that for this situation, the true circle does not

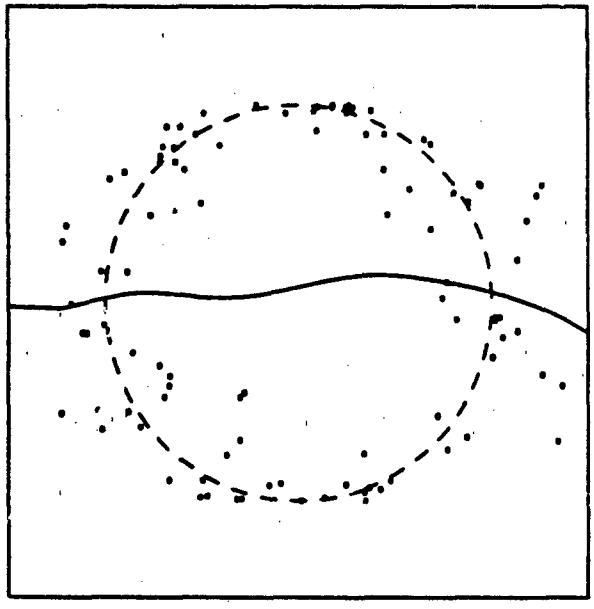


Figure 3.7a The dashed curve is the usual scatterplot smooth. $D(S) = 3.35$

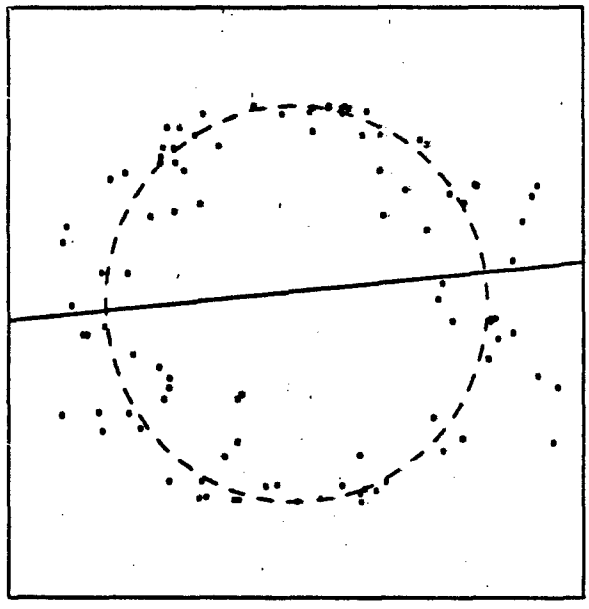


Figure 3.7b The dashed curve is the principal component line. $D(\hat{f}^{(0)}) = 3.43$

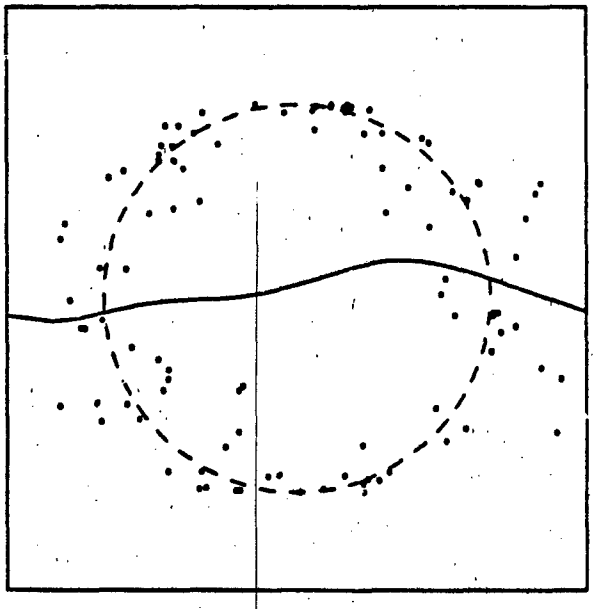


Figure 3.7c $D(\hat{f}^{(1)}) = 3.34$

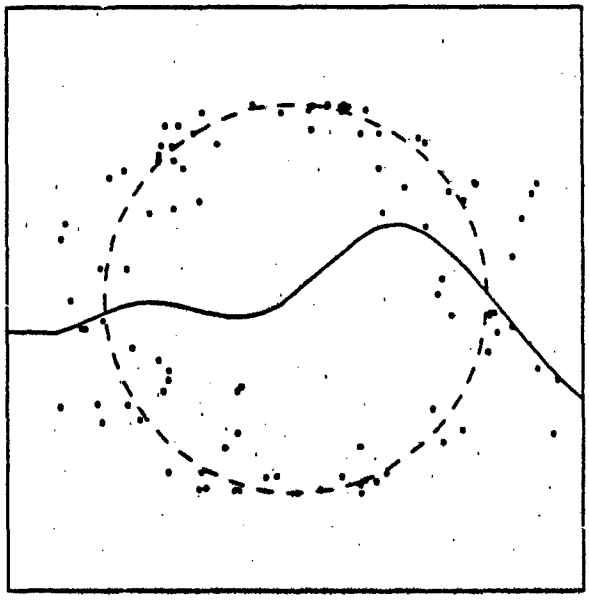


Figure 3.7d $D(\hat{f}^{(2)}) = 3.03$

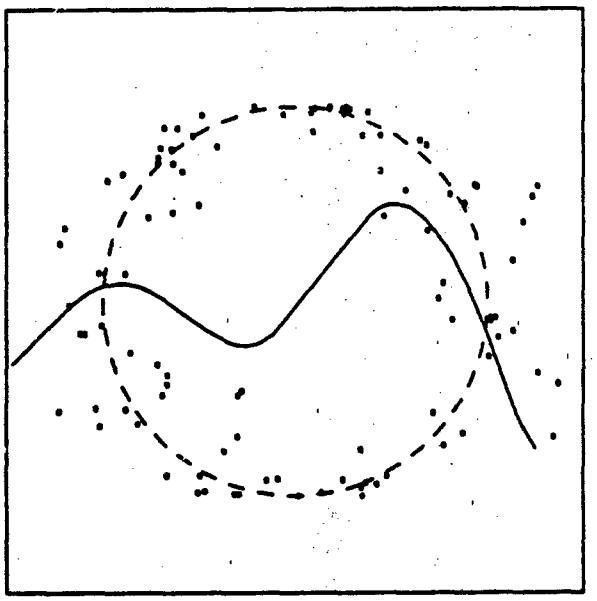


Figure 3.7e $D(\hat{f}^{(3)}) = 2.64$

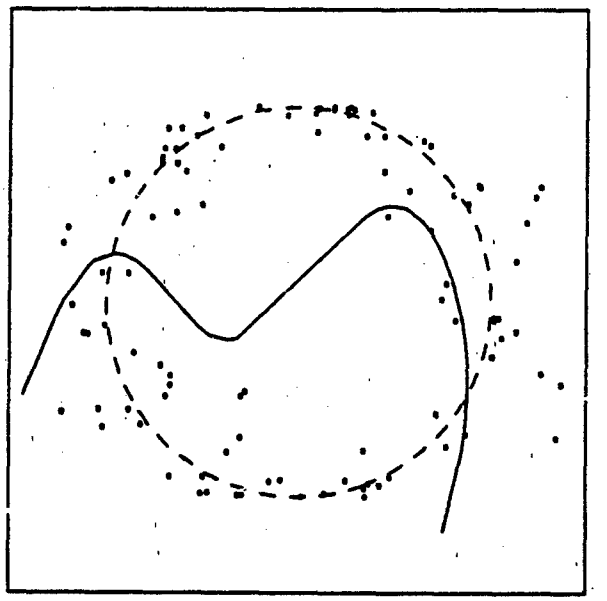


Figure 3.7f $D(\hat{f}^{(4)}) = 2.37$

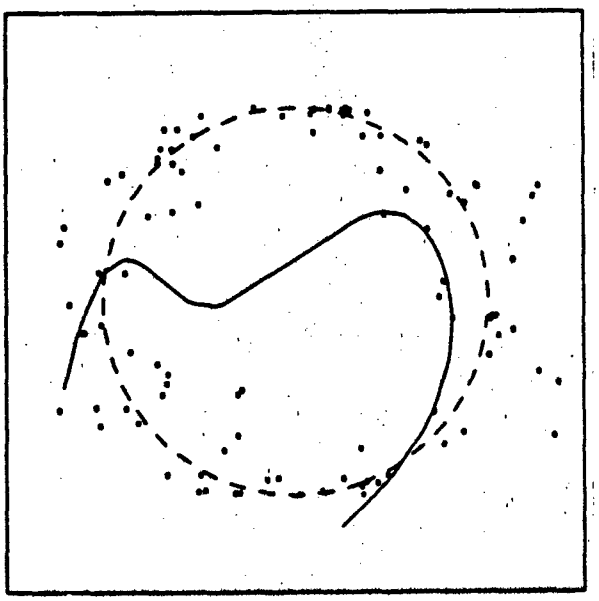


Figure 3.7g $D(\hat{f}^{(5)}) = 2.25$

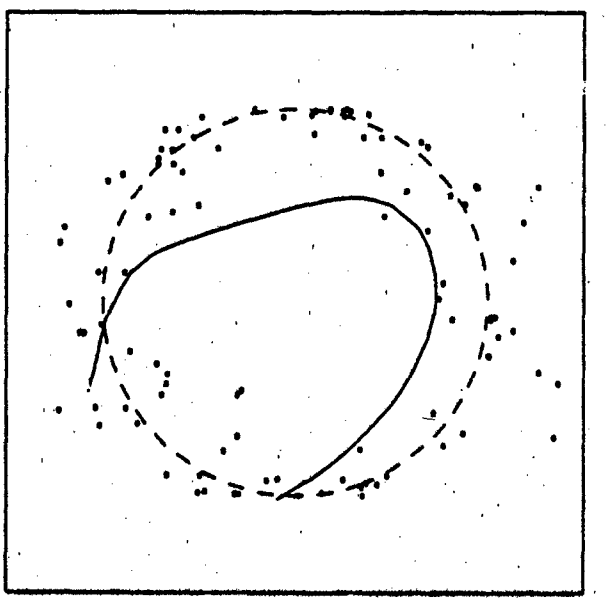


Figure 3.7h $D(\hat{f}^{(6)}) = 1.91$

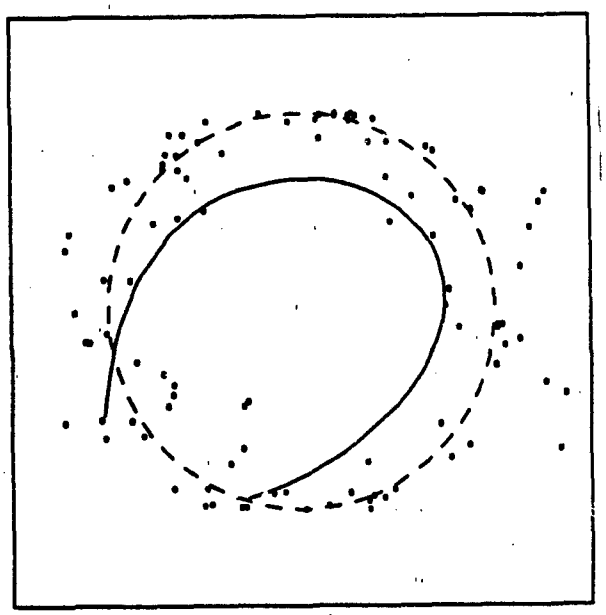


Figure 3.7i $D(\hat{f}^{(7)}) = 1.64$

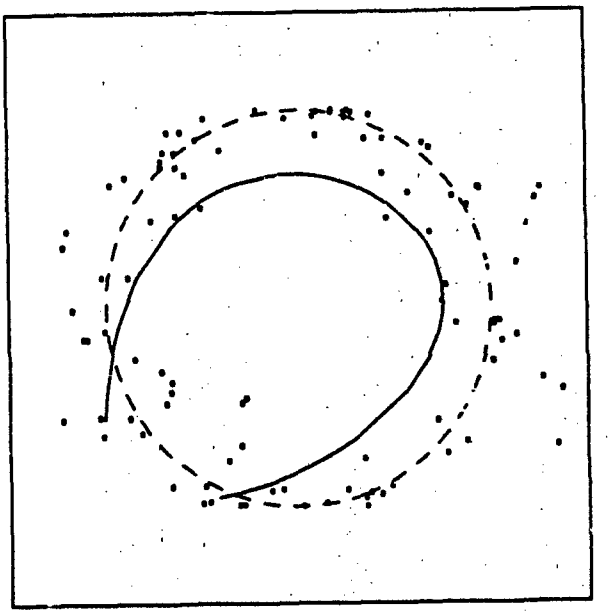


Figure 3.7j $D(\hat{f}^{(8)}) = 1.60$

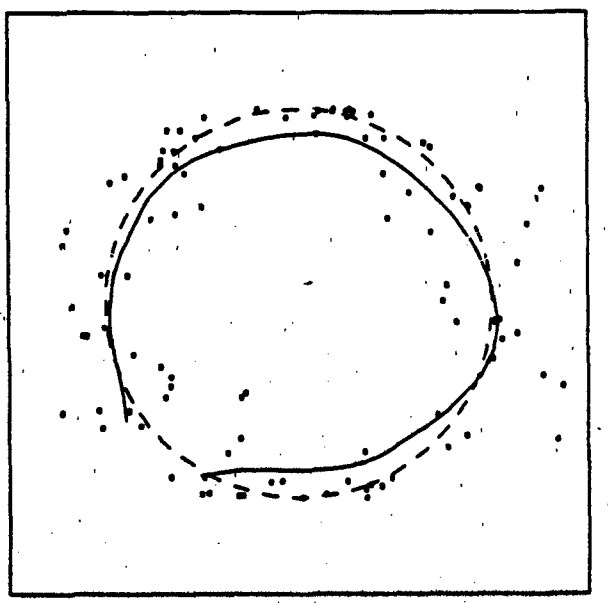


Figure 3.7k $D(\hat{f}^{(9)}) = 0.97$. **The span is automatically reduced at this stage.**

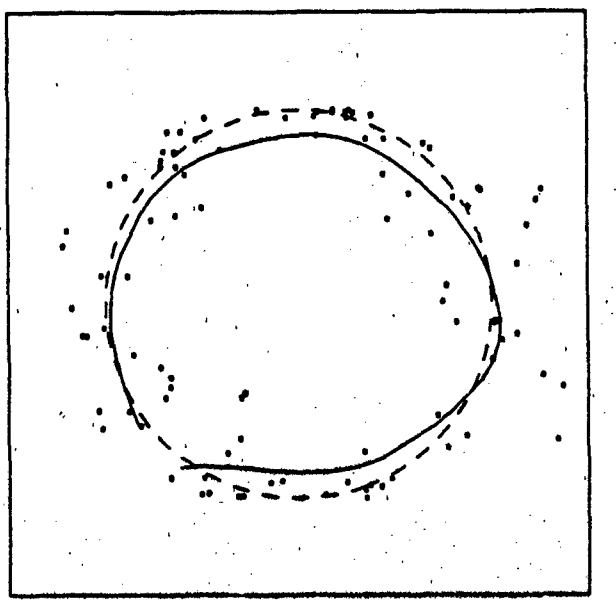


Figure 3.7l $D(\hat{f}^{(10)}) = 0.96$

minimize the distance, but rather a circle with slightly larger radius. Then the minimizing distance is approximately $\sigma^2(1 - 1/4\rho^2) = .99$. Our final distance is even lower. We still have to adjust for the overfit factor or number of parameters used up in the fitting procedure. This deflation factor is of the order $n/(n-q)$ where q is the number of parameters. In linear principal components we know q. In chapter 6 we suggest some rule of thumb approximations for q in this non-parametric setting.

This example presents the principal curve procedure with a particularly tough job. The starting value is wholly inappropriate and the projection of the points onto this line does not nearly represent the final ordering of the points projected onto the solution curve. At each iteration the coordinate system for the $\hat{\lambda}^{(j)}$ is transferred from the previous curve to the current curve. Points initially project in a certain order on the starting vector, as depicted in figure 3.8a. The new curve is a function of $\hat{\lambda}^{(0)}$ measured along this vector as in figure 3.8b obtained by averaging the coordinates of points local in $\hat{\lambda}^{(0)}$. The new $\hat{\lambda}^{(1)}$ values are found by projecting the points onto the new curve. It can be seen that the ordering of the projected points along the new curve can be very different to the ordering along the previous curve. This enables the successive curves to bend to shapes that could not be parametrized in the original principal component coordinate system.

3.5.2. The half-sphere in three-space.

Figure 3.9 shows 150 points generated from the surface of the half-sphere in 3-D. The simulated model in polar co-ordinates is

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5\sin(\lambda_1)\cos(\lambda_2) \\ 5\cos(\lambda_1)\cos(\lambda_2) \\ 5\sin(\lambda_2) \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix} \tag{3.11}$$

for $\lambda_1 \in [0, 2\pi]$ and $\lambda_2 \in [0, \pi/2)$. The vector e of errors is simulated from a $N(0, I)$ distribution, and the values of λ_1 and λ_2 are chosen so that the points are distributed uniformly in the surface. Figure 3.9a shows the data and the generating surface. The expected distance of the points from the generating half-sphere is to first order 1, which is the expected squared length of the residual when projecting a spherical standard gaussian 3-vector onto a plane through the origin. Ideally we would display this example on a motion graphics workstation in order to see the 3 dimensions.*

---

* This dissertation is accompanied by a motion graphics movie called *Principal Curves and Surfaces*. The half-sphere is one of 4 examples demonstrated in the movie.

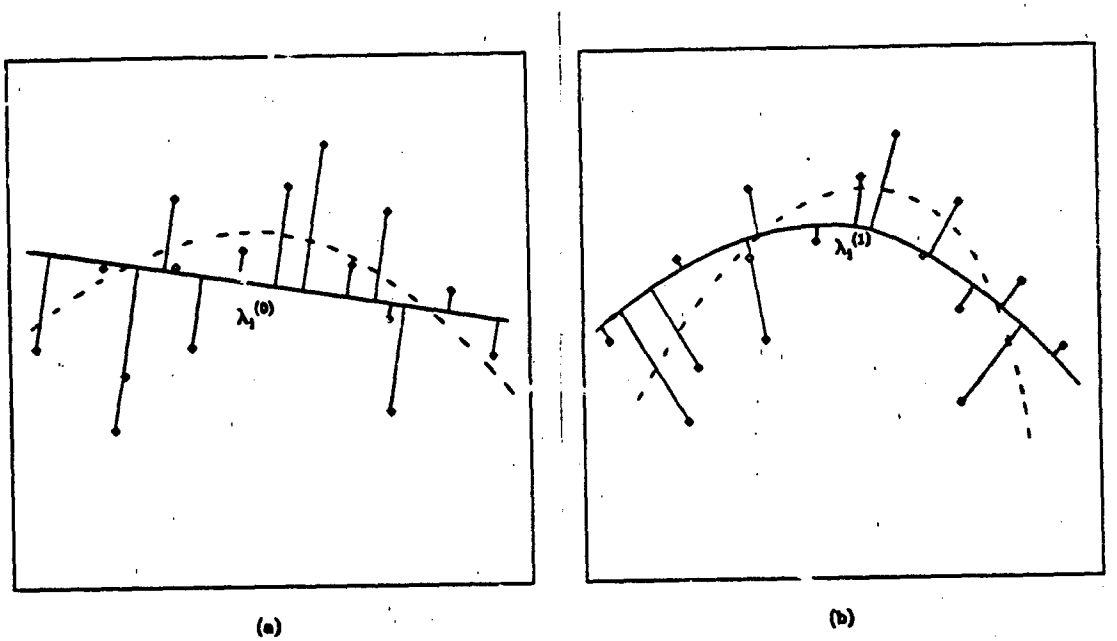


Figure 3.8 The curve of the the first iteration is a function of $\hat{\lambda}^{(0)}$ measured along the starting vector (a). The curve of the the second iteration is a function of $\hat{\lambda}^{(1)}$ measured along the curve of the first iteration (b).

3.6. Principal surfaces and principal components.

In this section we draw some comparisons between the principal curve and surface models and their linear counterparts in addition to those already mentioned.

3.6.1. A Variance decomposition.

Usually linear principal components are approached via variance considerations. The first component is that linear combination of the variables with the largest variance. The second component is uncorrelated with the first and has largest variance subject to this constraint. Another way of saying this is that the total variance in the plane spanned by the first two components is larger than that in any other plane. By total variance we mean the sum of the variances of the data projected onto any orthonormal basis of the subspace defined by the plane. The following treatment is for one component, but the ideas easily generalize to two.

