Principal Curves and Surfaces

Trevor Hastie
and
Werner Stuetzle

TECHNICAL REPORT No. 56
January 1985

Department of Statistics
University of Washington
Seattle, Washington 98195
Principal Curves and Surfaces

Trevor Hastie
Department of Statistics
Stanford University
and
Computational Group
Stanford Linear Accelerator Center

Werner Stuetzle
Department of Statistics
University of Washington

Abstract

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

In this paper we develop some theory for principal curves and surfaces, and present an algorithm for estimating them. The main theorem states that principal curves are critical values of the expected squared distance between the points and the curve. We compare the principal curve and surface procedures to other generalizations of principal components in the literature; the usual generalizations transform the space, whereas we transform the model. There are also strong ties with multidimensional scaling.

Keywords and Phrases: Principal Components, Smoother, Non-parametric, Non-linear, Errors in variables, Symmetric.

* This work was supported by the Department of Energy under contracts DE-AC03-76SF and DE-AT03-81-ER10843, and by the Office of Naval Research under contract ONR N00014-81-K-0340, and by the U.S. Army Research Office under contract DAAG29-82-K-0056.
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 1a. It is natural to try and summarize the joint behaviour exhibited by the points in the scatterplot. The form of summary we choose 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.

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 usually estimated by least squares. This procedure is equivalent to finding the line that minimizes the sum of vertical squared deviations, as depicted in figure 1a.

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 1a 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 1b does just this — it is found by minimizing the orthogonal deviations.

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 1982), and spline smoothers (Reinsch 1967). In general scatterplot smoothers produce a smooth curve that attempts to minimize the vertical deviations as depicted in figure 1c. The non-parametric versions listed above allow the data to dictate the form of the non-linear dependency.

In this paper 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 1d. 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 formally define Principal Curves to be those smooth curves that are self consistent for a distribution or data set. This means that if we pick any point on the curve, collect all the data that "project" to this point and average them, then this average coincides with the point on the curve.

The algorithm for finding principal curves is equally intuitive. It starts with any smooth curve, usually a straight line such as the first principal component. It then tests to if it is self consistent by going through all the steps in the definition, i.e. projecting and averaging. The curve defined by the averaging either is or is not a principal curve. If not, the operation is repeated using it as a starting guess. This is iterated until (hopefully) convergence.
Figure 1a  The linear regression line minimizes the sum of squared errors in the response variable.

Figure 1b  The principal component line minimizes the sum of squared errors in all the variables.

Figure 1c  The smooth regression curve minimizes the sum of squared errors in the response variable, subject to smoothness constraints.

Figure 1d  The principal curve minimizes the sum of squared errors in all the variables, subject to smoothness constraints.
If 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. This is also the *total least squares* model of Golub and van Loan (1979).

- 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, as well as those already mentioned above (Anderson 1982).

In all the above situations the model can be written as

$$x_i = u_0 + a\lambda_i + e_i$$

where $u_0 + a\lambda_i$ is the systematic component and $e_i$ the random component. If we assume that $\text{Cov}(e_i) = \sigma^2 I$, then the least squares estimate of $a$ is the first linear principal component. See Hastie (1984) for more details of these models.

A natural generalization of (1) is the nonlinear model

$$x_i = f(\lambda_i) + e_i.$$  \hspace{1cm} (2)

This might then be a non-linear errors in variables, factor analysis or structural model. In the same spirit as above, where we used the first linear principal component to estimate (1), we propose using a principal curve as the estimate of the systematic component in (2). This broadens the scope and use of such curves considerably.

This paper 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. A 2-dimensional Principal Surface passes through the middle of the data, or is a self-consistent surface for a p-dimensional distribution. It can also be used to estimate the systematic component in the model

\[ x_i = f(\lambda_i) + e_i \]

where \( \lambda_i \) is a 2-dimensional vector.

2. The Principal Curve model.

2.1. 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.

2.1.1 One dimensional curves.

A one dimensional curve \( f \) is a vector of functions of a single variable, which we denote by \( \lambda \). These functions are called the coordinate functions, and \( \lambda \) 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 \( \lambda \), 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 \( \lambda_0 \) to \( \lambda_1 \) 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 \( \lambda \) is also in those units. The vector \( f'(\lambda) \) is tangent to the curve at \( \lambda \) and is sometimes called the velocity vector at \( \lambda \). 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 \( \lambda \), and for a unit speed curve, it is easy to check that it is orthogonal to the tangent vector. In this case \( f''/\| f'' \| \) is called the principal normal of the curve at \( \lambda \). 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. 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

\[ r_f(\lambda) \overset{\text{def}}{=} \text{radius of curvature of } f \text{ at } \lambda \]

\[ = 1/\| f''(\lambda) \|. \]
The radius of curvature is the radius of the circle tangent to the curve with the same acceleration as the curve.