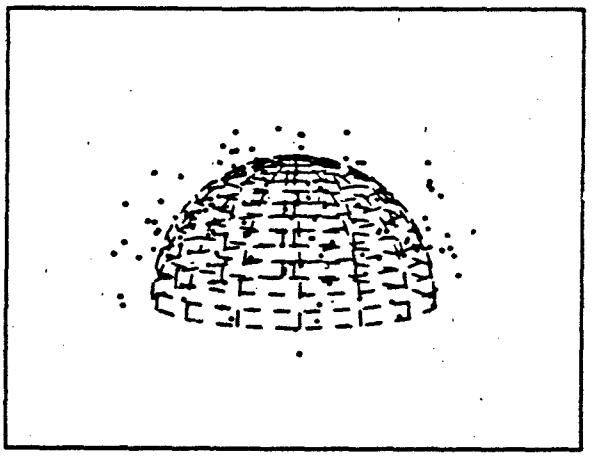


Figure 3.9a. The generating surface and the data. $D(S) = 1.0$

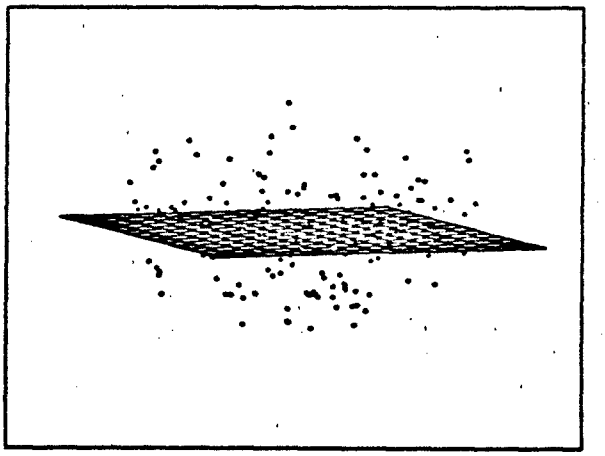


Figure 3.9b. The principal component plane. $D(\hat{f}^{(0)}) = 1.59$

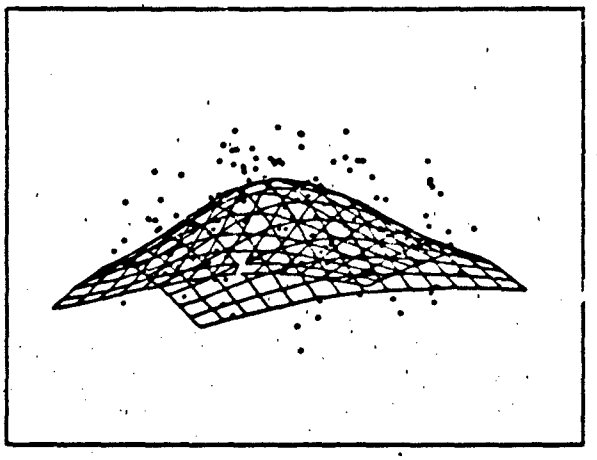


Figure 3.9c. $D(\hat{f}^{(1)}) = 1.20$

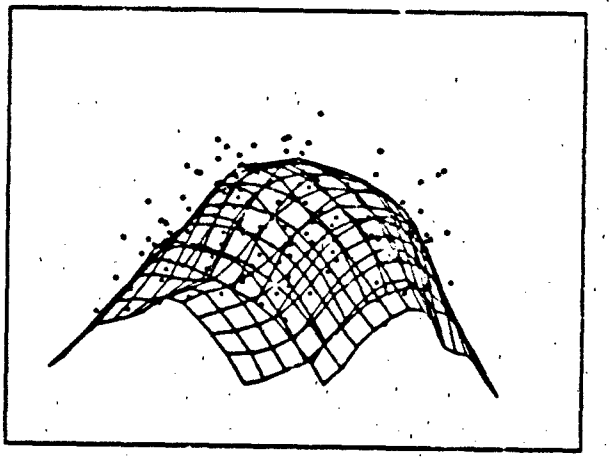


Figure 3.9d. $D(\hat{f}^{(4)}) = 0.78$

If $\lambda = (\lambda_1, \ldots, \lambda_n)'$ is the first principal component of X, a $n \times p$ data matrix, and a is the corresponding direction vector, then the following variance decomposition is easily derived:

$$\sum_{j=1}^{p} \mathrm{Var}(x_j) = \mathrm{Var}(\lambda) + E\|x - a\lambda\|^2 \tag{3.12}$$

where $\mathrm{Var}(\cdot)$ and $E(\cdot)$ refer to sample variance and expectation. If the principal component was defined in the parent population then the result is still true and $\mathrm{Var}(\cdot)$ and $E(\cdot)$ have their usual meaning. The second term on the right of (3.12) is the expected squared distance of a point to its projection onto the principal direction.*

The total variance in the original p variables is decomposed into two components: the variance explained by the linear projection and the residual variance in the distances from the points to their projections. We would like to have a similar decomposition for principal curves and surfaces.

Let w now be any random variable. Standard results on conditional expectation show that:

$$\sum_{j=1}^{p} \mathrm{Var}(x_j) = \sum_{j=1}^{p} \mathbf{E}(x_j - \mathbf{E}(x_j \mid w))^2 + \sum_{j=1}^{p} \mathrm{Var}(\mathbf{E}(x_j \mid w)). \tag{3.13}$$

If $w = \lambda_f(x)$ and f is a principal curve so that $\mathbf{E}(x_j \mid \lambda_f(x)) = f_j(\lambda_f(x))$, we have

$$\sum_{j=1}^{p} \mathrm{Var}(x_j) = \mathbf{E}\left\|x - f(\lambda_f(x))\right\|^2 + \sum_{j=1}^{p} \mathrm{Var}(f_j(\lambda_f(x))). \tag{3.14}$$

This gives us an analogous result to (3.12) in the distributional case. That is, the total variance in the p coordinates is decomposed into the variance explained by the true curve and the residual variance in the expected squared distance from a point to its true position on the curve. The sample version of (3.14) holds only approximately:

$$\sum_{j=1}^{p} \mathrm{Var}(x_j) \approx \sum_{i=1}^{n} \left\|x_i - \hat{f}(\hat{\lambda}_i)\right\|^2 + \sum_{j=1}^{p} \mathrm{Var}(\hat{f}_j(\hat{\lambda}_i)). \tag{3.15}$$

The reason for this is that most practical scatterplot smoothers are not projections, whereas conditional expectations are.

We make the following observations:

---

* We keep in mind that X is considered to be centered, or alternativly that $\mathbf{E}(x) = 0$. The above results are still true if this is not the case, but the equations are messier.

- if $f_j(\lambda) = a_j\lambda$, the linear principal component function, then

$$\sum_{j=1}^{p} \mathrm{Var}(f_j(\lambda_f(x))) = \sum_{j=1}^{p} a_j^2 \mathrm{Var}(\lambda_a(x)) = \mathrm{Var}(\lambda)$$

since a has length 1. Here we have written λ for the function $\lambda_a(x) = a'x$.

- if the f_j are approximately linear we can use the *Delta* method to obtain

$$\sum_{j=1}^{p} \mathrm{Var}(f_j(\lambda_f(x))) \approx \sum_{j=1}^{p} (f_j'(\mathrm{E}(\lambda_f(x)))^2 \mathrm{Var}(\lambda_f(x)) = \mathrm{Var}(\lambda_f(x))$$

since we restrict our curves to be unit speed and thus we have have $\|f'\| = 1$.

3.6.2. The power method.

We already mentioned that when the data is ellipsoidal the principal curve procedure yields linear principal components. We now show that if our smoother fits straight lines, then once again the principal curve procedure yields linear principal components irrespective of the starting line.

Theorem 3.1

If the smoother in the principal curve procedure produces least squares straight line fits, and if the initial functions describe a straight line, then the procedure converges to the first principal component.

Proof

Let $a^{(0)}$ be any starting vector which has unit length and is not orthogonal to the largest principal component of X, and assume X is centered. We find $\lambda_i^{(0)}$ by projecting x_i onto $a^{(0)}$ which we denote collectively by

$$\lambda^{(0)} = Xa^{(0)}$$

where $\lambda^{(0)}$ is a n vector with elements $\lambda_i^{(0)}$, $i = 1, \ldots, n$. We find $a_j^{(1)}$ by regressing or projecting the vector $x_j = (x_{1j}, \ldots, x_{nj})'$ onto $\lambda^{(0)}$:

$$a_j^{(1)} = \frac{\lambda^{(0)\prime} x_j}{\lambda^{(0)\prime}\lambda^{(0)}}$$

or

$$a^{(1)} = \frac{\lambda^{(0)\prime} X}{\lambda^{(0)\prime}\lambda^{(0)}}$$

$$= \frac{X'Xa^{(0)}}{a^{(0)\prime}X'Xa^{(0)}}$$

and $a^{(1)}$ is renormalized. It can now be seen that iteration of this procedure is equivalent to finding the largest eigenvector of $X'X$ by the power method (Wilkinson 1965). ∎

Chapter 4

Theory for principal curves and surfaces

In this chapter we prove the results referred to in chapter 3. In most cases we deal only with the principal curve model, and suggest the analogues for the principal surface model.

4.1. The projection index is measureable.

Since the first thing we do is condition on $\lambda_f(X)$, it might be prudent to check that it is indeed a random variable. To this end we need to show that the function $\lambda_f : \mathbb{R}^p \mapsto \mathbb{R}^1$ is measureable. *

Let $f(\lambda)$ be a unit speed parameterized continuous curve in p-space, defined for $\lambda \in [\lambda_0, \lambda_1] = \Lambda$. Let

$$D(x) = \inf_{\lambda \in \Lambda} \{d(x, f(\lambda))\} \quad \forall x \in \mathbb{R}^p$$

where

$$d(x, f(\lambda)) = \|x - f(\lambda)\|,$$

the usual euclidean distance between two vectors. Now set

$$M(x) = \{\lambda ; d(x, f(\lambda)) = D(x)\}.$$

Since Λ is compact, $M(x)$ is not empty. Since f, and hence $d(x, f(\lambda))$ is continuous, $M^c(x)$ is open, and hence $M(x)$ is closed. Finally, for each x in $\mathbb{R}^p$ we define the projection index:

$$\lambda_f(x) = \sup M(x)$$

$\lambda_f(x)$ is attained because $M(x)$ is closed, and we have avoided ambiguities.

Theorem 4.1

$\lambda_f(x)$ is a measureable function of x.

---

* I am grateful to H. Künsch of ETH, Zürich, for getting me started on this proof.

Proof

In order to prove that $\lambda_f(x)$ is measureable we need to show that for any $c \in \Lambda$, the set $\{x \,|\, \lambda_f(x) \leq c\}$ is a measureable set.

Now $x \in \{x \,|\, \lambda_f(x) \leq c\} \iff$ for any $\lambda \in (c, \lambda_1]$ there exists a $\lambda' \in [\lambda_0, c]$ such that $d(x, f(\lambda)) > d(x, f(\lambda'))$. (i.e. if there was equality then by our convention we choose $\lambda_f(x) = \lambda > c$.) In symbols we have

$$\{x \,|\, \lambda_f(x) \leq c\} = \bigcap_{\lambda \in (c,\lambda_1]} \bigcup_{\lambda' \in [\lambda_0, c]} \{x \,|\, d(x, f(\lambda)) > d(x, f(\lambda'))\}$$
$$\stackrel{\text{def}}{=} A_c$$

The first step in the proof is to show that

$$B_c \stackrel{\text{def}}{=} \bigcap_{\lambda \in (c,\lambda_1]} \bigcup_{\lambda'_q \in [\lambda_0, c] \cap Q} \{x \,|\, d(x, f(\lambda)) > d(x, f(\lambda'_q))\}$$
$$= A_c$$

where Q is the set of rational numbers. Since for each λ

$$\bigcup_{\lambda' \in [\lambda_0, c]} \{x \,|\, d(x, f(\lambda)) > d(x, f(\lambda'))\} \supseteq \bigcup_{\lambda'_q \in [\lambda_0, c] \cap Q} \{x \,|\, d(x, f(\lambda)) > d(x, f(\lambda'_q))\},$$

it follows that $B_c \subseteq A_c$. We need to show that $B_c \supseteq A_c$. Suppose $x \in A_c$ i.e. for any given $\lambda \in (c, \lambda_1]$ $\exists\, \lambda' \in [\lambda_0, c]$ such that

$$d(x, f(\lambda)) > d(x, f(\lambda')).$$

For any given such λ and λ' we can find an $\epsilon > 0$ such that

$$d(x, f(\lambda)) = d(x, f(\lambda')) + \epsilon$$

Now since f is continuous and the rationals are dense in $\mathbf{R}^1$ we can find a $\lambda'_q \in Q$ such that $\lambda'_q \leq \lambda'$ and $d(f(\lambda'), f(\lambda'_q)) < \epsilon$. (If $\lambda' \in Q$ we need go no further). This implies that $d(x, f(\lambda)) > d(x, f(\lambda'_q))$ by the pythagorean property of euclidean distance. This in turn implies that $x \in B_c$ and thus $A_c \subseteq B_c$, and therefore $A_c = B_c$.

The second step is to show that

$$D_c \stackrel{\text{def}}{=} \bigcap_{\lambda_q \in (c,\lambda_1] \cap Q} \bigcup_{\lambda'_q \in [\lambda_0, c] \cap Q} \{x \,|\, d(x, f(\lambda_q)) > d(x, f(\lambda'_q))\}$$
$$= B_c$$

Now clearly $B_c \subseteq D_c$. Suppose then that $x \in D_c$, i.e. for every $\lambda_q \in (c, \lambda_1] \cap Q$, there is a $\lambda'_q \in [\lambda_0, c] \cap Q$ such that $d(x, f(\lambda_q)) > d(x, f(\lambda'_q))$. Once again by continuity of f and because the rationals are dense in $\mathbb{R}^1$ we can find another $\lambda^*_q \in Q$, $\lambda^*_q > \lambda_q$ such that

$$d(x, f(\lambda)) > d(x, f(\lambda'_q))\}$$

for all $\lambda \in [\lambda_q, \lambda^*_q]$. This means that

$$x \in \bigcap_{\lambda \in [\lambda_q, \lambda^*_q]} \bigcup_{\lambda'_q \in [\lambda_0, c] \cap Q} \{x \,|\, d(x, f(\lambda)) > d(x, f(\lambda'_q))\}$$
$$\stackrel{\text{def}}{=} E_{\lambda_q, \lambda^*_q}$$

for every $\lambda_q \in (c, \lambda_1] \cap Q$. In other words

$$x \in \bigcap_{\lambda_q \in (c,\lambda_1] \cap Q} E_{\lambda_q, \lambda^*_q}$$
$$= B_c$$

and we have that $D_c = B_c$. Finally, each of the sets in D_c is a half space, and thus measureable, D_c is a countable union and intersection of measurable sets, and is thus itself measurable. ∎

4.2. The stationarity property of principal curves.

We first prove a result for straight lines. This will lead into the result for curves. The straight line theorem says that a principal component line is a critical point of the expected distance from the points to itself. The converse is also true.

We first establish some more notation. Suppose $f(\lambda) : \Lambda \mapsto \mathcal{G}$ is a unit speed continuously differentiable parametrized curve in $\mathbb{R}^p$, where Λ is an interval in $\mathbb{R}^1$. Let $g(\lambda)$ be defined similarly, without the unit speed restriction. An ϵ perturbed version of f is $f_\epsilon \stackrel{\text{def}}{=} f(\lambda) + \epsilon g(\lambda)$. Suppose X has a continuous density in $\mathbb{R}^p$ which we denote by h, and

let $D^2(h, f_\epsilon)$ be defined as before by

$$D^2(h, f_\epsilon) = \mathbf{E}_h \left\| X - f_\epsilon(\lambda_{f_\epsilon}(X)) \right\|^2$$

where $\lambda_{f_\epsilon}(X)$ parametrizes the point on f_ϵ closest to X.

Definition

The curve f is a *critical point of the distance function* in the class $\mathcal{G}$ iff

$$\left.\frac{dD^2(h, f_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = 0 \ \forall g \in \mathcal{G}.$$

(We have to show that this derivative exists.)

Theorem 4.2

Let $f(\lambda) = \mathbf{E}X + \lambda v_0$ with $\|v_0\| = 1$, and suppose we restrict $g(\lambda)$ to be linear as well. So $g(\lambda) = \lambda v$, $\|v\| = 1$ and $\mathcal{G} = \mathcal{L}$, the class of all unit speed straight lines. Then f is a critical point of the distance function in $\mathcal{L}$ iff v_0 is an eigenvector of $\Sigma = \mathbf{COV}(X)$.

Note:

- WLOG we assume that $\mathbf{E}X = 0$.
- $\|v\| = 1$ is simply for convenience.

Proof

The closest point from x to any line λw through the origin is found by projecting x onto w and has parameter value

$$\lambda_w(x) = \frac{x'w}{\|w\|}$$

Th n

$$\begin{aligned} d^2(x, \lambda w) &= \left\| x - \frac{ww'x}{\|w\|^2} \right\|^2 \\ &= \|x\|^2 - \frac{[w'x]^2}{w'w} \end{aligned}$$

Upon taking expected values we get

$$D^2(h, \lambda w) = \operatorname{tr} \Sigma - \frac{w'\Sigma w}{w'w}. \tag{4.1}$$

We now apply the above to f_ϵ instead of w, but first make a simplifying assumption. We can assume w.l.o.g that $v_0 = e_1$ since the problem is invariant to rotations.

We split v into a component $v_e = ce_1$ along e_1 and an orthogonal component v^*. Thus $v = cv_e + v^*$ where $e_1' v^* = 0$. So $f_\epsilon = \lambda((1+c\epsilon)e_1 + \epsilon v^*)$. We now plug this into (4.1) to get

$$\begin{aligned} D^2(h, f_\epsilon) &= \text{tr } \Sigma - \frac{((1+c\epsilon)e_1 + \epsilon v^*)'\Sigma((1+c\epsilon)e_1 + \epsilon v^*)}{(1+c\epsilon)^2 + \epsilon^2} \\ &= \text{tr } \Sigma - \frac{(1+c\epsilon)^2 e_1'\Sigma e_1 + 2\epsilon(1+c\epsilon)e_1'\Sigma v^* + \epsilon^2 v^{*\prime}\Sigma v^*}{(1+c\epsilon)^2 + \epsilon^2} \end{aligned} \tag{4.2}$$

Differentiating w.r.t. ϵ and setting $\epsilon = 0$ we get

$$\left.\frac{dD^2(h, f_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = -2e_1'\Sigma v^*.$$

If e_1 is a principal component of Σ then this term is zero for all v^* and hence for all v. Alternatively, if this term, and hence the derivative, is zero for all v and hence all $v^{*\prime}e_1 = 0$, we have

$$\begin{aligned} &v^{*\prime}\Sigma e_1 = 0 \ \forall\, v^{*\prime}e_1 = 0 \\ \Rightarrow &\Sigma e_1 = ce_1 \\ \Rightarrow &e_1 \text{ is an eigenvector of } \Sigma \end{aligned}$$

■

Note:

Suppose v is in fact another eigenvector of Σ, with eigenvalue d, then

$$D^2(h, f_\epsilon) - D^2(h, f) = \frac{\epsilon^2}{1+\epsilon^2}(\sigma_1^2 - d^2)$$

This shows that f might be a maximum, a minimum or a saddle point.

Theorem 4.3

Let $\mathcal{G}$ be the class of unit speed differentiable curves defined on Λ, a closed interval of the form $[a, b]$. The curve f is a principal curve of h iff f is a critical point of the distance function in the class $\mathcal{G}$.

We make some observations before we prove theorem 4.3. Figure 4.1illustrates the situation. The curve f_ϵ wiggles about f and approaches f as ϵ approaches 0. In fact, we can see that the curvature of f_ϵ is close to that of f for small ϵ. The curvature of f_ϵ is given by

$$1/r_{f_\epsilon}(\lambda) = \frac{f_\epsilon''(\lambda) \cdot N(\lambda)}{\|f_\epsilon'(\lambda)\|^2}$$

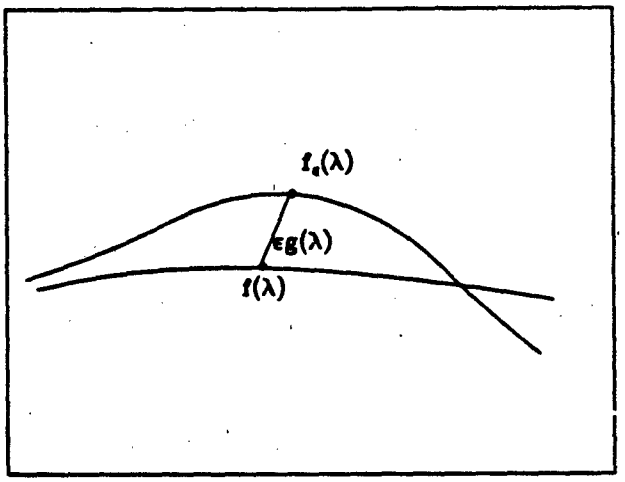


Figure (4.1) $f_\epsilon(\lambda)$ depicted as a function of $f(\lambda)$.

where $N(\lambda)$ is the normal vector to the curve at λ. Thus $1/r_{f_\epsilon}(\lambda) \leq \|f_\epsilon''(\lambda)\| / \|f_\epsilon'(\lambda)\|^2$ since the curve is not unit speed and so the acceleration vector is slightly off normal. Therefore we have $r_{f_\epsilon}(\lambda) \geq \|f'(\lambda) + \epsilon g'(\lambda)\|^2 / \|f''(\lambda) + \epsilon g''\|$ which converges to $r_f(\lambda)$ as $\epsilon \to 0$.

The theorem is stated only for curves f defined on compact sets. This is not such a restriction as it might seem at first glance. The notorious *space filling* curves are excluded, but they are of little interest anyway. If the density h has infinite support, we have to *box* it in $\mathbb{R}^p$ in order that f, defined on a compact set, can satisfy either statement of the theorem. (We show this later.) In practice this is not a restriction.

Proof of theorem 4.3.

We use the dominated convergence theorem (Chung, 1974 pp 42) to show that we can interchange the orders of integration and differentiation in the expression

$$\frac{d}{d\epsilon} D^2(h, f_\epsilon) = \frac{d}{d\epsilon} \mathbb{E}_h \left\| X - f_\epsilon(\lambda_{f_\epsilon}(X)) \right\|^2. \tag{4.3}$$

We need to find a random variable Y which is integrable and dominates almost surely the absolute value of

$$Z_\epsilon = \frac{\left\| X - f_\epsilon(\lambda_{f_\epsilon}(X)) \right\|^2 - \left\| X - f(\lambda_f(X)) \right\|^2}{\epsilon}$$

for all $\epsilon \geq 0$. Notice that by definition

$$\lim_{\epsilon \to 0} Z_\epsilon = \frac{d}{d\epsilon} \left\| X - f_\epsilon(\lambda_{f_\epsilon}(X)) \right\|^2 \Big|_{\epsilon=0}$$

if this limit exists. Now

$$Z_\epsilon \leq \frac{\left\| X - f_\epsilon(\lambda_f(X)) \right\|^2 - \left\| X - f(\lambda_f(X)) \right\|^2}{\epsilon}.$$

Expanding the first norm we get

$$\left\| X - f_\epsilon(\lambda_f(X)) \right\|^2 = \left\| X - f(\lambda_f(X)) \right\|^2 + \epsilon^2 \left\| g(\lambda_f(X)) \right\|^2 - 2\epsilon \left(X - f(\lambda_f(X)) \right) \cdot g(\lambda_f(X)),$$

and thus

$$\begin{aligned} Z_\epsilon &\leq -2 \left(X - f(\lambda_f(X)) \right) \cdot g(\lambda_f(X)) + \epsilon \left\| g(\lambda_f(X)) \right\|^2 \\ &\leq Y_1 \end{aligned}$$

where Y_1 is some bounded random variable.

Similarly we have

$$Z_\epsilon \geq \frac{\left\| X - f_\epsilon(\lambda_{f_\epsilon}(X)) \right\|^2 - \left\| X - f(\lambda_{f_\epsilon}(X)) \right\|^2}{\epsilon}.$$

We expand the first norm again, and get

$$\begin{aligned} Z_\epsilon &\geq -2 \left(X - f(\lambda_{f_\epsilon}(X)) \right) \cdot g(\lambda_{f_\epsilon}(X)) + \epsilon \left\| g(\lambda_{f_\epsilon}(X)) \right\|^2 \\ &\geq Y_2 \end{aligned}$$

where Y_2 is once again some bounded random variable. These two bounds satisfy the conditions of the dominated convergence theorem, and so the interchange is justified. However, from the form of the two bounds, and because f and g are continuous functions, we see that the limit $\lim_{\epsilon \to 0} Z_\epsilon$ exists whenever $\lambda_{f_\epsilon}(X)$ is continuous in ϵ at $\epsilon = 0$. Moreover, this limit is given by

$$\begin{aligned} \lim_{\epsilon \to 0} Z_\epsilon &= \frac{d}{d\epsilon} \left\| X - f_\epsilon(\lambda_{f_\epsilon}(X)) \right\|^2 \Big|_{\epsilon=0} \\ &= -2 \left(X - f(\lambda_f(X)) \right) \cdot g(\lambda_f(X)). \end{aligned}$$

We show in lemma 4.3.1 that this continuity condition is met almost surely.

We denote the distribution function of $\lambda_f(X)$ by h_λ, and get

$$\frac{d}{d\epsilon}D^2(h, f_\epsilon)\Big|_{\epsilon=0} = -2\,\mathbf{E}_{h_\lambda}\Big(\mathbf{E}(X\,|\,\lambda_f(X)=\lambda) - f(\lambda)\Big)\cdot g(\lambda). \tag{4.4}$$

If $f(\lambda)$ is a principal curve of h, then $\mathbf{E}(X\,|\,\lambda_f(X)=\lambda) = f(\lambda)$ for all λ in the support of h_λ, and thus

$$\frac{d}{d\epsilon}D^2(h, f_\epsilon)\Big|_{\epsilon=0} = 0 \quad \forall \text{ differentiable } g.$$

Alternatively, suppose that

$$\mathbf{E}_{h_\lambda}\Big(\mathbf{E}(X - f(\lambda)\,|\,\lambda_f(X)=\lambda)\cdot g(\lambda)\Big) = 0 \tag{4.5}$$

for all differentiable g. In particular we could pick $g(\lambda) = \mathbf{E}(X\,|\,\lambda_f(X)=\lambda) - f(\lambda)$. Then

$$\mathbf{E}_\lambda\Big\|\,\mathbf{E}(X\,|\,\lambda_f(X)=\lambda) - f(\lambda)\Big\|^2 = 0$$

and consequently f is a principal curve. This choice of g, however, might not be differentiable, so some approximation is needed.

Since (4.5) holds for all differentiable g we can use different g's to *knock off* different pieces of $\mathbf{E}(X\,|\,\lambda_f(X)=\lambda) - f(\lambda)$. In fact we can do it one co-ordinate at a time. For example, suppose $\mathbf{E}(X_1\,|\,\lambda_f(X)=\lambda)$ is positive for almost every $\lambda \in (\lambda_0, \lambda_1)$. We suggest why such an interval will always exist. We will show that $\lambda_f(x)$ is continuous at almost every x. The set $\{X\,|\,\lambda_f(X) = \lambda \in (\lambda_0, \lambda_1)\}$ is the set of X which exist in an open connected set in the normal plane at λ, and these normal planes vary smoothly as we move along the curve. Since the density of X_1 is smooth, it does not change much as we move from one normal plane to the next, and thus its expectation does not change much either. We then pick a differentiable g_1 so that it is also positive in that interval, and zero elsewhere, and set $g_2 \equiv \cdots \equiv g_p \equiv 0$. We apply the theorem and get $\mathbf{E}(X_1\,|\,\lambda_f(X)=\lambda) = f_1(\lambda)$ for $\lambda \in (\lambda_0, \lambda_1)$. We can do this for all such intervals, and for each co-ordinate, and thus the result is true. ∎

Corollary

If a principal curve is a straight line, then it is a principal component.

Proof

If f is a principal curve, then theorem 4.3 is true for all g, in particular for $g(\lambda) = \lambda v$. We then invoke theorem 4.2. ∎

In order to complete the proof, we need to prove the following

Lemma 4.3.1

The projection function $\lambda_{f_\epsilon}(x)$ is continuous at $\epsilon = 0$ for almost every x in the support of h.

Proof

Let us consider first where it will not be continuous. Suppose there are two points on f equidistant from x, and no other points on f are as close to x. Thus $\exists\ \lambda_0 > \lambda_1$, $\lambda_f(x) = \lambda_0$ and $\|x - f(\lambda_0)\| = \|x - f(\lambda_1)\|$. It is easy to pick g in this situation such that $\lambda_{f_\epsilon}(x)$ is not continuous at $\epsilon = 0$. We call such points ambiguous. However, we prove in lemma 4.3.2 that the set of all ambiguity points for a finite length differentiable curve has measure zero. We thus exclude them.

Suppose $\omega > 0$ is given, and there is no point on the curve as close to x as $f(\lambda_f(x)) = f(\lambda_0)$. Thus $\|x - f(\lambda_0)\| < \|x - f(\lambda_1)\|\ \forall\ \lambda_1 \in [a,b] \cap (\lambda_0 - \omega, \lambda_0 + \omega)^c$. (Notice that at the boundaries the ω interval can be suitably redefined.) Since this interval is compact, and the distance functions are differentiable, we can find a $\delta > 0$ such that $\|x - f(\lambda_0)\| \leq \|x - f(\lambda_1)\| - \delta$. Let $M = \sup_{\lambda \in [a,b]} \|g(\lambda)\|$ and $\epsilon_0 = \delta/(2M)$. Then $\|x - f_\epsilon(\lambda_0)\| < \|x - f_\epsilon(\lambda_1)\|\ \forall\ \lambda_1 \in [a,b] \cap (\lambda_0 - \omega, \lambda_0 + \omega)^c$ and $\forall \epsilon \leq \epsilon_0$. This implies that $\lambda_{f_\epsilon}(x) \in (\lambda_0 - \omega, \lambda_0 + \omega)$, and the continuity is established. ∎

Lemma 4.3.2

The set of ambiguity points has probability measure zero.

Proof

We prove the lemma for a curve in 2-space, but the proof generalizes to higher dimensions. Referring to figure 4.2, suppose a is an ambiguity point for the curve f at λ. We draw the circle with center a and tangent to f at λ. This means that f must be tangent to the circle somewhere else, say at $f(\lambda')$. If b on the normal at $f(\lambda)$ is also an ambiguity point, we can draw a similar circle for it. But this contradicts the fact that $f(\lambda)$ is the closest point to a,

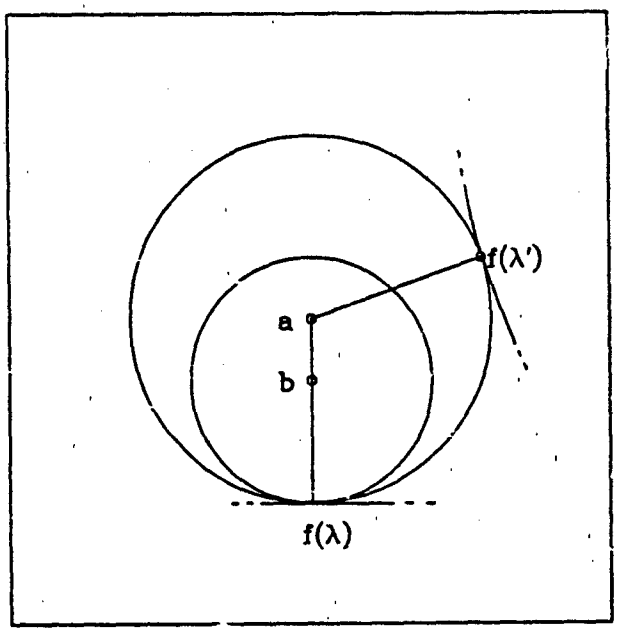


Figure 4.2 There are at most two ambiguity points on the normal to the curve; one on either side of the curve.

since the circle for b lies entirely inside the circle for a, and by the ambiguity of b we know the curve must touch this inner circle somewhere other than at $f(\lambda)$.

Let $I(X)$ be an indicator function for the set of ambiguity points. Since there are at most two at each λ, we have that $\mathbf{E}(I(X) \mid \lambda_f(X) = \lambda) = 0$. But this also implies that the unconditional expectation is zero. ∎

Corollary

The projection index $\lambda_f(x)$ is continuous at almost every x.

Proof

We show that if $\lambda_f(x)$ is not continuous at x, then x is an ambiguity point. But this set has measure zero by lemma 4.3.2.

If $\lambda_f(x)$ is not continuous at x, there exists a $\epsilon_0 > 0$ such that for every $\delta > 0$ $\exists$ z_δ such that $\|x - z_\delta\| < \delta$ but $\left|\lambda_f(x) - \lambda_f(z_\delta)\right| > \epsilon_0$. Letting δ go to zero, we see that x must

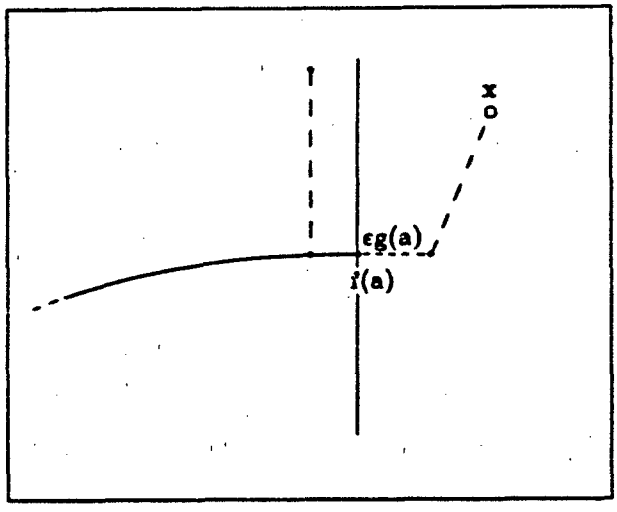


Figure 4.3 The set of points to the right of $f(a)$ that project there has measure zero.

be equidistant to $\lambda_f(x)$ and at least one other point on the curve with projection index at least ϵ_0 from $\lambda_f(x)$. ∎

Theorem 4.3 proves the equivalence of two statements: f is a principal curve and f is a critical value of the distance function. We needed to assume that f is defined on a compact set A. This means that the curve has two ends, and any data beyond the ends might well project at the endpoints. This leaves some doubt as to wether the endpoint can be the average of these points. The next lemma shows that for either statement of the theorem to be true, some truncation of the support of h might be necessary (if the support is unbounded).

Lemma 4.3.3

If f is a principal curve, then $\Big(x - f(\lambda_f(x))\Big) \cdot f'(\lambda_f(x)) = 0$ a.s. for x in the support of h. If $\frac{dD^2(h, f_\epsilon)}{d\epsilon}\Big|_{\epsilon=0} = 0 \ \forall$ differentiable g, then the same is true. By $f'(a)$ we mean the derivative from the right, and similarly from the left for $f'(b)$.

Proof

If $\lambda_f(x) \in (a, b)$ the proof is immediate. Suppose then that $\lambda_f(x) = a$. Rotate the coordinates so that $f'(a) = e_1$. No points to the left of $f(a)$ project there. Suppose f is a

principal curve. This then implies that the set of points that are to the right of $f(a)$ and project at $f(a)$ has conditional measure zero, else the conditional expectation would be to the right. Thus they also have unconditional measure zero.

Alternatively, suppose that there is a set of x of positive measure to the right of $f(a)$ that projects there. We can construct g such that $g(a) = f'(a)$, and zero everywhere else. For such a choice of g it is clear that the derivative cannot be zero. However, this choice of g is not continuous. But we can construct a version of g that is differentiable and does the same job as g. We have then reached a contradiction to the claim that $\frac{dD^2(h,f_\epsilon)}{d\epsilon}\Big|_{\epsilon=0} = 0 \ \forall$ differentiable g. ∎

4.3. Some results on the subclass of smooth principal curves.

We have defined a subset $\mathcal{F}_c(h)$ of principal curves. These are principal curves for which $\lambda_f(x)$ is a continuous function at each x in the support of h. In the previous section we showed that if $\lambda_f(x)$ is not continuous at x, then x is an ambiguity point. We now prove the converse: no points of continuity are ambiguity points. This will prove that the continuity constraint indeed avoids ambiguities in projection.

In figure 4.4a the curve is smooth but it wraps around so that points close together might project to completely different parts of the curve. This reflects a global property of the curve and presents an ambiguity that is unsatisfactory in a summary of a distribution.

Theorem 4.4

If $\lambda_f(x)$ is continuous at x, then x is not an ambiguity point.

Proof

We prove by contradiction. Suppose we have an x, and $\lambda_1 \neq \lambda_2$ such that

$$\begin{aligned} \|x - f(\lambda_1)\| &= \|x - f(\lambda_2)\| \\ &= d(x, f) \end{aligned}$$

It is easy to see that if λ_1 yields the closest point on the curve for x, then λ_1 is the position that yields the minimum for all $x_{\alpha_1} = \alpha_1 f(\lambda_1) + (1 - \alpha_1)x$ for $\alpha \in (0, 1)$. Similarly for λ_2. Now the idea is to let α_1 and α_2 get arbitrarily small, and thus $\|x_{\alpha_1} - x_{\alpha_2}\|$ gets small, but $\lambda_f(x_{\alpha_1}) - \lambda_f(x_{\alpha_2}) =$ *constant* and this violates the continuity of $\lambda_f(\cdot)$ ∎

Figure 4.4b represents the other ambiguous situation, this time caused by a local property of the curve. We consider only points *inside* the curve. If such points can occur at

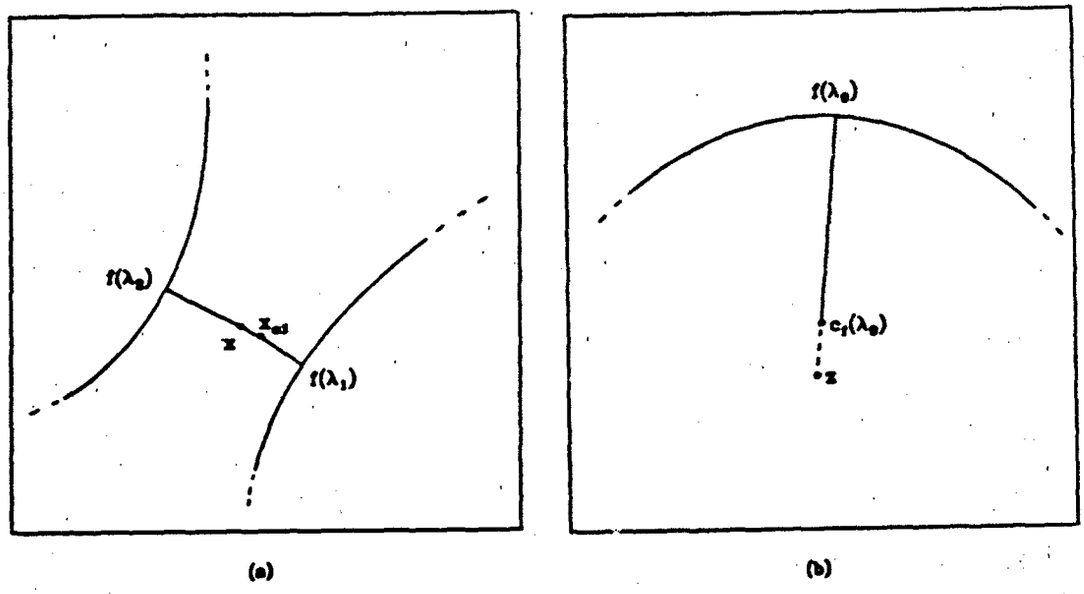


Figure 4.4 The continuity constraint avoids global ambiguities (a) and local ambiguities (b) in projection.

the center of curvature, then there is no unique point of projection on the curve. By *inside* we mean that the inner product $(x - f(\lambda_f(x))) \cdot (e_f(\lambda_f(x)) - f(\lambda_f(x)))$ is non-negative, where $e_f(\lambda)$ is the center of curvature of f at the point $f(\lambda)$.

Theorem 4.5

If $\lambda_f(x)$ is continuous at x, then x is not at the center of curvature of f at λ.

Proof

The idea of the proof is illustrated in figure 4.4b. If a point at $e_f(\lambda)$ projects at λ, then it will project at many other points immediately around λ, since locally $f(\lambda)$ behaves like the arc of a circle with center $e_f(\lambda)$. This would contradict the continuity of λ_f. Furthermore, if a point at x beyond $e_f(\lambda)$ projects at λ, we would expect that points on either side of x would project to different parts of the curve, and this would also contradict the continuity of λ_f.

We now make these ideas precise. Assume x projects at $\lambda_f(x) = \lambda_0$, where

$$x = f(\lambda_0) + \frac{f''(\lambda_0)}{\|f''(\lambda_0)\|}\left(\frac{1}{\|f''(\lambda_0)\|} + \delta\right)$$

and $\delta \geq 0$. Thus x is on or beyond the center of curvature of f at λ_0. Let $q(\lambda) \stackrel{\text{def}}{=} \|f(\lambda) - x\|$. By hypothesis $q(\lambda) \geq q(\lambda_0)$ with equality holding iff $\lambda = \lambda_0$. (Otherwise there would be at least two points on the curve the same distance from x and this would violate the continuity of λ_f). This implies that

(1) $q'(\lambda_0) = 0$

(2) $q''(\lambda_0) > 0$ for a strict minimum to be achieved.

We evaluate these two conditions:

$$\begin{aligned} q'(\lambda_0) &= f'(\lambda_0) \cdot (f(\lambda_0) - x) \\ &= f'(\lambda_0) \cdot -\frac{f''(\lambda_0)}{\|f''(\lambda_0)\|}\left(\frac{1}{\|f''(\lambda_0)\|} + \delta\right) \\ &= 0 \end{aligned}$$

$$\begin{aligned} q''(\lambda_0) &= f''(\lambda_0) \cdot (f(\lambda_0) - x) + f'(\lambda_0) \cdot f'(\lambda_0) \\ &= f''(\lambda_0) \cdot -\frac{f''(\lambda_0)}{\|f''(\lambda_0)\|}\left(\frac{1}{\|f''(\lambda_0)\|} + \delta\right) + 1 \\ &= -\|f''(\lambda_0)\|\,\delta \\ &\leq 0 \end{aligned}$$

which contradicts (2) above. ∎

4.4. Some results on bias.

The principal curve procedure is inherently biased. There are two forms of bias that can occur concurrently. We identify them as *model bias* and *estimation bias*.

Model bias occurs in the framework of a functional model, where the data is generated from a model of the form $x = f(\lambda) + e$, and we wish to recover $f(\lambda)$. In general, starting at $f(\lambda)$, the principal curve procedure will not have $f(\lambda)$ as its solution curve, but rather a biased version thereof. This bias goes to zero with the ratio of the noise variance tothe radius of curvature.

Estimation bias occurs because we use scatterplot smoothers to estimate conditional expectations. The bias is introduced because we average over neighborhoods, and this usually has a flattening effect.

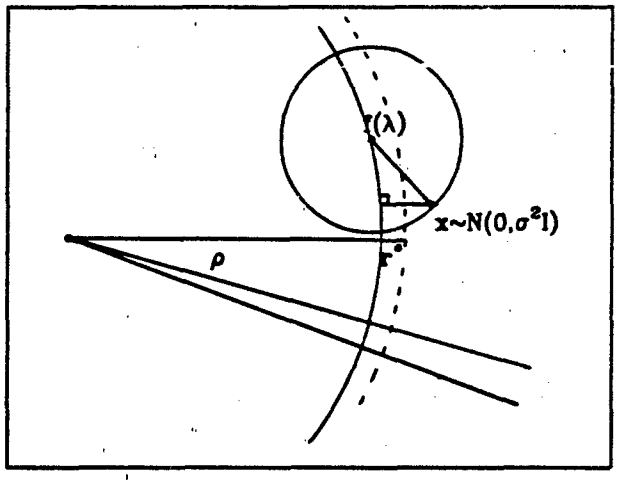


Figure 4.5 The data is generated from the arc of a circle with radius ρ and with iid $N(0, \sigma^2 I)$ errors. The location on the circle is selected uniformly.

4.4.1. A simple model for investigating bias.

The scenario we shall consider is the arc of a circle in 2-space. This can be parametrized by a unit speed curve $f(\lambda)$ with constant curvature $1/\rho$, where ρ is the radius of the circle:

$$f(\lambda) = \begin{pmatrix} \rho\cos(\lambda/\rho) \\ \rho\sin(\lambda/\rho) \end{pmatrix}, \tag{4.6}$$

for $\lambda \in [-\lambda_f, \lambda_f] \subseteq [-\pi\rho, \pi\rho]$. For the remainder of this section we will denote intervals of the type $[-\lambda_\theta, \lambda_\theta]$ by A_θ.