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

2.1.2 Definition of principal curves.

Denote by $X$ a random vector in $p$-space with continuous probability density $h(x)$. Let $G$ be the class of differentiable 1-dimensional curves in $\mathbb{R}^p$, parametrized over some compact interval. 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 $G$. For $f \in G$ and $z \in \mathbb{R}^p$, we define the projection index $\lambda_f : \mathbb{R}^p \mapsto \Lambda_f$ by

$$
\lambda_f(z) = \sup_{\lambda} \{ \lambda : \|z - f(\lambda)\| = \inf_{\mu} \|z - f(\mu)\| \}. 
$$

(3)

The projection index $\lambda_f(z)$ of $z$ is the value of $\lambda$ for which $f(\lambda)$ is closest to $z$. There might be a number of such points (suppose $f$ is a circle and $z$ is at the center), so we pick the largest such value of $\lambda$. It can be shown (Hastie, 1984) that $\lambda_f(z)$ is a measureable mapping from $\mathbb{R}^p$ to $\mathbb{R}^1$, and thus $\lambda_f(X)$ is a random variable.

Definition

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

$$
E(X | \lambda_f(X) = \lambda) = f(\lambda) \text{ for a.e. } \lambda \in \Lambda_f
$$

We call the class of principal curves $F(h)$.

Figure 3 illustrates the definition of a principal curve. At any particular location $\lambda_0$ on the curve,
Figure 3. Each point on a principal curve is the average of the points that project there.

we collect all the points in $p$ space that have $f(\lambda_0)$ as their closest point on the curve. Then $f(\lambda_0)$ is the average of these points. Any curve that has this property is called a principal curve. One might say that principal curves are their own conditional expectation. In the next section we prove that 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 when we look at the bias associated with principal curves.

2.2. The distance property of principal curves.

The principal components are critical points of the average 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. We first rephrase the result for principal components, and then prove the rephrased result for both principal components and principal curves.

Let $d(z, f)$ denote the usual euclidean distance from a point $z$ to its projection on the curve $f$:

$$d(z, f) \overset{\text{def}}{=} \|z - f(\lambda f(z))\|$$  \hspace{1cm} (4)
and define the function $D^2 : \mathcal{G} \rightarrow \mathbb{R}^1$ by

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

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 any curves and we form $f_\epsilon = f + \epsilon g$ (see figure 4), then:

**Definition**

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

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

This means that they are minima, maxima or saddle points of this distance function.

**Theorem 1**

Suppose $f(\lambda) = \mathbb{E} X + \lambda v_0$ with $\|v_0\| = 1$, and $g(\lambda) = \lambda v$, $\|v\| = 1$. Thus $\mathcal{G} = \mathcal{L}$, 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 $\Sigma = \text{COV}(X)$.

If we restrict $f$ and $g$ to be members of the subset $\mathcal{G}$ of curves defined on a compact $\Lambda$, then principal curves have this property as well. In this case $f_\epsilon$ describes a class of curves about $f$ that shrink in as $\epsilon$ gets small. The corresponding result is:
Theorem 2

Let \( \mathcal{G} \) be the class of unit speed differentiable curves defined on a compact interval. The curve \( f \) is a principal curve of \( h \) iff \( f \) is a critical point of the distance function in the class \( \mathcal{G} \).

This is a key property and is an essential link to all the previous models and motivation in section 2. This property is similar to that enjoyed by conditional expectations or projections; the residual distance is minimized. Figure 5 illustrates the idea, and in fact is almost a proof in one direction.

![Figure 5](image)

Figure 5 The conditional expectation curve gets at least as close to the points as the original curve.

Suppose \( k \) is not a principal curve. Then the curve defined by \( f(\lambda) = \mathbb{E}(X | \lambda_k(X) = \lambda) \) certainly gets as close or 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 \( \lambda_k \) 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.

Both the above theorems will be proved in the appendix. 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 straight line fits.

2.3. An algorithm for finding principal curves.

We are still in the situation of finding principal curves for a probability distribution. When seeking principal curves 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

\[ D^2(f) = E \left\| X - f(\lambda_f(X)) \right\|^2 \]

\[ = E_{\lambda_f(X)} E \left[ \left\| X - f(\lambda_f(X)) \right\|^2 | \lambda_f(X) \right]. \quad (5) \]

We can write this as a minimization problem in \( f \) and \( \lambda \): find \( f \) and \( \lambda(\cdot) \) such that

\[ D^2_1(f, \lambda) = E \| X - f(\lambda) \|^2 \]  

is a minimum. Clearly, given any candidate solution \( f \) and \( \lambda \), \( f \) and \( \lambda_f \) is at least as good. Two key ideas emerge from this:

- If we know \( f \) as a function of \( \lambda \), then we can minimize (7) by picking \( \lambda(x) = \lambda_f(x) \) at each point \( x \) in the support of \( h \).
- If we know \( \lambda(x) \), then the distance

\[ D^2_2(f, \lambda) = E_{\lambda(X)} \sum_{j=1}^{p} E[(X_j - f_j(\lambda(X)))^2 | \lambda(X)] \]  

is minimized by the set of \( p \) conditional expectations

\[ f_j(\lambda) = E(X_j | \lambda(X) = \lambda). \]  

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 (9) will define a smooth function. The above preamble motivates the following iterative algorithm.