The points x are generated as follows: First a λ is selected uniformly from A_f. Given this value of λ we pick the point x from some smooth symmetric distribution with first two moments $(f(\lambda), \sigma^2 I)$ where σ has yet to be specified. Intuitively it seems that more mass gets put *outside* the circle than inside, and so the circle, or arc thereof, that gets closest to the data has radius larger than ρ. Consider the points that project onto a small arc of the circle (see figure 4.5). They lie in a segment which fans out from the origin. As we shrink this arc down to a point, the segment shrinks down to the normal to the curve at that point, but there is always more mass outside the circle than inside. So when we take conditional expectations, the mean lies *outside* the circle.

One would hope that the principal curve procedure, operating in distribution space

and starting at the true curve, would converge to this minimizing distance circle in this idealized situation. It turns out that this is indeed the case.

Figure 4.5 depicts the situation. We have in mind situations where the ratio σ/ρ is small enough to guarantee that $\mathbf{P}(|e| > \rho) \approx 0$. This effectively keeps the points local; they will not project to a region on the circle too far from where they were generated.

Theorem 4.6

Let $f(\lambda), \lambda \in \Lambda_f$ be the arc of a circle as described above. The parameter λ is distributed uniformly in the arc, and given λ, $x = f(\lambda) + e$ where the components of e are iid with mean 0 variance σ^2. We concentrate on a smaller arc Λ_θ inside Λ_f, and assume that the ratio σ/ρ is small enough to guarantee that all the points that project into Λ_θ actually originated from somewhere within Λ_f.

Then

$$\mathbf{E}(x \mid \lambda_f(x) \in \Lambda_\theta) = \begin{pmatrix} r_\theta \\ 0 \end{pmatrix}$$

where

$$r_\theta = r^* \frac{\sin(\theta/2)}{\theta/2}, \tag{4.7}$$

$\lambda_\theta/\rho = \theta/2$ and

$$\begin{aligned} r^* &= \lim_{\theta \to 0} r_\theta \\ &= \mathbf{E}\sqrt{(\rho + e_1)^2 + e_2^2}. \end{aligned}$$

Finally $r^* \to \rho$ as $\sigma/\rho \to 0$.

Lemma 4.6.1

Suppose $\lambda_f = \pi\rho$. (We have a full circle.) The radius of the circle, with the same center as $f(\lambda)$, that minimizes the expected squared distance to the points is*

$$\begin{aligned} r^* &= \mathbf{E}\sqrt{(\rho + e_1)^2 + e_2^2} \\ &> \rho. \end{aligned}$$

Also $r^* \to \rho$ as $\sigma/\rho \to 0$.

---

* I thank Art Owen for suggesting this result.

Proof of lemma 4.6.1

The situation is depicted in Figure 4.5. For a given point x the squared distance from a circle with radius r is the radial distance and is given by

$$d^2(x,r) = (\|x\| - r)^2.$$

The expected drop in the squared distance using a circle with radius r instead of ρ is given by

$$\begin{aligned} \mathbf{E}\Delta D^2(x,r,\rho) &= \mathbf{E}d^2(x,\rho) - \mathbf{E}d^2(x,r) \\ &= \mathbf{E}(\|x\| - \rho)^2 - \mathbf{E}(\|x\| - r)^2 \end{aligned} \tag{4.8}$$

We now condition on $\lambda = 0$ and expand (4.8) to get

$$\mathbf{E}\Delta D^2(x,r,\rho \,|\lambda = 0) = \rho^2 - r^2 + 2(r-\rho)\,\mathbf{E}\sqrt{(\rho+e_1)^2 + e_2^2}$$

Differentiating w.r.t. r we see that a maximum is achieved for

$$\begin{aligned} r &= r^* \\ &= \mathbf{E}\sqrt{(\rho+e_1)^2 + e_2^2} \end{aligned}$$

$$\begin{aligned} r_* &= \rho\,\mathbf{E}\sqrt{(1+e_1/\rho)^2 + (e_2/\rho)^2} \\ &\geq \rho\,\mathbf{E}\,|1+e_1/\rho| \\ &\geq \rho|\,\mathbf{E}(1+e_1/\rho)| \quad \text{(Jensen)} \\ &= \rho \end{aligned}$$

with strict inequality iff $\mathbf{Var}(e_i/\rho) = \sigma^2/\rho^2 = 0$. Note that

$$\mathbf{E}\Delta D^2(x,r^*,\rho) = (\rho - \mathbf{E}\sqrt{(\rho+e_1)^2 + e_2^2})^2 \tag{4.9}$$

which is non-negative.

When we condition on some other value of λ, we can rotate the system around so that $\lambda = 0$ since the distance is invariant to such rotations, and thus for each value of λ the same r^* maximizes $\mathbf{E}\Delta D^2(x,r,\rho\,|\lambda)$, and thus maximizes $\mathbf{E}\Delta D^2(x,r,\rho)$. ∎

Note: We can write the expression for r^* as

$$r^* = \rho\,\mathbf{E}\sqrt{(1+\epsilon_1)^2 + \epsilon_2^2} \tag{4.10}$$

where $\epsilon_i = e_i/\rho$, $\epsilon_i \sim (0, \delta)$, and $\delta = \sigma/\rho$. Expanding the square root expression using the Taylor's expansion we get

$$r^* \approx \rho + \sigma^2/(2\rho). \tag{4.11}$$

This yields an expected squared distance of

$$\mathbf{E}d^2(X, r^*) \approx \sigma^2 - \sigma^4/(4\rho^2)$$

which is smaller than the usual σ^2. This expression was also obtained by Efron (1984).

Proof of theorem 4.6.

We will show that in a segment of size ϕ the expected distance from the points in the segment to their mean converges to the expected radial distance as $\phi \to 0$. If we consider all such segments of size ϕ, the conditional expectations will lie on the circumference of a circle. By definition the conditional expectations minimize the squared distances to the points in their segments, and hence in the limit the radial distance in each segment. But so did r^*, and the results follow.

Suppose that ϕ is chosen so that $2\pi/\phi$ is a positive integer. We divide the circle up into segments each with arc angle ϕ. Consider $\mathbf{E}(x \,|\, \lambda_f(x) \in \Lambda_\phi)$, where Λ_ϕ and λ_ϕ are defined above.

Figure 4.6depicts the situation. The points are symmetrical about the x_1-axis, so the expectation will be of the form $(r, 0)'$. By the rotational invariance of the problem, if we find these conditional expectations for each of the segments in the circle, we end up with a circle of points, spaced ϕ degrees apart with radius r.

We first show that as $\phi \to 0$, $r \to r*$. In order to do this, let us compare the distance of points from their mean vector $r = (r, 0)'$ in the segment, to their radial distance from the circle with radius r. If we let $r(x)$ denote the radial projection of x onto the circle, we have

$$\begin{aligned} \mathbf{E}[(x - \mathbf{E}(x \,|\, \lambda_f(x) \in \Lambda_\phi)^2 \,|\, \lambda_f(x) \in \Lambda_\phi] &= \mathbf{E}[(x - r)^2 \,|\, \lambda_f(x) \in \Lambda_\phi] \\ &\geq \mathbf{E}[(x - r(x))^2 \,|\, \lambda_f(x) \in \Lambda_\phi] \end{aligned} \tag{4.12}$$

Also, we have

$$\begin{aligned} &\mathbf{E}[(x - r)^2 \,|\, \lambda_f(x) \in \Lambda_\phi] \\ &\quad = \mathbf{E}[(x - r(x))^2 \,|\, \lambda_f(x) \in \Lambda_\phi] + \mathbf{E}[(r(x) - r)^2 \,|\, \lambda_f(x) \in \Lambda_\phi] \\ &\quad\quad - 2\,\mathbf{E}\left(|r(x) - r|\,|x - r(x)|\cos(\psi(x)) \,|\, \lambda_f(x) \in \Lambda_\phi\right) \end{aligned} \tag{4.13}$$

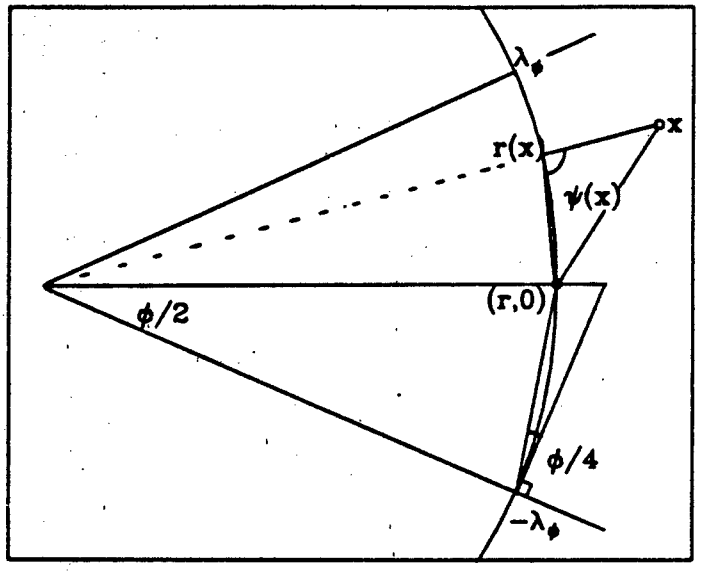


Figure 4.6 The conditional expectation of x, given $\lambda_f(x) \in \Lambda_\phi$.

where $\psi(x)$ are the angles as depicted in figure 4.6. The second term on the right of (4.13) is smaller than $(r\phi/2)^2$. We treat separately the case when x is inside the circle, and when x is outside.

- When x is inside the circle, $\psi(x)$ is acute and hence $\cos(\psi(x)) > 0$. Thus

$$\begin{aligned} &\mathbf{E}[(\mathbf{r}-r)^2 \,|\, \lambda_f(x) \in \Lambda_\phi] \\ &\qquad \leq \mathbf{E}[(x - r(x))^2 \,|\, \lambda_f(x) \in \Lambda_\phi] + O(\phi) \end{aligned} \tag{4.14}$$

- When x is outside the circle, $\psi(x)$ is obtuse and $\cos(\psi(x)) < 0$. Since $-\cos(\psi(x)) = \sin(\psi(x) - \pi/2)$ and from the figure $\psi(x) - \pi/2 \leq \phi/4$, we have that $-\cos(\psi(x)) \leq \sin(\phi/4) = O(\phi)$. Now $\mathbf{E}[(|r(x) - r| \cdot |x - r(x)|) \,|\, \lambda_f(x) \in \Lambda_\phi]$ is bounded since the errors are assumed to have finite second moments. Thus (4.14) once again holds.

So from (4.12) and (4.14), as $\phi \to 0$, the expected squared radial distance in the segment and the expected squared distance to the mean vector converge to the same limit. Suppose

$$\begin{aligned} \mathbf{E}(x \,|\, \lambda_f(x) = 0) &= r^{**} \\ &= \begin{pmatrix} r^{**} \\ 0 \end{pmatrix} \end{aligned}$$

Since the conditional expectation r^{**} minimizes the expected squared distance in the segment, this tells us that a circle with radius r^{**} minimizes the radial distance in the segment. Since, by rotational symmetry, this is true for each such segment, we have that r^{**} minimizes

$$\mathbf{E}_\phi \, \mathbf{E}(\|x\| - r)^2 \,|\, \lambda_f(x) = \phi) = \mathbf{E}(\|x\| - r)^2.$$

This then implies that $r^{**} = r^*$ by lemma 4.6.1 and thus

$$\begin{aligned} \lim_{\phi \to 0} \mathbf{E}(x \,|\, \lambda_f(x) \in \Lambda_\phi) &= \mathbf{E}(x \,|\, \lambda_f(x) = 0) \\ &= r^* \end{aligned}$$

This is the conditional expectation of points that project to a an arc of size 0 or simply a point. In order to get the conditional expectation of points that project onto an arc of size θ, we simply integrate over the arc:

$$\mathbf{E}(x \,|\, \lambda_f(x) \in \Lambda_\theta) = \mathbf{E}_{\lambda_f(x) \in \Lambda_\theta} \, \mathbf{E}(x \,|\, \lambda_f(x) = \lambda)$$

Suppose λ corresponds to an angle x, then

$$\mathbf{E}(x \,|\, \lambda_f(x) = \lambda) = \begin{pmatrix} r^* \cos(x) \\ r^* \sin(x) \end{pmatrix}$$

Thus

$$\begin{aligned} \mathbf{E}(x \,|\, \lambda_f(x) \in \Lambda_\theta) &= \begin{pmatrix} \int_{-\theta/2}^{\theta/2} \frac{r^* \cos(x)}{\theta} \, dx \\ \int_{-\theta/2}^{\theta/2} \frac{r^* \sin(x)}{\theta} \, dx \end{pmatrix} \\ &= \begin{pmatrix} r^* \frac{\sin(\theta/2)}{\theta/2} \\ 0 \end{pmatrix} \end{aligned} \tag{4.15}$$

■

Corollary

The above results generalize exactly for the situation where data is generated from a sphere in $\mathbf{R}^3$. The sphere that gets closest to the data has radius

$$r^* = \mathbf{E}\sqrt{(\rho + e_1)^2 + e_2^2 + e_3^2}$$

and this is exactly the conditional expectation of x_1 for points whose projection is at $(\rho, 0, 0)'$.

Corollary

If the data is generated from the circumference of a circle as above, the principal curve procedure converges after one iteration if we start at the model. This is also true for the principal surface procedure if the data is generated from the surface of a sphere.

Proof

After one iteration, we have a circle with radius r^*. All the points project at exactly the same position, and so the conditional expectations are the same. This is also true for the principal surface procedure on the sphere. ∎

4.4.2. From the circle to the helix.

The circle gives us insight into the behaviour of the principal curve procedure, since we can imagine any smooth curve as being made up of many arcs of circles. Equation (4.15) clearly separates and demonstrates the two forms of bias:

- Model bias since $r^* \geq \rho$.
- Estimation bias since the co-ordinate functions are shrunk by a factor $\sin(\theta/2)/(\theta/2)$ when we average within arcs or spans of size θ.

For a sufficiently large span, the estimation bias will dominate. Suppose that in the present setup, $\sigma = \rho/4$. Then from (4.11) we have that $r^* = 1.031\rho$. From (4.7) we see that a smoother with span corresponding to 0.27π or 14% of the observations will cancel this effect. This is considered a small span for moderate sample sizes. Usually the estimation bias will tend to flatten out curvature. This is not always the case, as the circle example demonstrates. In this special setup, the center of curvature remains fixed and the result of flattening the co-ordinate functions is to reduce the radius of the circle. The central idea is still clear: model bias is in a direction away from the center of curvature, and estimation bias towards the center.

We can consider a circle to be a flattened helix. We show that as we unflatten the helix, the effect of estimation bias changes from reducing the radius of curvature to increasing it.

To fix ideas we consider again the circle in $\mathbb{R}^2$. As we have observed the result of estimation and model bias is to reduce the expected radius from 1 to r (for a non-zero span

smoother such that $r < 1$). Thus we have

$$\hat{f}_0 = \begin{pmatrix} r\cos(\lambda) \\ r\sin(\lambda) \end{pmatrix},$$

with $\left\|\hat{f}_0'(\lambda)\right\| \equiv r$. The reparameterized curve is given by

$$\hat{f} = \begin{pmatrix} r\cos(\lambda/r) \\ r\sin(\lambda/r) \end{pmatrix},$$

and by definition the radius of curvature is $r < 1$. Here the center of curvature remains the same, but this is not usually the case.

A unit speed helix in $\mathbb{R}^3$ can be represented by

$$f(\lambda) = \begin{pmatrix} \cos(\lambda/c) \\ \sin(\lambda/c) \\ b\lambda/c \end{pmatrix}$$

where $c^2 = 1 + b^2$. It is easy to check that $r_f = 1 + b^2$, so even though the helix looks like a circle with radius 1 when we look down the center, it has a radius of curvature larger than 1. This is because the *osculating plane*, or plane spanned by the normal vector and the velocity vector, makes an angle with the $x_1 - x_2$ plane. In the case of a circle, the effect of the smoothing was to shrink the co-ordinates by a factor r. For a certain span smoother, the helix co-ordinates will become $(r\cos(\lambda/c), r\sin(\lambda/c), b\lambda/c)'$. Notice that straight lines are preserved by the smoother. Thus the new unit speed curve is given by

$$\hat{f}(\lambda) = \begin{pmatrix} r\cos(\lambda/c^*) \\ r\sin(\lambda/c^*) \\ b\lambda/c^* \end{pmatrix},$$

where $c^* = r^2 + b^2$. The radius of curvature is now $(r^2 + b^2)/r$. If we look at the difference in the radii we get

$$\begin{aligned} r_{\hat{f}} - r_f &= \frac{r^2 + b^2}{r} - 1 + b^2 \\ &= \frac{(1-r)(b^2 - r)}{r} \\ &> 0 \text{ if } b^2 > r \end{aligned}$$

This satisfies our intuition. For small b the helix is almost like a circle and so we expect circular behaviour. When b gets large, the helix is stretched out and the smoothed version has a larger radius of curvature.

4.4.3. One more bias demonstration.

We conclude this section with one further example. So far we have discussed bias in a rather oversimplified situation of constant curvature.

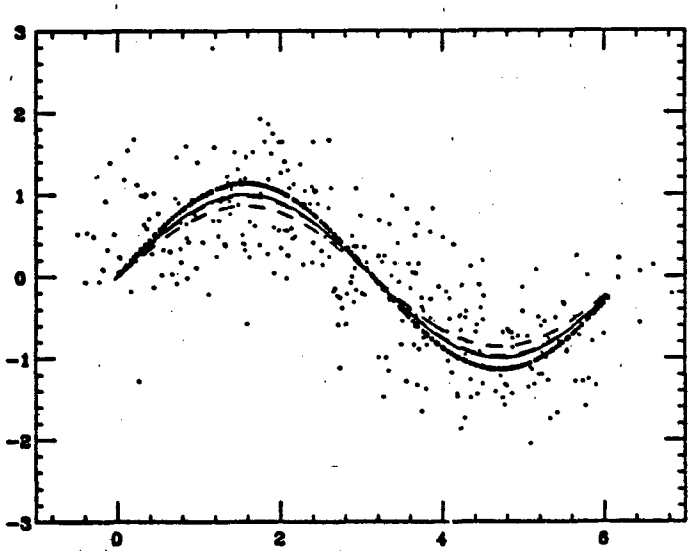


Figure 4.7 The thick curve is the the principal curve using conditional expectations at the model, and shows the *model bias*. The two dashed curves show the compounded effect of model and *estimation* bias at spans of 30% and 40%.

A sine wave in $\mathbf{R}^2$ does not have constant curvature. In parametric form we have

$$f(\lambda) = \begin{pmatrix} \lambda\pi \\ \sin(\lambda\pi) \end{pmatrix}.$$

A simple calculation shows that the radius of curvature $r_f(\lambda)$ is given by

$$\frac{1}{r_f(\lambda)} = \frac{\sin(\lambda\pi)}{(1+\cos^2(\lambda\pi))^{3/2}},$$

and achieves a minimum radius of 1 unit. The model for the data is $X = f(\lambda) + e$ where $\lambda \sim U[0,2]$ and $e \sim \mathcal{N}(0, I/4)$ independent of λ. Figure 4.7shows the true model (solid curve), and the points are a sample from the model, included to give an idea of the error structure. The thick curve is $\mathbf{E}(X \mid \lambda_f(X) = \lambda)$. Here is a situation where the model bias results in a curve with more curvature, namely a minimum radius of 0.88 units. This

curve was found by simulation, and is well approximated by $1/0.88\sin(\lambda\pi)$. There are two dashed curves in the figure. They represent $\mathbf{E}(X \mid \lambda_f(X) \in \Lambda_s(\lambda))$, where $\Lambda_s(\lambda)$ represents a symmetric interval of length $s\Lambda$ about λ (Boundary effects were eliminated by cyclically extending the range of λ.) We see that at $s = 30\%$ the estimation bias approximately cancels out the model bias, whereas at $s = 40\%$ there is a residual estimation bias.

4.5. Principal curves of elliptical distributions.

We have seen that for elliptical distributions the principal components are principal curves. Are there any more principal curves? We first of all consider the uniform disc with no *holes*. For this distribution we propose the following:

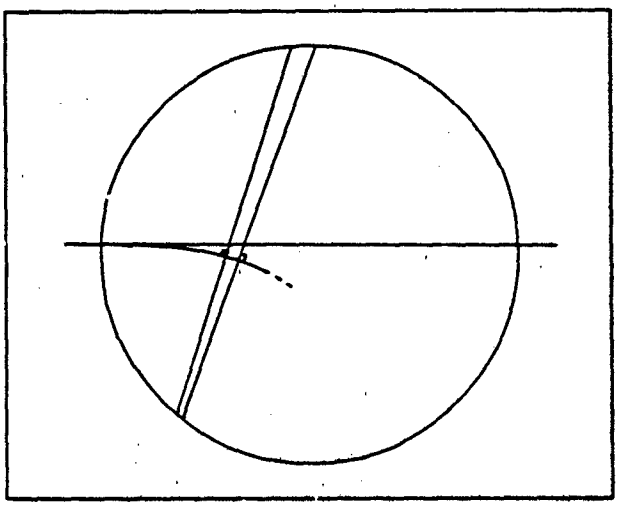


Figure (4.8) The only principal curves in $\mathcal{F}_c(h)$ of a uniform disk are the principal components.

Proposition

The only principal curves in $\mathcal{F}_c(h)$ are straight lines through the center of the disk.

An informal proof of this claim is as follows:

- Any principal curve must enter the disk once and leave it once. This must be true since if it were to remain inside it would have to circle around. But this would violate the continuity constraint imposed by $\mathcal{F}_c(h)$ since there would have to exist points at

the centers of curvature of the curve at some places. Furthermore, it cannot end inside the disk for reasons similar to those used in lemma 4.3.3.

- The curve enters and leaves the disk normal to the circumference. For symmetry reasons this must be true. As it enters the disk there must be equal mass on both sides.

- The curve never bends (see figure 4.8). At the first point of curvature, the normal to the curve will be longer on one side than the other. The set of points that project at this spot will not be conditionally uniformly distributed along the normal. This is because the set is the limit of a sequence of segments with center at the center of curvature of the curve at the point in question. Also, all points in the segment will project onto the arc that generates the segment; if not the continuity constraint would be violated. So in addition to the normal being longer, it will have more mass on the long side as well. This contradicts the fact that the mean lies on the curve.

Thus the only curves allowed are straight lines, and they will then have to pass through the center of the disk.

Suppose now that we have a convex combination of two disks of different radii but the same centers. A similar argument can be used to show that once again the only principal curves are the lines through the center. This then generalizes to any mixture of uniform disks and hence to any spherically symmetric distribution of this form.

We conjecture that for ellipsoidal distributions the only principal curves are the principal components.

Chapter 5

Algorithmic details

In this chapter we describe in more detail the various constituents of the principal curve and surface algorithms.

5.1. Estimation of curves and surfaces.

We described a simple smooth or local averaging procedure in chapter 4. There it was convenient to describe the smoother as a method of averaging in p space, although it has been pointed out that we can do the smoothing co-ordinate wise. That simplifies the treatment here, since we only need to discuss smoothers in their more usual regression context.

Usually a scatterplot smoother is regarded as an estimate of the conditional expectation $\mathbf{E}(Y \mid X)$, where Y and X are random variables. For our purposes X may be one or two dimensional. We will discuss one dimensional smoothers first, since they are easier to implement than two dimensional smoothers.

5.1.1. One dimensional smoothers.

The following subset of smoothers evolved naturally as estimates of conditional expectation, and are listed in order of complexity and computational cost.

5.1.1.1 Moving average smoothers.

The simplest and most natural estimate of $E(Y \mid X)$ is the *moving average* smoother. Given a sample (y_i, x_i), $i = 1, \ldots, n$, with the x_i in ascending order, we define

$$\text{Smooth}_s(y \mid x_i) = \frac{1}{2k+1} \sum_{x_j \in [x_{i-k}, x_{i+k}]} y_j \tag{5.1}$$

where $k = [(ns-1)/2]$ and $s \in (0,1]$ is called the span of the smoother. An estimate of the conditional expectation at x_i is the average of the y_j for all those observations with x value equal to x_i. Since we usually only have one such observation, we average the y_j for

all those obeervations with x value close to x_i. In the definition above, close is defined in the ordinal scale or in ranks. We can also use the interval scale or simply distance, but this is computationally more expensive. This moving average smoother suffers from a number of drawbacks. It does not produce very smooth fits and does not even reproduce straight lines unless the x_i are equispaced. It also suffers from bias effects on the boundaries.

5.1.1.2 Local linear smoothers.

An improvement on the moving average smoother is the *local linear* smoother of Friedman and Stuetzle (1981). Here the smoother estimates the conditional expectation at x_i by the fitted value from the least squares line fit of y on x using only those points for which $x_j \in (x_{i-k}, x_{i+k})$. This suffers less from boundary bias than the moving average and always reproduces straight lines exactly. The cost of computation for both of the above smoothers is $O(n)$ operations. Of course we can think of fitting local polynomials as well, but in practice the gain in bias is small relative to the extra computational burden.

5.1.1.3 Locally weighted linear smoothers.

Cleveland (1979) suggested using the local linear smoother, but also suggested weighting the points in the neighborhood according to their distance in x from x_i. This produces even smoother curves at the expense of an increased computation time of $O(kn)$ operations. (In the local linear smoother, we can obtain the fitted value at x_{i+1} from that at x_i by applying some simple updating algorithm to the latter. If local weighting is performed, we can no longer use updating formulae.)

5.1.1.4 Kernel smoothers.

The kernel smoother (Gasser and Muller, 1979) applies a weight function to every observation in calculating the fit at x_i. A variety of weight functions or kernels exist and a popular choice is the gaussian kernel centered at x_i. They produce the smoothest functions and are computationally the most expensive. The cost is $O(n^2)$ operations, although in practice the kernels have a bounded domain and this brings the cost down to $O(sn)$ for some s that depends on the kernel and the data.

In all but the kernel smoother, the span controls the smoothness of the estimated function. The larger the span, the smoother the function. In the case of the kernel smoother, there is a scale parameter that controls the spread of the kernel, and the larger the spread, the smoother the function. We will discuss the choice of spans in section 5.4.

For our particular application, it was found that the locally weighted linear smoother and the kernel smoother produced the most satisfactory results. However, when the sample size gets large, these smoothers become too expensive, and we have to sacrifice smoothness for computational speed. In this case we would use the faster local linear smoother.

5.1.2. Two dimensional smoothers.

There are substantial differences between one and two dimensional smoothers. When we find neighbors in two space, we immediately force some metric on the space in the way we define distance. In our algorithm we simply use the euclidean distance and assume the two variables are in the same scale.

It is also computationally harder to find neighbors in two dimensions than in one. The *k-d tree* (Friedman, Bently and Finkel, 1976) is an efficient algorithm and data structure for finding neighbors in k dimensions. The name arises from the data structure used to speed up the search time — a binary tree. The technique can be thought of as a multivariable version of the binary search routine. Friedman et al show that the computation required to build the tree is $O(kn \log n)$ and the expected search time for the m nearest neighbors of any point is $O(\log n)$.

5.1.3. The local planar surface smoother.

We wish to find $\textbf{Smooth}(y \mid x_0)$ where x_0 is a 2-vector not necessarily present in the sample. The following algorithm is analogous to the local linear smoother:

- Build the 2-d tree for the n pairs $(x_{11}, x_{21}), \cdots, (x_{1n}, x_{2n})$.
- Find the ns nearest neighbors of x_0, and fit the least squares plane through their associated y values.
- The smooth at x_0 is defined to be the fitted value at x_0.

This algorithm does not allow updating as in the one-dimensional local linear smoother. The computation time for one fitted value is $O(\log n + ns)$. For this reason, we can include weights at no extra order in computation cost. We use gaussian weights with covariance $h^2 I$ and centered at x_0, and h is another parameter of the procedure.

A simpler version of this smoother uses the (gaussian weighted) average of the y values for the ns neighbors. In the one dimensional case, we find that fitting local straight lines reduces the bias at the boundaries. In surface smoothing, the proportion of points on the

boundary increases dramatically as we go from one to two dimensions. This provides a strong motivation for fitting planes instead of simple averages.

5.2. The projection step.

The other step in the principal curve and surface procedures is to *project* each point onto the current surface or curve. In our notation we require $\hat{\lambda}^{(j)}(x_i)$ for each i. We have already described the exact approach in chapter 3 for principal curves, which we repeat here for completeness.

5.2.1. Projecting by exact enumeration.

We project x_i *into* the line segment joining every adjacent pair of fitted values of the curve, and find the closest such projection. Into implies that when projecting we do not go beyond the two points in question. This procedure is exact but computationally expensive ($O(n)$ operations per search.) Nonetheless, we have used this method on the smaller data sets (≤ 150 observations.) There is no analogue for the principal surface routine.

5.2.2. Projections using the k-d tree.

At each of the n values of $\hat{\lambda}$ we have a fitted p vector. This is true for either the principal curve or surface procedure. We can build a p-d tree, and for each x_i, find its nearest neighbor amongst these fitted values. We then proceed differently for curves and surfaces.

- For curves we project the point *into* the segments joining this nearest point and its left neighbor. We do the same for the right neighbor and pick the closest projection.
- For surfaces we find the nearest fitted value as above. Suppose this is at $\hat{f}^{(j)}(\hat{\lambda}_k^{(j-1)})$. We then project x_i onto the plane corresponding to this fitted value and get a new value λ^*. (This plane has already been calculated in the smoothing step and is stored.) We then evaluate $\hat{f}^{(j)}(\lambda^*)$ and check if it is indeed closer. (This precautionary step is similar to projecting x_i *into* the line segments in the case of curves.) If it is, we set $\hat{\lambda}_i^{(j)} = \lambda^*$, else we set $\hat{\lambda}_i^{(j)} = \hat{\lambda}_k^{(j-1)}$. One could think of iterating this procedure, which is similar to a gradient search. Alternatively one could perform a Newton-Raphson search using derivative information contained in the least squares planes. These approaches are expensive, and in the many examples tested, made little or no difference to the estimate.

5.2.3. Rescaling the λ's to arc-length.

In the principal curve procedure, as a matter of practice, we always rescale the λ's to arc-length. The estimated λ's are then measured in the same units as the observations. Let $\hat{\lambda}_i^r$ denotes the rescaled $\hat{\lambda}_i^{(j)}$'s, and suppose $\hat{\lambda}_i^{(j)}$ are sorted. We define $\hat{\lambda}_i^r$ recursively as follows:

- $\hat{\lambda}_1^r = 0.$
- $\hat{\lambda}_i^r = \hat{\lambda}_{i-1}^r + \left\| \hat{f}^{(j)}(\hat{\lambda}_i^{(j)}) - \hat{f}^{(j)}(\hat{\lambda}_{i-1}^{(j)}) \right\|.$

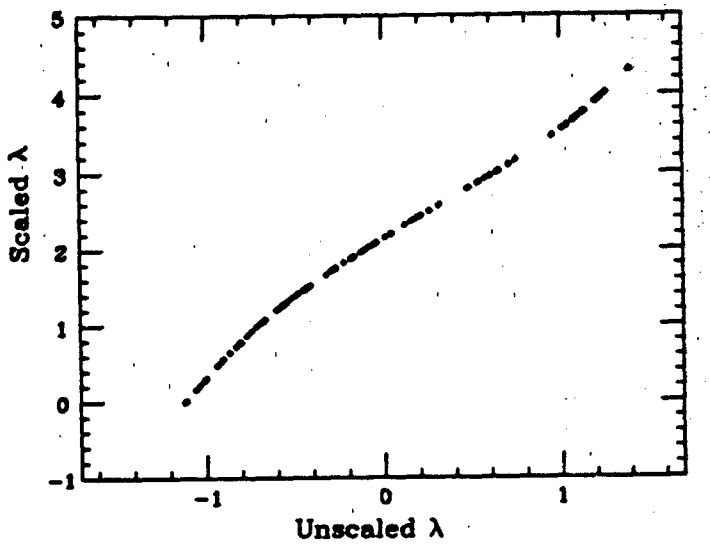


Figure (5.1) A λ plot for the circle example. Along the vertical axis we plot the final values for $\hat{\lambda}_i$, after rescaling the $\hat{\lambda}$'s at every iteration in the principal curve procedure. Along the horizontal axis we have the final $\hat{\lambda}$'s using the principal curve procedure with no rescaling.

In general there is no analogue of rescaling to arc-length for surfaces. Surface area is the corresponding quantity. We can adjust the parameters locally so that the area of a small region in parameter space has the same area as the region it defines on the surface. But this adjustment will be different in other regions of the surface having the same values for one of the parameters. The exceptions are surfaces with zero gaussian curvature. (These are surfaces that can be obtained by *smoothly denting* a hyperplane to form something like a corrugated sheet. One can imagine that such a rescaling is then possible).

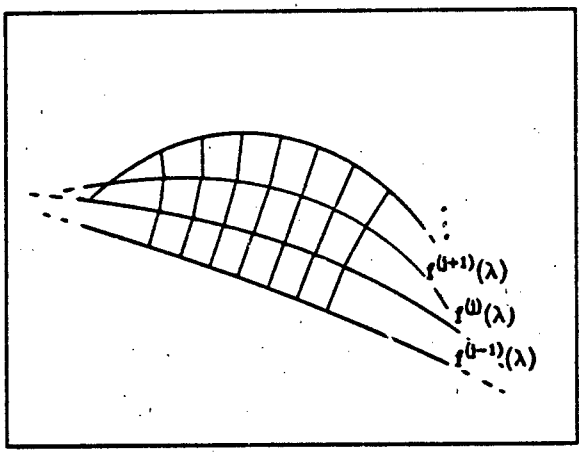


Figure (5.2) Each iteration approximately preserves the metric from the previous one. The starting curve is unit speed, and so the final curve is approximately so, up to a constant.

Even though it is not possible to do such a rescaling for surfaces, it would be comforting to know that our parametrization remains reasonably consistent over the surface as we go through the iterations.

Figure 5.1 demonstrates what happens if we use the principal curve procedure on the circle example, and do not rescale the parameter estimates at each iteration. The metric gets preserved, up to a scalar. Figure 5.2shows why this is so. The original metric gets transferred from one iteration to the next. As long as the curves do not change dramatically from one iteration to the next, there will not be much distortion.

5.3. Span selection.

We consider there to be two categories of spans corresponding to two distinct stages in the algorithm.

5.3.1. Global procedural spans.

The first guess for f is a straight line. In many of the interesting situations, the final curve will not be a function of the arc length of this initial curve. The final curve is reached by successively *bending* the original curve. We have found that if the initial spans of the smoother are too small, the curve will bend too fast, and may get lost! The most

successful strategy has been to initially use large spans, and then to decrease them slowly. In particular, we start with a span of $0.5n$, and let the procedure converge. We then drop the span to $0.4n$ and converge again. Finally the same is done at $0.3n$ by which time the procedure has found the general shape of the curve. We then switch to mean square error (MSE) span selection mode.

5.3.2. Mean squared error spans.

The procedure has converged to a self consistent curve for the span last used. If we reduce the span, the average distance will decrease. This situation arises in regression as well. In regression, however, there is a remedy. We can use cross-validation (Stone 1977) to select the span. We briefly outline the idea.

5.3.2.1 Cross-validation in regression.

Suppose we have a sample of n independent pairs (y_i, x_i) from the model $Y = f(X) + \epsilon$. A nonparametric estimate of $f(x_0)$ is $\hat{f}_s(x_0) = \text{Smooth}_s(y \mid x_0)$. The expected squared prediction error is

$$EPE = \mathbf{E}(Y - \hat{f}(X))^2 \tag{5.2}$$

where the expectation is taken over everything random (i.e. the sample used to estimate $\hat{f}(\cdot)$ and the future pairs (X,Y)). We use the residual sum of squares,

$$RSS(s) = \sum_{i=1}^{n} (y_i - \hat{f}_s(x_i))^2,$$

as the natural estimate of EPE. This is however, a biassed estimate, as can be seen by letting the span s shrink down to 0. The smooth then estimates y_i by itself, and RSS is zero. We call this *bias due to overfitting* since the bias is due to the influence y_i has in forming its own prediction. This also shows us that we cannot use RSS to help us pick the span. We can, however, use the cross-validated residual sum of squares (CVRSS). This is defined as

$$CV\,RSS(s) = \sum_{i=1}^{n} (y_i - \text{Smooth}_s^{(i)}(y \mid x_i))^2, \tag{5.3}$$

where $\text{Smooth}_s^{(i)}(y \mid x_i)$ is the smooth calculated from the data with the pair (y_i, x_i) removed, and then evaluated at x_i. It can be shown that this estimate is approximately unbiassed for the true prediction error. In minimizing the prediction error, we also mini-

mize the integrated mean square error EMSE given by

$$EMSE(s) = \mathbf{E}(\hat{f}_s(X) - f(X))^2$$

since they differ by a constant. We can decompose this expression into a sum of a variance and bias terms, namely

$$\begin{aligned} EMSE(s) &= \mathbf{E}[\ \mathrm{Var}(\hat{f}_s(X)] + \mathbf{E}[(\ \mathbf{E}(\hat{f}_s(X)\,|X) - f(X))^2] \\ &= VAR(s) + BIAS^2(s). \end{aligned}$$

As s gets smaller the variance gets larger (averaging over less points) but the bias gets smaller (width of the neighborhoods gets smaller), and vice versa. Thus if we pick s to minimize CVRSS(s) we are trying to minimize the true prediction error or equivalently to find the span which optimally mixes bias and variance.

Getting back to the curves, one thought is to cross-validate the orthogonal distance function. This, however, will not work because we would still tend to use span zero. (In general we have more chance of being close to the interpolating curve than any other curve). Instead, we cross-validate the co-ordinates separately.

5.3.2.2 Cross-validation for principal curves.

Suppose f is a principal curve of h, for which we have an estimate $\hat{f}$ based on a sample $x_1, \ldots, x_n$.

A natural requirement is to choose s to minimize $EMSE(s)$ given by

$$\begin{aligned} EMSE(s) &= \mathbf{E}_h \left\| f(\lambda_f(X)) - \hat{f}_s(\lambda_f(X)) \right\|^2 \\ &= \sum_{j=1}^{p} \mathbf{E}_{h_\lambda}(\ \mathrm{Var}(\hat{f}_s(\lambda_f(X)) \,|\, \lambda_f(x)) + \mathbf{E}_{h_\lambda} \left\| f(\lambda_f(X)) - \hat{f}_s(\lambda_f(X)) \right\|^2 \quad (5.4) \end{aligned}$$

which is once again a trade-off between bias and variance. Notice that were we to look at the closest distance between these curves, then the interpolating curve would be favored. As in the regression case, the quantity $EPE(s) = \mathbf{E}_h \left\| X - \hat{f}_s(\lambda_f(X)) \right\|^2$ estimates $EMSE(s) + D(f)$, where $D(f) = \mathbf{E} \left\| X - f(\lambda_f(X)) \right\|^2$. It is thus equivalent to choose s to minimize $EMSE(s)$ or $EPE(s)$. As in the regression case, the cross-validated estimate

$$CVRSS(s) = \sum_{j=1}^{p} \left[\sum_{i=1}^{n} (x_{ji} - \mathrm{Smooth}_s^{(i)}(x_j \,|\, \lambda_i))^2 \right], \quad (5.5)$$

where $\lambda_i = \lambda_f(x_i)$, attempts to do this. Since we do not know λ_i, we pick $\lambda_i = \lambda_{\hat{f}^{(k)}}(x_i)$ where $\hat{f}^{(k)}$ is the (non cross-validated) estimate of f. In practice, we evaluate $CVRSS(s)$ for a few values of s and pick the one that gives the minimum.

From the computing angle, if the smoother is linear one can easily find the cross-validated fits. In this case $\hat{y} = Cy$ for some smoother matrix C, and the cross-validated fit $\hat{y}_{(i)}$ is given by $\hat{y}_{(i)} = \sum_{j \neq i} \frac{c_{ij} y_j}{1 - c_{ii}}$ (Wahba 1975).

There are a number of issues connected with the algorithms that have not yet been mentioned, such as a robustness and outlier detection, what to display and how to do it, and bootstrap techniques. The next chapter consists of many examples, and we will deal with these issues as they arise.

Chapter 6

Examples

This chapter contains six examples that demonstrate the procedures on real and simulated data. We also introduce some ideas such as bootstrapping, robustness, and outlier detection.

Example 6.1. Gold assay pairs.

This real data example illustrates:

- A principal curve in 2-space,
- non-linear errors in variables regression,
- co-ordinate function plots, and
- bootstrapping principal curves.

A California based company collects computer chip waste in order to sell it for its content of gold and other precious metals. Before bidding for a particular cargo, the company takes a sample in order to estimate the gold content of the the whole lot. The sample is split in two. One sub-sample is assayed by an outside laboratory, the other by their own inhouse laboratory. (The names of the company and laboratory are withheld by request). The company wishes to eventually use only one of the assays. It is in their interest to know which laboratory produces on average lower gold content assays for a given sample.

The data in figure 6.1a consists of 250 pairs of gold assays. Each point is represented by

$$\boldsymbol{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}$$

where x_{ji} = log(1 + assay yield for ith assay pair for lab j) and where $j = 1$ corresponds to the inhouse lab and $j = 2$ the outside lab. The log transformation tends to stabilize the variance and produce a more even scatter of points than in the untransformed data. (There were many more small assays (1 oz per ton) than larger ones (> 10 oz per ton)).

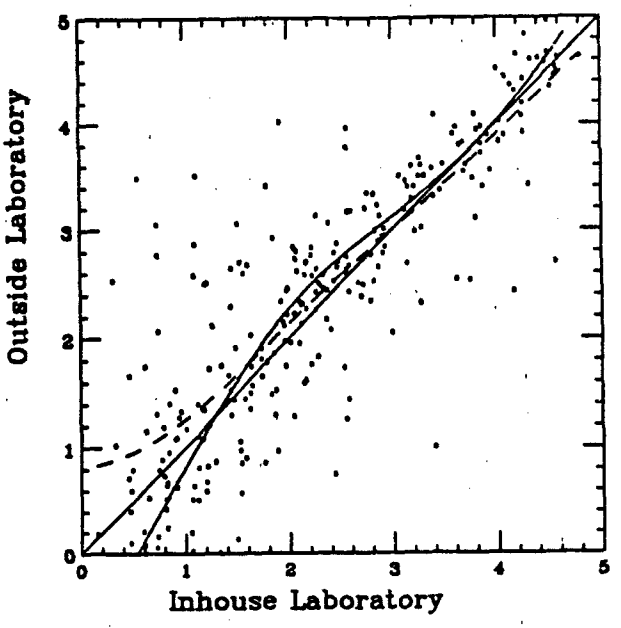


Figure 6.1a Plot of the log assays for the inhouse and outside labs. The solid curve is the principal curve, the dashed curve the scatterplot smooth.

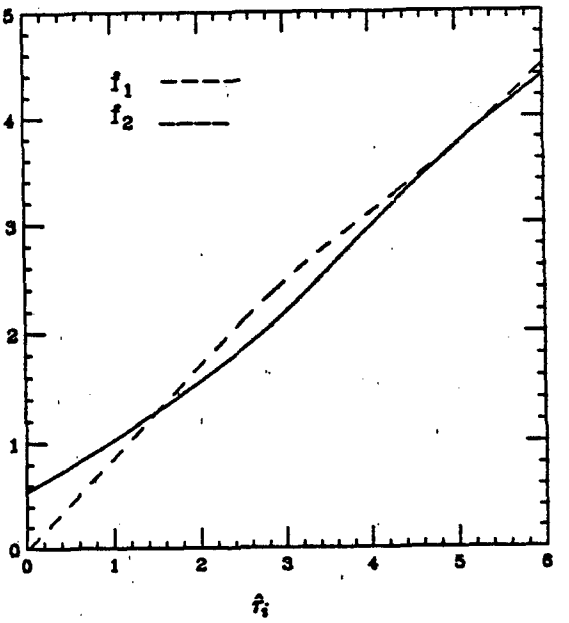


Figure 6.1b Estimated coordinate functions. The dashed curve is the outside lab, the solid curve the inhouse lab.

A standard analysis might be a paired t-test for an overall difference in assays. This would not reflect local differences which can be of great importance since the higher the level of gold the more important the difference.

The data was actually analyzed by smoothing the differences in log assays against the average of the two assays. This can be considered a form of symmetric smoothing and was suggested by Cleveland (1983). We discuss the method further in chapter 7.

The model presented here for the above data is

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} f_1(r_i) \\ f_2(r_i) \end{pmatrix} + \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \tag{6.1}$$

where r_i is the unknown true gold content for sample i (or any monotone function thereof), $f_j(r_i)$ is the expected assay result for lab j, and e_{ji} is measurement error. We wish to analyze the relationship between f_1 and f_2 for different true gold contents.

This is a generalization of the errors in variables model or the structural model (if we

regard the r_i themselves as unobservable random variables), or the functional model (if the r_i are considered fixed). This model is traditionally expressed as a linear model:

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} \alpha + \beta z_i \\ z_i \end{pmatrix} + \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \tag{6.2}$$

where $f_2(r_i) = z_i$ and

$$\begin{aligned} f_1(r_i) &= f_1 \circ f_2^{-1}(z_i) \qquad \text{(assuming } f_2 \text{ is monotone)} \\ &= \alpha + \beta z_i \end{aligned}$$

It suffers, however, from the same drawback as the t-test in that only global inference is possible.

We assume that the e_{ji} are pairwise independent and that *

$$\mathbf{Var}(e_{1i}) = \mathbf{Var}(e_{2i}) \quad \forall\, i.$$

The model is estimated using the principal curve estimate for the data and is represented by the solid curve in figure 6.1a. The dashed curve is the usual scatterplot smooth of x_2 against x_1 and is clearly misleading as a scatterplot summary. The curve lies above the 45° line in the interval 1.4 ɔ 4 which represents an untransformed assay interval of 3 to 15 oz/ton. In this interval the inhouse average assay is lower than that of the outside lab. The difference is reversed at lower levels, but this is of less practical importance since at these levels the cargo is less valuable. This is more clearly seen by examining the estimated coordinate function plots in figure 6.1b.

A natural question arising at this point is wether the kink in the curve is real or not. If we had access to more data from the same population we could simply calculate the principal curves for each and see how often the kink is reproduced. We could then perhaps construct a 95% confidence tube for the true curve.

In the absence of such repeated samples, we use the bootstrap (Efron 1981, 1982) to simulate them. We would like to, but cannot, generate samples of size n from F, the true distribution of x. Instead we generate samples of size n from $\hat{F}$, the empirical or estimated distribution function, which puts mass $1/n$ on each of the sample points x_i. Each such sample, which samples the points x_i with replacement, is called a bootstrap sample.

---

* In the linear model one usually requires that $Var(e_{ji}) = constant_j$. This assumption can be relaxed here.

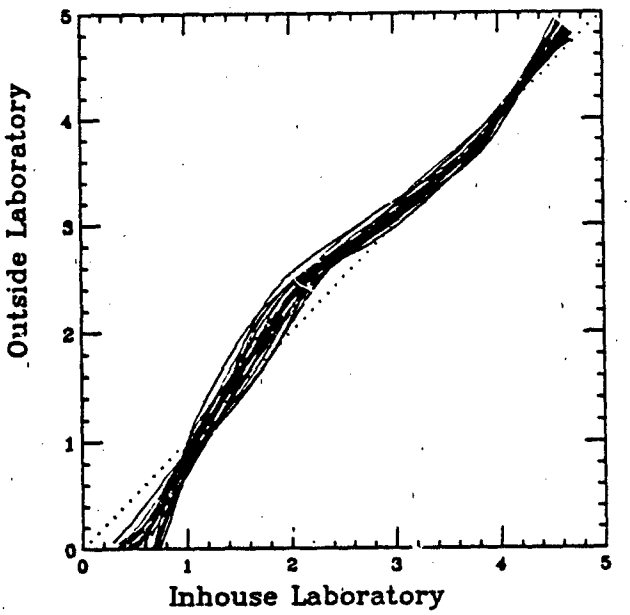


Figure 6.1*c* 25 bootstrap curves. The data X is sampled 25 times with replacement, each time yielding a bootstrap sample X^*. Each curve is the principal curve of such a sample.

Figure 6.1c shows the principal curves obtained for 25 such bootstrap samples. The 45° line is included in the figure, and we see that none of the curves cross the line in the region of interest. This provides strong evidence that the kink is indeed real.

When we compute a particular bootstrap curve, we use the principal curve of the original sample as a starting value. Usually one or two iterations are all that is required for the procedure to converge. Also, since each of the bootstrap points occurs at one of the sample sites, we know where they project onto this initial curve.

It is tempting to extract from the procedure estimates of $\hat{r}_i$, the true gold level for sample i. However, $\hat{r}_i$ need not be the true gold level at all. It may be any variable that orders the pairs $f(\hat{r}_i)$ along the curve, and is probably some monotone function of the true gold level. It is clear that both labs could consistently produce biased estimates of the true gold level and there is thus no information at all in the data about the true level.

Estimates of r_i do provide us with a good summary variable for each of the pairs, if

that is required:

$$\hat{r}_i = h(x_i)$$

since we obtain $\hat{r}_i$ by projecting the point x_i onto the curve. Finally we observe that the above analysis could be extended in a straightforward way to include 3 or more laboratories. It is hard to imagine how to tackle the problem using standard regression techniques.

Example 6.2. The helix in three-space.

This is a simulated example illustrating:

- A principal curve in 3-space,
- co-ordinate plots, and
- cross-validation and span selection.

We looked at the bias of the principal curve procedure in estimating the helix in chapter 4. We now demonstrate the procedure by generating data from that model. We have

$$f(\lambda) = \begin{pmatrix} \sin(4\pi\lambda) \\ \cos(4\pi\lambda) \\ 4\lambda \end{pmatrix} + e,$$

where $\lambda \sim U[0,1]$ and $e \sim N(0, .2I)$. This situation does not present the principal curve procedure with any real problems. The reason is that the starting vector passes down the middle of the helix and the data projects onto it in nearly the correct order. Table 6.1shows the steps in the iterations as the procedure converges at each of the *procedural spans* shown. At a span of $s = .2$ we use cross-validation to find the minimum *mse span*.

Figure 6.2c shows the $CVRSS$ curve used to select the span, which is 0.1 with a value of $CVRSS$ of 0.1944. One more step is performed and the procedure is terminated. Figure 6.2d shows the estimated co-ordinate functions for this choice of span. We see that the estimate of the linear co-ordinate is rather wiggly. It is clear that a small span was required to estimate the sinusoidal co-ordinates, but a large span would suffice for the linear co-ordinate. This suggests a different scheme for cross-validation—choosing the spans separately for each co-ordinate. The results are shown in figures 6.2e and 6.2f. As predicted, a larger span is chosen for the linear co-ordinate, and its estimate is no longer wiggly. This is the final model referred to in the table and represented in figure 6.2.

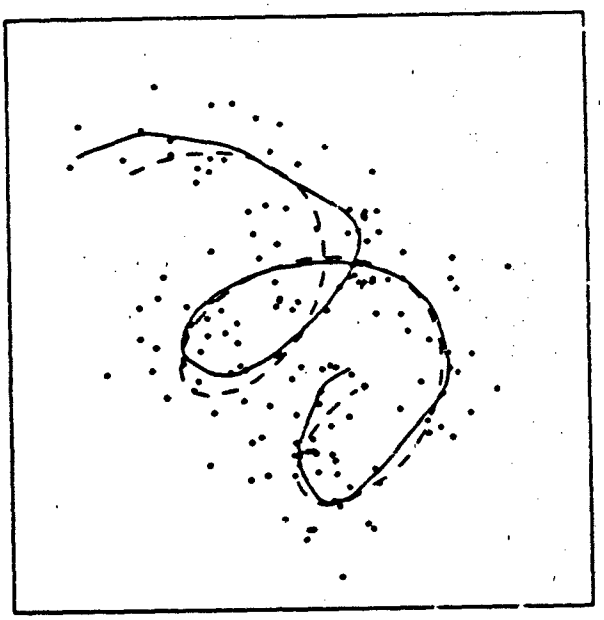


Figure 6.2a Data generated from a helix with independent errors on each coordinate. The dashed curve is the original helix, the solid curve is the principal curve estimate.

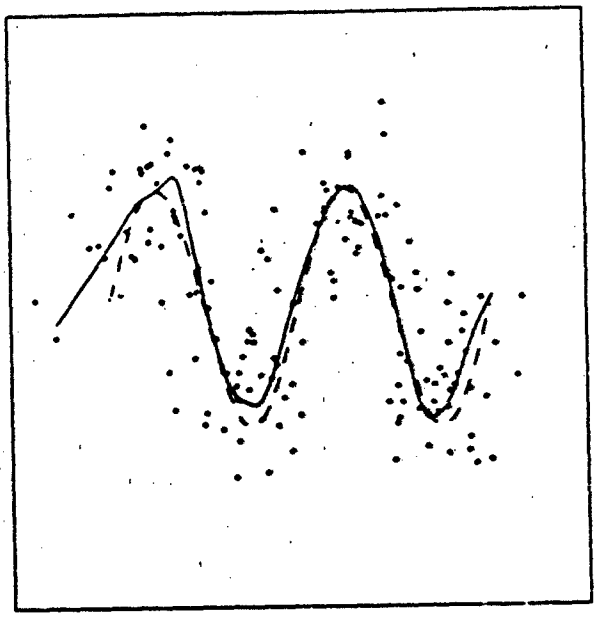


Figure 6.2b Another view of the helix, the data and the principal curve.

Table 6.1. The steps in the iterations. Initially the procedure goes through a regimen of procedural spans. Then the final span is found by cross-validation.

Iteration #	Span	D^2	d.o.f.	Comments
	procedural spans			
start	1.0	1.110	2.0	principal component line
1	0.4	0.740	4.2	initial span
2	0.4	0.565	4.6	
3	0.4	0.550	4.7	
4	0.4	0.549	4.7	converged
5	0.3	0.376	5.7	reduce span
6	0.3	0.361	5.4	
7	0.3	0.360	5.4	converged
8	0.2	0.222	7.3	reduce span
9	0.2	0.217	6.9	
10	0.2	0.217	6.9	converged
	mse spans			
final	0.07, 0.09, 0.35	0.162 0.189	9.7	cross-validated

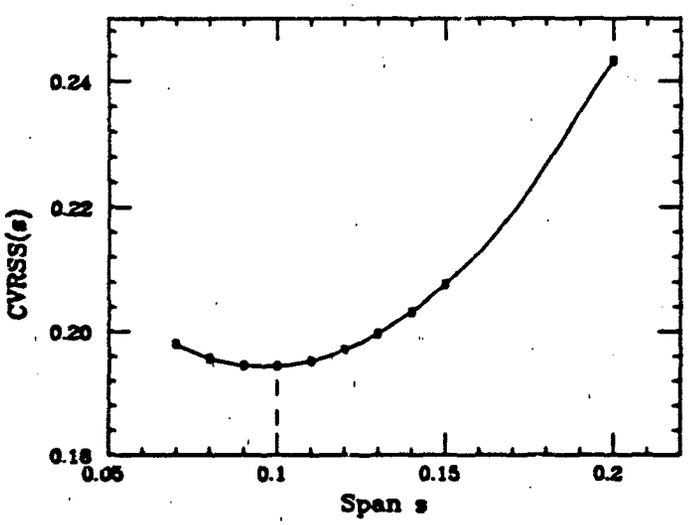


Figure 6.2c The cross-validation curve shows $CVRSS(s)$ as a function of the span s. One span is used for all 3 co-ordinates.

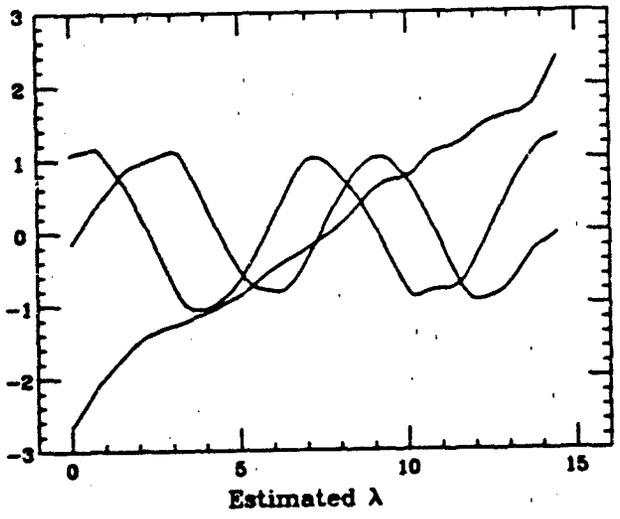


Figure 6.2d The estimated co-ordinate functions for the helix, using the span found in figure 6.2c.

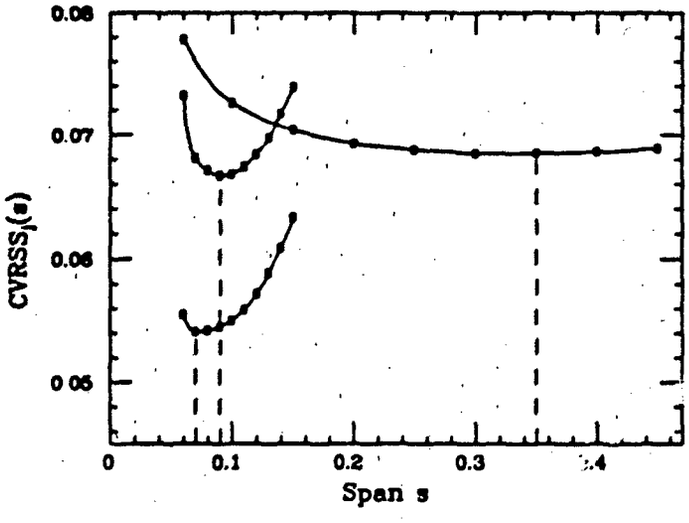


Figure 6.2e The cross-validation curve shows $CVRSS_j(s)$ as a function of the span s. A separate span is found for each co-ordinate.

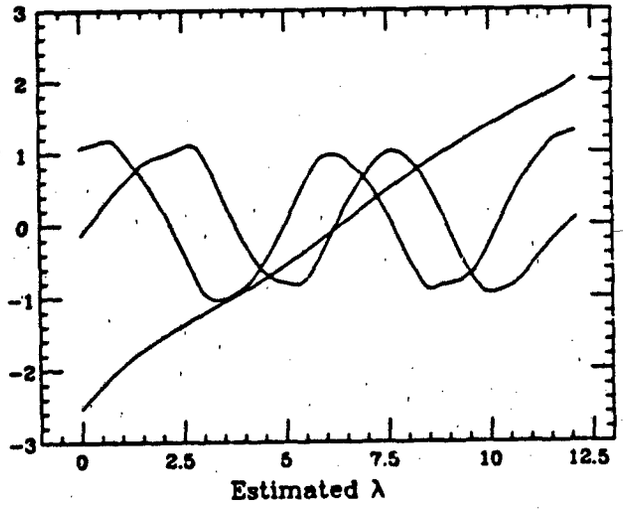


Figure 6.2f The estimated co-ordinate functions for the helix, using the spans found in figure 6.2f.

The entry labelled d.o.f. in table 6.1is an abbreviation for degrees of freedom. In linear regression the number of parameters used in the fit is given by $\mathrm{tr}(H)$ where H is the projection or *hat* matrix. If the response variables y_i are iid with variance σ^2, then

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{Var}(\hat{y}_i) &= \sum_{i=1}^{n} \mathrm{Var}(h_i' y) \\ &= \sigma^2 \mathrm{tr}(H'H) \\ &= \sigma^2 \mathrm{tr}(H)\end{aligned}$$

We can do the same calculation for a linear smoother matrix C, and in fact for the local straight lines smoother we even have $\mathrm{tr}(C'C) = \mathrm{tr}(C)$. As the span decreases, the diagonal entries of C get larger, and thus the variance of the estimates increases, as we would expect. One can also approach this from the other side by looking at the residual sum of squares. In the absence of bias we have

$$\begin{aligned}\mathbf{E}\,RSS &= \mathbf{E}\,\|(I-C)y\|^2 \\ &= \mathbf{E}\,y'(I-C)'(I-C)y \\ &= \mathrm{tr}[(I-C)'(I-C)\ \mathrm{Cov}(y)] \\ &= (n - \mathrm{tr}(C))\sigma^2\end{aligned} \tag{6.3}$$

if $\mathrm{tr}(C'C) = \mathrm{tr}(C)$. * More motivation for regarding $\mathrm{tr}(C)$ as the number of parameters or d.o.f. can be found in Cleveland (1979) and Tibshirani (1984). Some calculations similar to those in 3.5.1 show that the expected squared distance of X from the true f is $D^2 \approx 2\sigma^2$, or more precisely $D^2 \approx 2\sigma^2 - \sigma^4/(4\rho^2)$ where ρ is the radius of curvature, which in our example is $1 + 1/\pi^2$. Thus $D^2 \approx 0.18$. The cross validated residual estimate $\sum CVRSS_j$ was found to be 0.189. The orthogonal distance from the final curve is $D^{2(11)} = 0.162$. This is deflated due to overfitting. The average value of d.o.f for the final curve is (one for each co-ordinate) 9.7, or a total of 29.1. Some simple heuristics show that the we should scale this value up by by $2n/(2n - d.o.f) = 300/(300 - 29.1) = 1.11$. We then get $2n/(2n - d.o.f)D^{2(11)} = 0.179$ which is back in the correct ballpark.

It is more convenient to view the 3 dimensional examples on a color graphics system (such as the Chromatics system of the Orion group, Stanford University). This allows one to rotate the points in real time and thus see the 3rd dimension.

---

* For our smoothers, each row of C is the row of a projection matrix, and hence $c_i' c_i = c_{ii}$.

Example 6.3. Geological data.

This real data example illustrates:

- Data modelling in 3 dimensions,
- non-linear factor analysis, and
- outlier detection and robust fitting.

The data in this example consists of measurements of the mineral content of 64 core samples, each taken at different depths (Chernoff, 1973). Measurements were made of 10 minerals in each sample. We simply label the minerals $X_1, \cdots, X_{10}$, and analyze the first three.

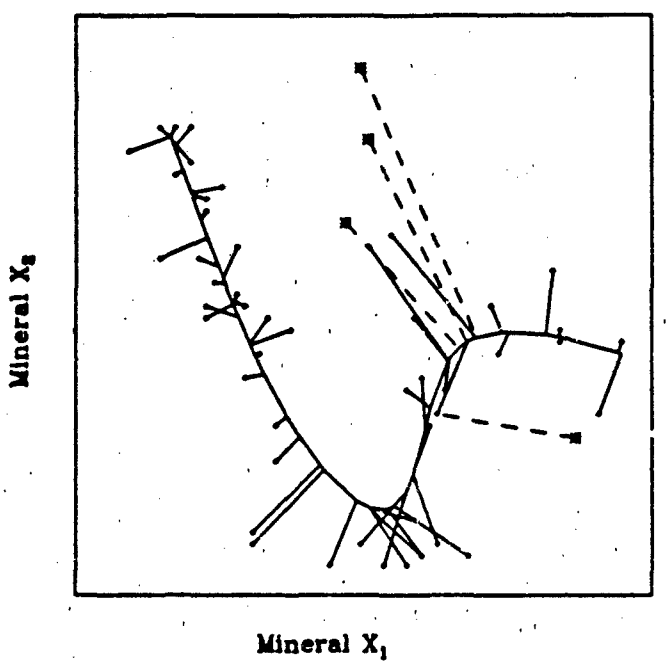


Figure 6.3a The principal curve for the mineral data. (Variable X_3 is into the page). The spikes join the points to their projection on the curve. The 4 outliers are joined to the curve with the broken lines.

Figure 6.3a shows the data and the solution curve. (A final span of 0.35 was manually selected.) In 3-D the picture looks like a dragon with its tail pointing to the left and the

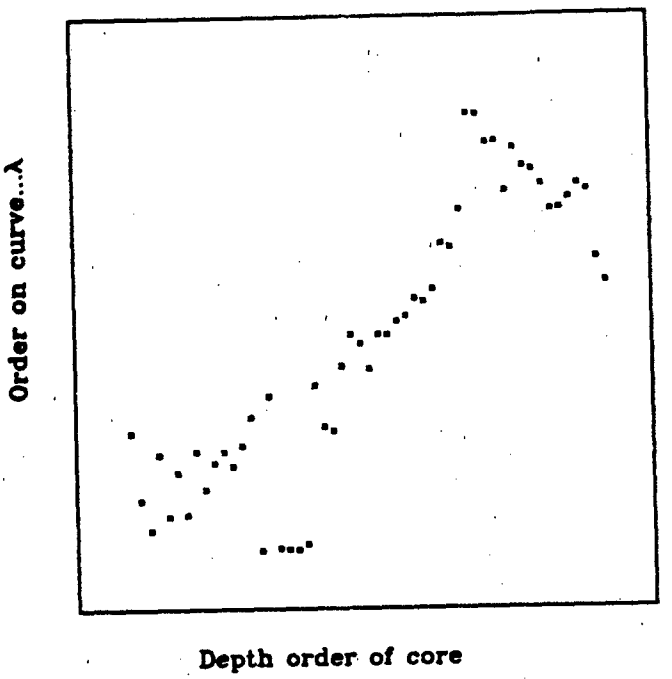


Figure 6.3b The values $\lambda_f(x_i)$ are plotted against the depth order of the core samples.

long (outlier) spikes could be a mane. The linear principal component explains 55% of the variance, whereas this solution explains 82%.

The spikes join the obeervations to their closest projections on the curve. This is a useful device for spotting outliers. A robust version of the principal curve procedure was used in this example. After the first iteration, points receive a weight which is inversly proportional to their distance from the curve. In the smoothing step, a weighted smooth is used, and if the weight is below a certain threshhold, it is set to 0. Four points were identified as outliers, and are labelled differently in figure 6.3*a* . We would really consider them model outliers, since in that region of the curve the model does not appear to fit very well.

Figure 6.3*b* shows the relationship between the order of the points on the curve, and the depth order of the core samples. The curve appears to recover this variable for the most part. The area where it does not recover the order is where the curve appears to fit the data badly anyway. So here we have uncovered a hidden variable or factor that we are able to validate with the additional information we have about the ordering. The co-ordinate

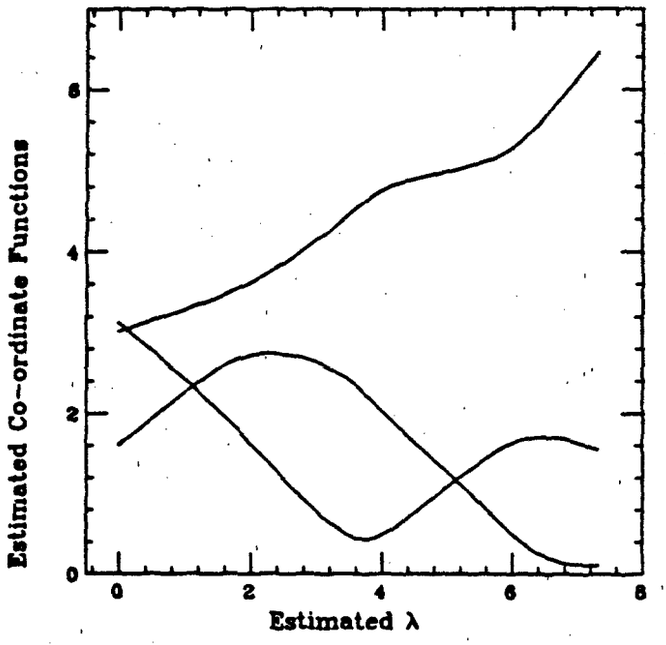


Figure 6.3c The estimated co-ordinate functions or *factor loading curves* for the three minerals.

plots would then represent the mean level of the particular mineral at different depths (see figure 6.3c). Usually one would have to use these co-ordinate plots to identify the factors, just as one uses the factor loadings in the linear case.

Example 6.4. The uniform ball.

This example illustrates:

- A principal surface in 3 space, and
- a connection to multidimensional scaling.

The data is artificially constructed, with no noise, by generating points uniformly from the surface of a sphere. It is the same data used by Shepard and Carroll (1966) to demonstrate their parametric mapping algorithm. (see reference and chapter 7). We simply use it here to demonstrate the ability of the principal surface algorithm to produce surfaces that are not a function of the starting plane (in analogy to the circle example in chapter 3).

There are 61 data points, as shown in figure 6.4a. One point is placed at each intersection of 5 equally spaced parallels and 12 equally spaced meridians. The extra point

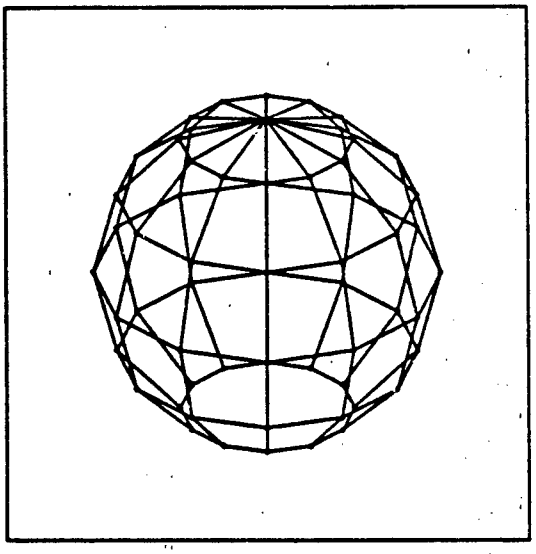


Figure 6.4a The data points are placed in a uniform pattern on the surface of a sphere. The south pole is missing.

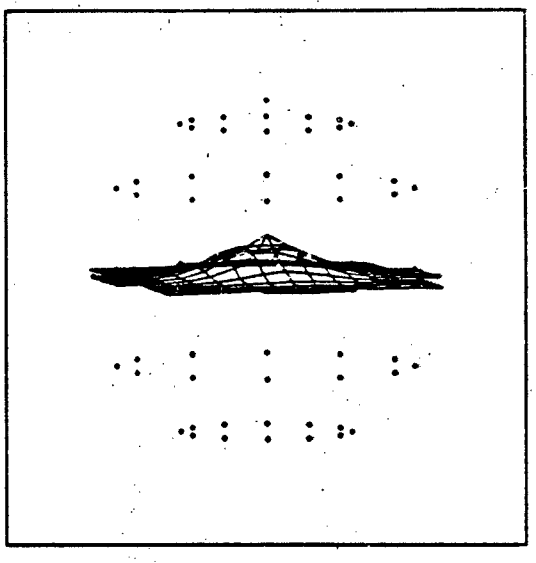


Figure 6.4b The second iteration of the principal surface procedure finds a surface that is a function of the first iteration.

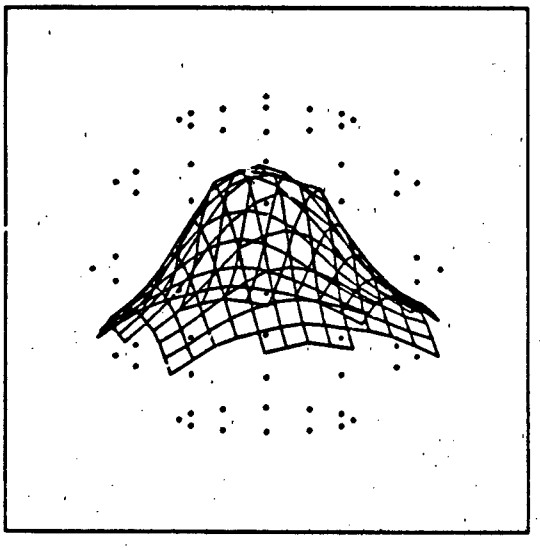


Figure 6.4c An intermediate stage in the iterations.

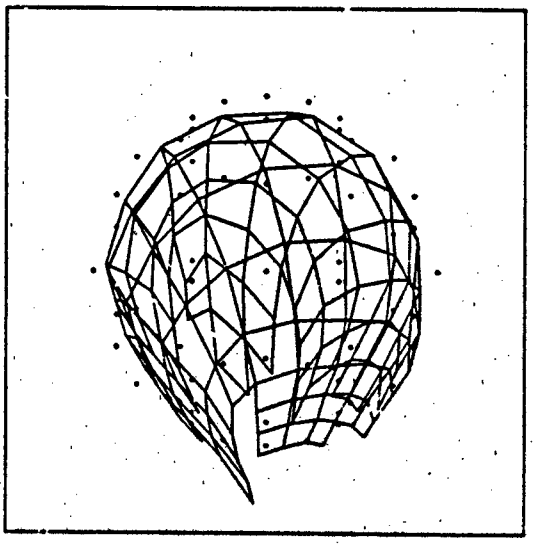


Figure 6.4d The final surface produced by the principal surface routine.

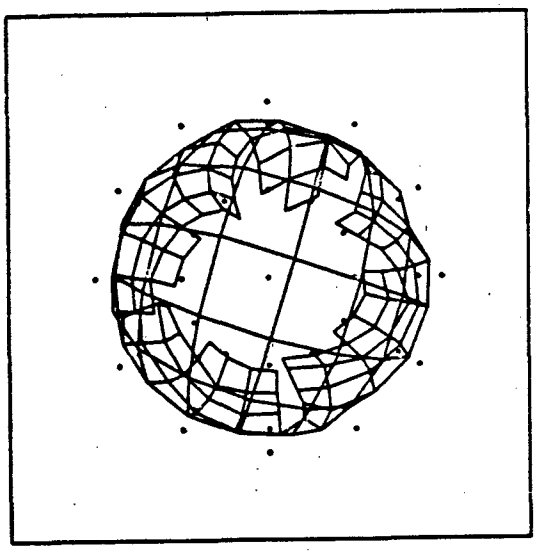


Figure 6.4e Another view of the final principal surface.

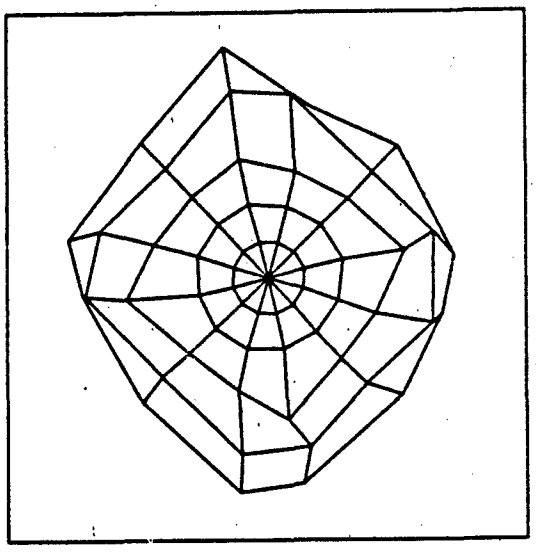


Figure 6.4f The λ map is a two dimensional summary of the data. It resembles a stereographic map of the world.

is placed at the north pole. (If we placed a point at the south pole the principal surface procedure would never move from the starting plane, which is in fact a principal surface.) Figures 6.4b to 6.4d show various stages in the iterative procedure, and figure 6.4e shows another view of the final surface. Figure 6.4f is a parameter map of the two dimensional $\hat{\lambda}$. It resembles a stereographic map of the earth. (A stereographic map is obtained by placing the earth, or a model thereof, on a piece of paper. Each point on the surface is mapped onto the paper by extrapolating the line segment joining the north pole to the point until it reaches the paper.) Points in the southern hemisphere are mapped on the inside of a circle, points in the northern hemisphere on the outside, and there is a discontinuity at the north pole. Points close together on this map are close together in the original space, but the converse is not necessarily true. This map provides a two dimensional summary of the original data. If we are presented with any new observations, we can easily locate them on the map by finding their closest position on the surface.

Example 6.5. One dimensional color data.

This almost real data example illustrates:

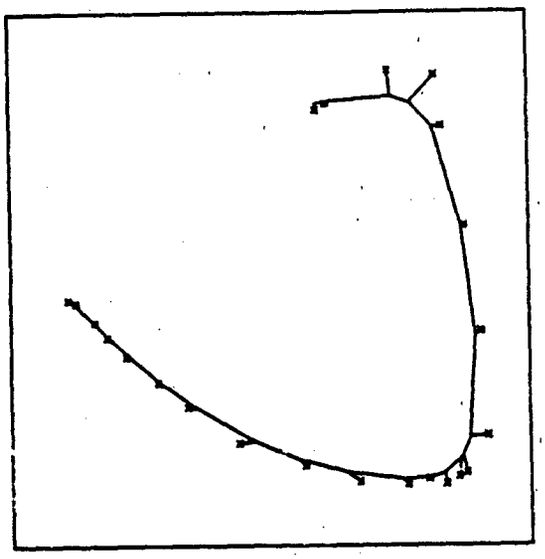


Figure 6.5a The 4 dimensional color data projected onto the first principal component plane. The principal curve, found in the original four-space, is also projected onto this plane.

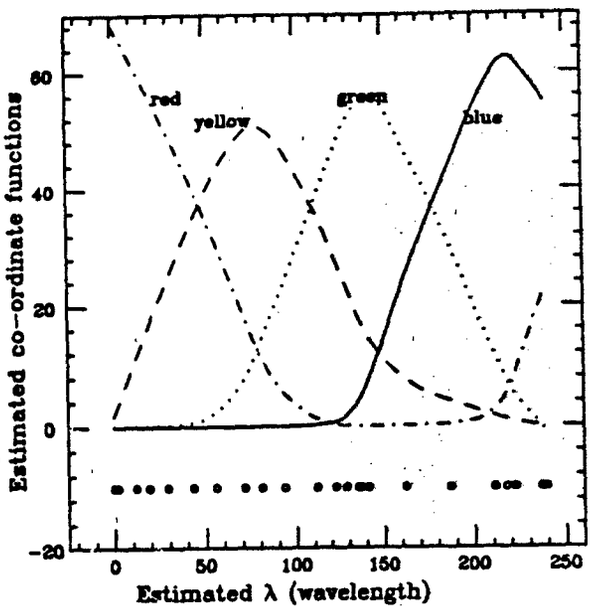


Figure 6.5b The estimated co-ordinate functions plotted against the arc length of the principal curve. This $\hat{\lambda}$ will be monotone with the true wavelength.

- A principal curves in 4-space, and
- a one dimensional MDS example.

These data were used by Shepard and Carroll (1966) (who cite the original source as Boynton and Gordon (1965)) to illustrate a version of their parametric data representation techniques called proximity analysis. We give more details of this technique in chapter 7.

Each of the 23 observations represents a spectral color at a specific wavelength. Each observation has 4 psychological variables associated with it. They are the relative frequencies with which 100 observers named the color *blue, green, yellow* and *red*. As can be seen in figure 6.5a, there is very little error in this data, and it is one dimensional by construction. Since the color changes slowly with wavelength, so should these relative frequencies, and they should thus fall on a one dimensional curve, as they do. The data, by construction lies in a 3 dimensional simplex since the four variables add up to 1. The pictures we show are projections of this simplex onto the 2-D subspace spanned by the first two linear principal components. Figure 6.5a shows the solution curve and figure 6.5b shows the recovered parameters and co-ordinate functions. This solution is in qualitative agreement with the data and with the solution produced by Shepard and Carroll.

Example 6.6. Lipoprotein data.

This real data example illustrates:

- A principal surface in 3 space with some interpretations,
- a principal curve suggested by the surface, and
- co-ordinate plots for surfaces.

Williams and Krauss (1982) conducted a study to investigate the inter-relationships between the serum concentrations of lipoproteins at varying densities in sedentry men. We focus on a subset of the data, and consider the serum concentrations of LDL 3-4 (Low Density Lipoprotein with flotation rates between S_f3-4), LDL 7-8, and HDL 3 (High Density Lipoprotein) in the sample of 81 men. Figures 6.6a-d are different views of the principal surface found for the data. Quantitively this surface explains 97.4% of the variability in the data, and accounts for 80% of the residual variance unexplained by the principal component plane. Qualitatively, we see that the surface has interesting structure in only two of the co-ordinates, namely LDL 3-4 and LDL 7-8. We can infer from the the surface that the bow shaped relationship between these two variables does not change for varying levels of HDL 3. It exhibits an independent behaviour. We have included a co-ordinate plot (figure 6.6e) of the estimated co-ordinate function for the variables LDL 7-8 which helps confirm this claim. The relationship between LDL 7-8 and $(\hat{\lambda}_1, \hat{\lambda}_2)$ depends mainly on the level of $\hat{\lambda}_1$. Similar information is conveyed by the other co-ordinate plots, or can be seen from the estimated surface directly. This suggests a model of the form

$$\begin{pmatrix} \text{LDL 3-4} \\ \text{LDL 7-8} \\ \text{HDL 3} \end{pmatrix} = \begin{pmatrix} f_1(\lambda_1) \\ f_2(\lambda_1) \\ f_3(\lambda_2) \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}.$$

As specified λ_2 is confounded with HDL 3, and is thus unidentifiable. We need to estimate the first two components of the model. This is a principal curve model, and figure 6.6f shows the estimated curve. It exhibits the same dependence between LDL 7-8 and LDL 3-4 as did the surface. The curve explains 92.6% of the variance in the two variables, whereas the principal component line explains only 80%.

Williams and Krauss performed a similar analysis looking at pairs of variables at a time. We discuss their techniques in chapter 7. Their results are qualitatively the same as ours for the LDL pair.

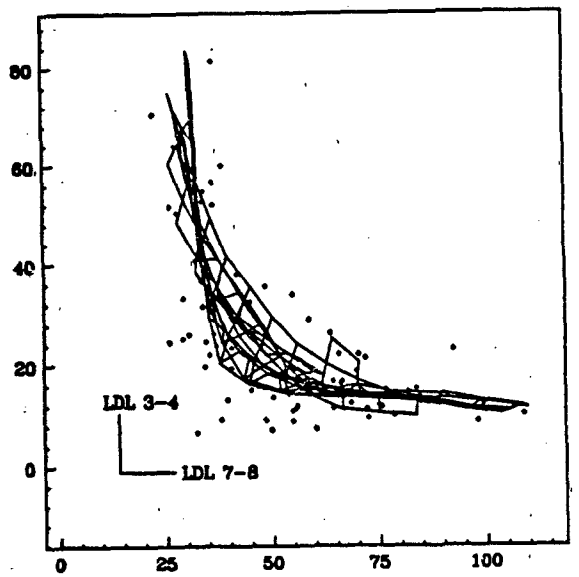


Figure 6.6a The principal surface for the serum concentrations LDL 7-8, LDL 3-4 and HDL 3 in a sample of 81 sedentary men. Variable HDL 3 is into the page.

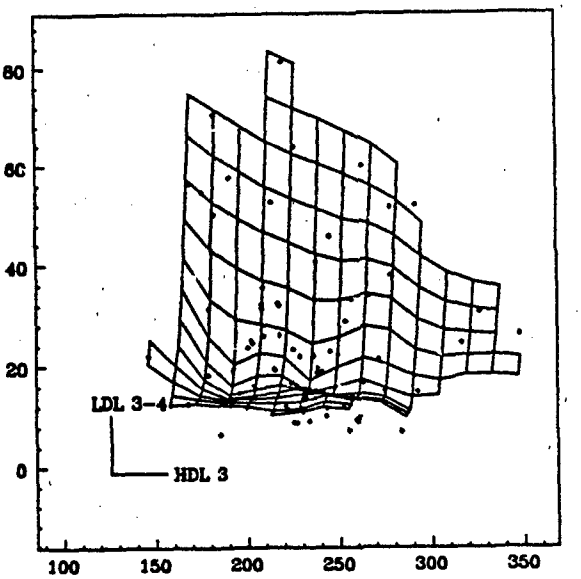


Figure 6.6b The principal surface as in figure 6.6a from a different viewpoint. Variable LDL 7-8 is into the page.

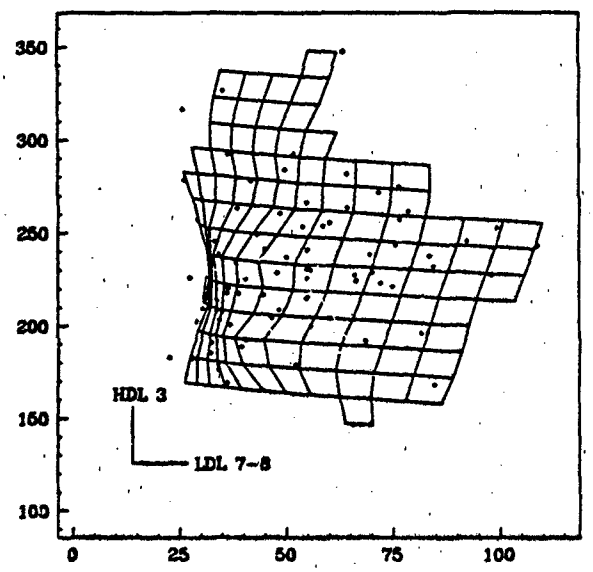


Figure 6.6c The principal surface as in figure 6.6a from a different viewpoint. Variable LDL 3-4 is into the page.

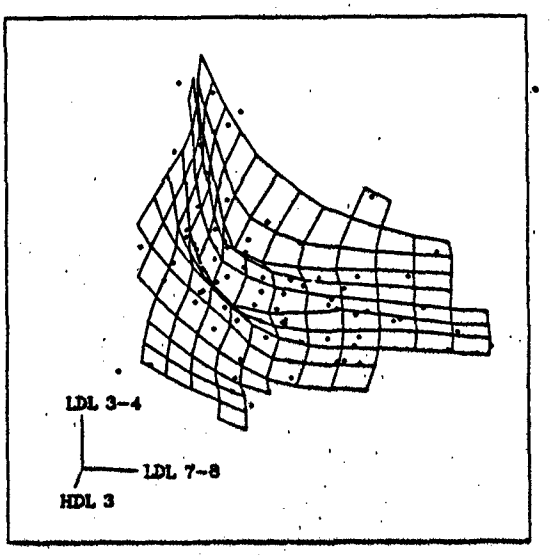


Figure 6.6d The principal surface as in figure 6.6a from a slightly oblique perspective.

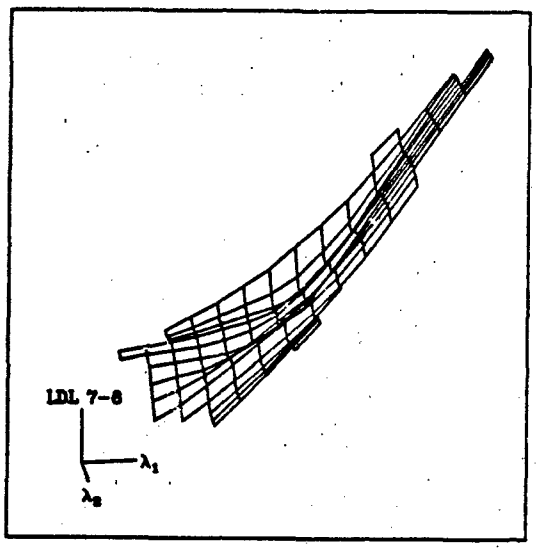


Figure 6.6e The estimated co-ordinate function for LDL 7-8 versus $\hat{\lambda}$. $\hat{\lambda}_2$ has little effect.

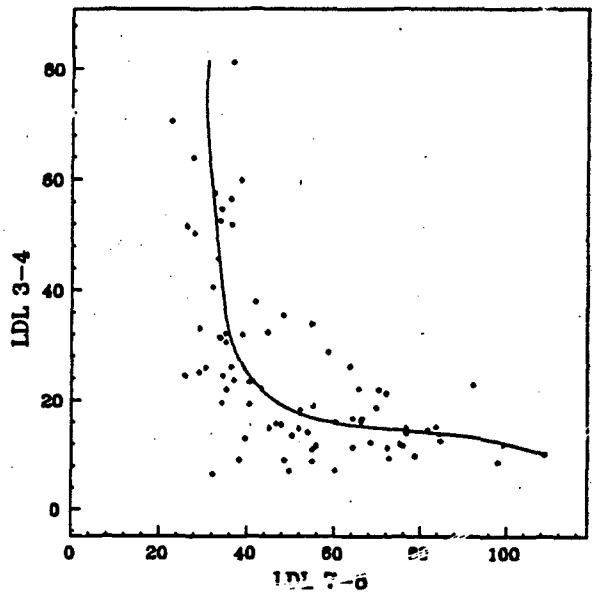


Figure 6.6f The principal curve for the serum concentrations LDL 7-8 and LDL 3-4 in a sample of 81 sedentry men.

Chapter 7

Discussion and conclusions

In this chapter we discuss some of the existing techniques for symmetric smoothing, as well as the various generalizations of principal components and factor analysis. We compare these techniques with the methodology developed here. The chapter concludes with a summary of the uses of principal curves and surfaces.

7.1. Alternative techniques.

Other non-linear generalizations of principal components exist in the literature. They can be broadly classified according to two dichotomies.

- We can estimate either the non-linear manifold or the non-linear constraint that defines the manifold. In linear principal components the approaches are equivalent.
- The non-linearity can be achieved by transforming the space or by transforming the model.

The principal curve and surface procedures model the non-linear manifold by transforming the model.

7.1.1. Generalized linear principal components.

This approach corresponds to modeling either the nonlinear constraint or the manifold by transforming the space. The idea here is to introduce some extra variables, where each new variable is some non-linear transformation of the existing co-ordinates. One then seeks a subspace of this non linear co-ordinate system that models the data well. The subspace is found by using the usual linear eigenvector solution in the new *enlarged* space. This technique was first suggested by Gnanadesikan & Wilk (1966, 1968), and a good description can be found in Gnanadesikan (1977). They suggested using polynomial functions of the original p co-ordinates. The resulting *linear* combinations are then of the form (for $p = 2$ and quadratic polynomials)

$$\lambda_j = a_{1j}x_1 + a_{2j}x_2 + a_{3j}x_1x_2 + a_{4j}x_1^2 + a_{5j}x_2^2 \qquad (7.1)$$

and the a_j will be eigenvectors of the appropriate covariance matrix.

This model has appeal mainly as a dimension reducing tool. Typically the linear combination with the smallest variance is set to zero. This results in an implicit non-linear constraint equation as in (7.1) where we set $\lambda = 0$. We then have a rank one reduction that tells us that the data lies close to a quadratic manifold in the original co-ordinates.

The model has been generalized further to include more general transformations of the co-ordinates other than quadratic, but the idea is essentially the same as the above; a linear solution is found in a transformed space. Young, Takane & de Leeuw (1978) and later Friedman (1983) suggested different forms of this generalization to include non-parametric transformations of the co-ordinates. The problem can be formulated as follows: Find a and $s'(x) = (s_1(x_1), \cdots, s_p(x_p))$ such that

$$\mathbf{E}\,\|s(x) - as'(x)\|^2 = \min! \tag{7.2}$$

or alternatively such that

$$\mathrm{Var}[a's(x)] = \max! \tag{7.3}$$

where $\mathbf{E}s_j(x_j) = 0$, $a'a = 1$ and $\mathbf{E}s_j^2(x_j) = 1$. The idea is to transform the coordinates suitably and then find the linear principal components. If in (7.3) we replaced *max* by *min* then we would be estimating the constraint in the transformed space.

The estimation procedure alternates between finding the $s_j(\cdot)$ and finding the linear principal components in the transformed space

- For a fixed vector of functions $s(\cdot)$, we chose a to be the first principal component of the covariance matrix $\mathbf{E}s(x)s(x)'$.
- For a known, (7.2) can be written in the form

$$k\,\mathbf{E}[s_1(x_1) - \sum_{j=2}^{p} b_{1j}s_j(x_j)]^2 + \cdots \text{ terms in } s_2(\cdot), \cdots, s_p(\cdot), \tag{7.4}$$

and b_{j1} are functions of a above. If $s_2, \ldots, s_p$ are known, equation (7.4) is minimized by

$$s_1(x_1) = \mathbf{E}(\sum_{j=2}^{p} b_{1j}s_j(x_j) \mid x_1)$$

This is true for any s_j, and suggests an inner iterative loop. This inner loop is very similar to the ACE algorithm (Breiman and Friedman, 1982), except the normalization

is slightly different. Breiman and Friedman proved that the ACE algorithm converges under certain regularity conditions in the distributional case.

The disadvantages of this technique are:

- The space is transformed, and in order to understand the resultant fit, we usually would need to transform back to the original space. This can only be achieved if the transformations are restricted to monotone functions. In the transformed space the estimated manifold is given by

$$\begin{pmatrix} \hat{s}_1(x_1) \\ \vdots \\ \hat{s}_p(x_p) \end{pmatrix} = aa's(x).$$

  Thus if the $s_j(\cdot)$ are monotone, we get untransformed estimates of the form

$$\begin{pmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_p \end{pmatrix} = \begin{pmatrix} s_1^{-1}(a_1 z) \\ \vdots \\ s_p^{-1}(a_p z) \end{pmatrix} \tag{7.5}$$

  where $z = a's(x)$. Equation (7.5) defines a parametrized curve. The curve is not completely general since the co-ordinate functions are monotone. For the same reason, Gnanadesikan (1978) expressed the desirability of having procedures for estimating models of the type proposed in this dissertation.

- We are estimating manifolds that are close to the data in the transformed co-ordinates. When the transformations are non-linear this can result in distortion of the error variances for individual variables. What we really require is a method for estimating manifolds that are close to the data in the original p co-ordinates. Of course, if the functions are linear, both approaches are identical.

An advantage of the technique is that it can easily be generalised to take care of higher dimensional manifolds, although not in an entirely general fashion. This is achieved by replacing a with A where A is $p \times q$. We then get a q dimensional hyperplane in the transformed space given by $AA's(x_i)$. However, we end up with a *number* of implicit constraint equations which are hard to deal with and interpret. Despite the problems associated with generalized principal components, it remains a useful tool for performing rank 1 dimensionality reductions.

7.1.2. Multi-dimensional scaling.

This is a technique for finding a low dimensional representation of high dimensional data. The original proposal was for data that consists of $\binom{n}{2}$ dissimilarities or distances between n objects. The idea is to find a m (m small, 1, 2 or 3) dimensional euclidean representation for the objects such that the inter-object distances are preserved as well as possible. The idea was introduced by Torgerson (1958), and followed up by Shepard (1962), Kruskal (1964a ,1964b), Shepard & Kruskal (1964) and Shepard & Carroll (1966). Gnanadcsikan (1978) gives a concise description.

The procedures have also been suggested for situations where we simply want a lower dimensional representation of high dimensional euclidean data. The lower dimensional representation attempts to reproduce the interpoint distances in the original space. We fit a principal curve to the color data in example 6.5; these data were originally analyzed by Shepard and Carroll (1966) using MDS techniques. Although there have been some intriguing examples of the technique in the literature, a number of problems exist.

- The solution consists of a vector of m co-ordinates representing the location of points on the low dimensional manifold, but *only for the n data points*. What we don't get, and often desire is a mapping of the whole space. We are unable, for example, to find the location of new points in the reduced space.

- The procedures are computationally expensive and unfeasible for large n ($nm > 300$ is considered large). They are usually expressed as non-linear optimization problems in nm parameters, and differ in the choice of criterion.

The principal curve and surface procedures partially overcome both the problems listed above; they are unable to find structures as general as those that can be found by the MDS procedures due to the averaging nature of the scatterplot smoothers, but they do provide a mapping for the space. We have demonstrated their ability to model MDS type data in examples 6.4 and 6.5. They do not, however, provide a model for dissimilarities which was the original intention of multidimensional scaling.

7.1.3. Proximity models.

Shepard & Carroll (1966) suggested a functional model similar in form to the model we suggest. They required only to estimate the n vectors of m parameters for each point, and considered the data to be functions thereof. The parameters (nm altogether) are found

by direct search as in MDS, with a different criterion to be minimized. Their procedure, however, was geared towards data without error, as in the ball data in example 6.4. This becomes evident when one examines the criterion they used, which measures the continuity of the data as a function of the parameters. When the data is not smooth, as is usually the case, we need to estimate functions that vary smoothly with the parameters, and are close to the data.

7.1.4. Non-linear factor analysis.

More recently, Etezadi-Amoli and McDonald (1983) approached the problem of non-linear factor analysis using polynomial functions. They use a model of the form

$$X = f(\lambda) + e$$

where f is a polynomial in the unknown parameters or factors. Their procedure for estimating the unknown factors and coefficients is similar to ours in this restricted setting. * Their emphasis is on the factor analysis model, and once the appropriate polynomial terms have been found, the problem is treated as an enlarged factor analysis problem. They do not estimate the λ's as we do, using the geometry of the problem, but instead perform a search in nq parameter space, where q is the dimension of λ and n is the number of observations. Our emphasis is on providing one and two dimensional summaries of the data. In certain situations, these summaries can be used as estimates of the appropriate non-linear functional and factor models.

7.1.5. Axis interchangeable smoothing.

Cleveland (1983) describes a technique for symmetrically smoothing a scatterplot which he calls *axis interchangeable smoothing* (which we will refer to as AI smoothing). We briefly outline the idea:

- standardize each coordinate by some (robust) measure of scale.
- rotate the coordinate axes by 45°. (if the correlation is positive, else rotate through $-45°$).
- smooth the transformed y against the transformed x.

---

* Their paper was published in the September, 1983 issue of Psychometrika, whereas Hastie (1983) appeared in July.

- rotate the axes back.
- unstandardize.

If the standardization uses regular standard deviations, then the rotation is simply a change of basis to the principal component basis. The resulting curve minimizes the distance from the points orthogonal to this principal component. It has intuitive appeal since the principal component is the line that is closest in distance to the points. We then allow the points to *tug in* the principal component line. It is simple and fast to compute the AI Smooth, and for many scatterplots it produces curves that are very similar to the principal curve solution. This is not surprising when we consider the following theorem:

Theorem 7.1

If the two variables in a scatterplot are standardized to have unit standard deviations, and if the smoother used is linear and reproduces straight lines exactly, then the axis interchangeable smooth is identical to the curve of the first iteration of the principal curve procedure.

Proof

Let the variables x and y be standardized as above. The AI Smooth transforms to two new variables

$$\begin{aligned} x^* &= \frac{(x+y)}{\sqrt{2}} \\ y^* &= \frac{(x-y)}{\sqrt{2}} \end{aligned} \tag{7.6}$$

Then the AI Smooth replaces (x^*, y^*) by $(x^*, \text{Smooth}(y^* \mid x^*))$. But $\text{Smooth}(x^* \mid x^*) = x^*$ since the smoother reproduces straight lines exactly.* Thus the AI Smooth transforms back to

$$\begin{aligned} \hat{x} &= \frac{(\text{Smooth}(x^* \mid x^*) + \text{Smooth}(y^* \mid x^*)}{\sqrt{2}} \\ \hat{y} &= \frac{(\text{Smooth}(x^* \mid x^*) - \text{Smooth}(y^* \mid x^*))}{\sqrt{2}} \end{aligned} \tag{7.7}$$

Since the smoother is linear, and in view of (7.6), (7.7) becomes

$$\begin{aligned} \hat{x} &= \text{Smooth}(x \mid x^*) \\ \hat{y} &= \text{Smooth}(y \mid x^*) \end{aligned} \tag{7.8}$$

---

* Any weighted local linear smoother has this property. Local averages, however, do not unless the predictors are evenly spaced.

This is exactly the curve found after the first iteration of the principal curve procedure, since $\hat{\lambda}^{(0)} = x^*$. ∎

Williams and Krauss (1982) extended the AI smooth by iterating the procedure. At the second step, the residuals are calculated locally by finding the tangent to the curve at each point and evaluating the residuals from these tangents. The new fit at that point is the smooth of these residuals against their projection onto the tangent. This procedure would probably get closer to the principal curve solution than the AI smooth (we have not implemented the Williams and Krauss smooth). Analytically one can see that the procedures differ from the second step on.

This particular approach to symmetric smoothing (in terms of residuals) suffers from several deficiencies :

- the type of curves that can be found are not as general as those found by the principal curve procedure.
- they are designed for scatterplots and do not generalize to curves in higher dimensions.
- they lack the interpretation of principal curves as a form of conditional expectation.

7.2. Conclusions.

In conclusion we summarize the role of principal curves and surfaces in statistics and data analysis.

- They generalize the one and two dimensional summaries of multivariate data usually provided by the principal components.
- When the principal curves and surface are linear, they are the principal component summaries.
- Locally they are the critical points of the usual distance function for such summaries; this gives an indication that there are not too many of them.
- They are defined in terms of conditional expectations which satisfies our mental image of a summary.
- They provide the least squares estimate for generalized versions of factor analysis, functional models and the errors in variables regression models. The non-linear errors

in variables model has been used successfully a number of times in practical data analysis problems (notably calibration problems).

- In some situations they are a useful alternative to MDS techniques, in that they provide a lower dimensional summary of the *space* as opposed to the *data set*.
- In some situations they can be effective in identifying outliers in higher dimensional space.
- They are a useful data exploratory tool. Motion graphics techniques have become popular for looking at 3 dimensional point clouds. Experience shows that it is often impossible to identify certain structures in the data by simply rotating the points. A summary such as that given by the principal curve and surfaces can identify structures that would otherwise be transparent, even if the data could be viewed in a real three dimensional model.

Acknowledgements

My great appreciation goes to my advisor Werner Stuetzle, who guided me through all stages of this project. I also thank Werner and Andreas Buja for suggesting the problem, and Andreas for many helpful discussions. Rob Tibshirani helped me a great deal, and some of the original ideas emerged whilst we were suntanning alongside a river in the Californian mountains. Brad Efron, as usual, provided many insightful comments. Thanks to Jerome Friedman for his ideas and constant support. In addition I thank Persi Diaconis and Iain Johnstone for their help and comments, and Roger Chaffee and Dave Parker for their computer assistance. Finally I thank the trustees of the Queen Victoria, the Sir Robert Kotas and the Sir Harry Crossley scholarships for their generous assistence.

Bibliography

Anderson, T.W. (1982), *Estimating Linear Structural Relationships,* Technical Report # 389, Institute for Mathematical studies in the Social Sciences, Stanford University, California.

Barnett, V. (Ed) (1981), *Interpreting Multivariate Data,* Wiley, Chichester.

Becker, R.A. and Chambers, J.M. (1984), *S: An Interactive Environment for Data Analysis and Graphics,* Wadsworth, California.

Boynton, R.M. and Gordon, J. (1965), *Bezold-Brüke Hue Shift Measured by Color-Naming Technique,* J. Opt. Soc. Amer, 55 , 78-86.

Breiman, L. and Friedman, J.H. (1982), *Estimating Optimal Transformations for Multiple Regression and Correlation,* Dept. of Statistics Tech. Rept, Orion 16, Stanford University.

Chernoff, H. (1973), *The use of Faces to Represent Points in k-dimensional Space Graphically,* Journal of the American Statistical Association, 68, #342, 361-368.

Chung, K.L. (1974), *A Course in Probability Theory,* Academic Press, New York.

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

Cleveland, W.S. (1983), *The Many Faces of a Scatterplot,* Submitted for publication.

Craven, P. and Wahba, G. (1979), *Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-validation,* Numer. Math., 31, 377-403.

do Carmo, M.P. (1976), *Differential Geometry of Curves and Surfaces,* Prentice Hall, New Jersy.

Etezadi-Amoli, J. and McDonald, R.P. (1983), *A Second Generation Nonlinear Factor Analysis,* Psychometrika, 48, #3, 315-342.

Efron, B. (1981), *Non-parametric Standard Errors and Confidence Intervals,* Canadian Journal of Statistics, 9, 139-172.

Efron, B. (1982), *The Jacknife, the Bootstrap and other Resampling Plans,* SIAM-CBMS, 38.

Efron, E. (1984), *Bootstrap Confidence Intervals for Parametric Problems,* Technical Report #90, Division of Biostatistics, Stanford University.

Friedman, J.H. (1983), *Personal communication.*

Friedman, J.H, Bently, J.L. and Finkel, R.I. (1976), *An Algorithm for Finding Best Matches in Logarithmic Expected Time.* STAN-CS-75-482, Stanford University.

Friedman, J.H. and Stuetzle, W. (1982), *Smoothing of Scatterplots,* Dept. of Statistics Tech. Rept. Orion 3, Stanford University.

Gasser, Th. and Muller, H.G. (1979), *Kernel Estimation of Regression Functions,* in *Smoothing Techniques for Curve Estimation,* Proceedings, Heidelberg, Springer Verlag.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations,* Wiley, New York.

Gnanadesikan, R. and Wilk, M.B. (1969), *Data Analytic Methods in Multivariate Statistical Analysis,* in Multivariate Analysis II (P.R. Krishnaiah, ed.), Academic Press, New York.

Golub, G.H. and Reinsch, C. (1970), *Singular Value Decomposition and Least Squares Solutions,* Numer. Math. 14, 403-420

Golub, G.H. and van Loan, C. (1979), *Total Least Squares,* in *Smoothing Techniques for Curve Estimation,* Proceedings, Heidelberg, Springer Verlag.

Greenacre, M. (1984), *Theory and Applications of Correspondence Analysis,* Academic Press, London.

Hastie, T.J. (1983), *Principal Curves,* Dept. of Statistics Tech. Rept. Orion 24, Stanford University.

Hastie, T.J. and Stuetzle, W. (1984), *Principal Curves and Surfaces,* (Motion Graphics Movie), Dept. of Statistics, Stanford University.

Hotelling, H. (1933), *Analysis of a Complex of Statistical Variables into Principal Components,* J. Educ. Psych., 24, 417-441, 498-520.

Kendall, M.G. and Stuart, A. (1961), *The Advanced Theory of Statistics,* Volume 2, Hafner, New York.

Kruskal, J.B. (1964a), *Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis,* Psychometrika, 29, #1, 1-27.

Kruskal, J.B. (1964b), *Nonmetric Multidimensional Scaling: a Numerical Method,* Psychometrika, 29, #2, 115-129.

Lindley, D.V. (1947), *Regression Lines and the Linear Functional Relationship,* Journal of the Royal Statistical Society, Supplement, 9, 219-244.

Madansky, A. (1959), *The Fitting of Straight Lines when both Variables are Subject to Error,* Journal of the American Statistical Society, 54, 173-205.

Mosteller, F. and Tukey, J. (1977), *Data Analysis and Regression,* Addison Wesley, Massachusetts.

Reinsch, C. (1967), *Smoothing by Spline Functions,* Numer. Math., 10, 177-183.

Shepard, R.N. (1962), *The Analysis of Proximities: Multidimensional Scaling with an unknown Distance Function,* Psychometrika, 27, 123-139, 219-246.

Shepard, R.N. and Carrol, J.D. (1966), *Parametric Representations of Non- Linear Data Structures,* in *Multivariate Analysis* (Krishnaiah, P.R.ed), Academic Press, NewYork.

Shepard, R.N. and Kruskal, J.B. (1964), *Non-metric Methods for Scaling and for Factor Analysis,* Amer. Psychologist, 19, 557-558.

Spearman, C. (1904), *General Intelligence, Objectively determined and Measures,* American Journal of Psychology, 15, 201-293.

Stone, M. (1977), *An Asymptotic choice of Model by Cross-validation and Akaike's Criterion,* Roy. Stat. Soc. B, 7, 44-47.

Thorpe, J.A. (1978), *Elementary Topics in Differential Geometry,* Springer-Verlag, New York. Undergraduate Text in Mathematics.

Tibshirani, R.J. (1984), *Bootstrap Confidence Intervals,* Technical Report #91, Division of Biostatistics, Stanford University.

Torgeson, W.S. (1958), *Theory and Methods of Scaling,* Wiley, New York.

Wilkinson, J.H. (1965), *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford.

Wahba, G. and Wold, S. (1975), *A Completely Automatic French Curve: Fitting Spline Functions by Cross-validation,* Comm. Statistics, 4, 1-7.

Williams, P.T. and Krauss, R.M. (1982), *Graphical Analysis of the Sectional Interrelationships among Subfractions of Serum Lipoproteins in Middle Aged Men,* unpublished manuscript, Stanford University.

Young, F.W, Takane, Y, and de Leuuw, J. (1978), *The Principal Components of Mixed Measurement Level Multivariate Data: an Alternating Least Squares Method with Optimal Scaling Features,* Psychometrika, 43, no.2.