**Principal curve algorithm**

**Initialization:** Set \( f^{(0)}(\lambda) = a\lambda \) where \( a \) is the first linear principal component of \( h \). Set \( \lambda^{(0)} = \lambda^{(0)}_f \).

**Repeat:** over iteration counter \( j \)

1) Set \( f^{(j)}(\cdot) = E(X | \lambda^{(j-1)}(X) = \cdot). \)
2) Choose \( \lambda^{(j)} = \lambda^{(j)}_f. \)
3) Evaluate \( D^2(\cdot) = D^2_1(f^{(j)}, \lambda^{(j)}). \)

**Until:** \( D^2(\cdot) \) fails to decrease.
In specific problems, there might exist starting curves more natural than the principal components. For example, it might be known that the data arises from a circle, in which case this would be more appropriate as a prior curve.

It is easy to check that the criterion $D^2(i)$ must converge. It is positive and bounded below by 0. Suppose we have $f^{(i-1)}$ and $\lambda^{(i-1)}$. Now $D^2(f^{(i)}, \lambda^{(i-1)}) \leq D^2(f^{(i-1)}, \lambda^{(i-1)})$ by the properties of conditional expectation. Also $D^2(f^{(i)}, \lambda^{(i)}) \leq D^2(f^{(i)}, \lambda^{(i-1)})$ since the $\lambda^{(i)}$ 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 curves is suggestive of the above algorithm. We want a smooth curve that is self consistent. So we start with the line. We then check if it is indeed self consistent by evaluating the conditional expectation. If not we have a new curve 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 curve. By theorem 2 above, this curve is a critical point of the distance function.

2.4. Principal curves for data sets.

So far we have considered the principal curves of a continuous multivariate probability distribution. In reality, we usually have a finite multivariate data set. How do we define the principal curves 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.

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 $\lambda_f$ and $f^{(0)}$. If some prior starting curve is more appropriate, it can be substituted at this point.

2.4.1 The projection step.

Given functions $f^{(i-1)}$ we wish to find for each $z_i$ in the sample a value $\lambda_i^{(i-1)} = \lambda_f^{(i-1)}(z_i)$. We have $f^{(i-1)}$ evaluated at $n$ values of $\lambda$, in fact at $\lambda_1^{(i-2)}$, $\lambda_2^{(i-2)}$, $\ldots$, $\lambda_n^{(i-2)}$; $f^{(i-1)}$ can be evaluated at other points by interpolation. For each point $i$ in the sample we can project $z_i$ onto the line joining each pair $(\lambda_k^{(i-2)}, f^{(i-1)}(\lambda_k^{(i-2)}))$. 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 $\lambda_{ik} \in [\lambda_k^{(i-2)}, \lambda_{k+1}^{(i-2)}]$. We then let $\lambda_i^{(i-1)}$ be the $\lambda_{ik}$ that corresponds to the smallest value of $d_{ik}$. This procedure is computationally expensive, using $O(n)$ operations per projection.

A more efficient approach is to use the k-d tree (Friedman, Bently and Finkel, 1976). This uses a binary tree data structure to find nearest neighbors in $k$ dimensions. 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 $O(\log n)$.
At each of the \( n \) values of \( \hat{\lambda} \) we have a fitted \( p \) vector. We can build a \( p \)-d tree, and for each \( z_i \), find its nearest neighbor amongst these fitted values. We then project \( z_i \) into the segments on either side of this closest point to refine the estimate.

2.4.2 The conditional expectation step.

We have to estimate \( f^{(j)}(\lambda) = \mathbb{E}(X \mid \lambda^{(j-1)} = \lambda) \). We restrict ourselves to estimating this quantity at only \( n \) values of \( \lambda^{(j-1)} \), namely \( \lambda_1^{(j-1)}, \ldots, \lambda_n^{(j-1)} \) which we have already estimated. In order to estimate \( \mathbb{E}(X \mid \lambda^{(j-1)} = \hat{\lambda}_i^{(j-1)}) \), 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 \( z_i \). 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 \( z_k \) 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 6 demonstrates this local averaging.

![Diagram of conditional expectation estimation](image)

**Figure 6** We estimate the conditional expectation \( \mathbb{E}(X \mid \lambda^{(j-1)} = \hat{\lambda}_i^{(j-1)}) \) by averaging the observations \( z_k \) for which \( \hat{\lambda}_k^{(j-1)} \) is close to \( \hat{\lambda}_i^{(j-1)} \).

These ideas of local averaging are not new. In the more usual regression context, scatterplot smoothers are used to estimate the regression function \( \mathbb{E}(Y \mid z) \) by some kind of local averaging. Some commonly used smoothers are the kernel smoother (e.g. Gasser and Muller, 1979), spline smoothers (Wahba and Wold, 1975), and more recently the fast smoother technology of Friedman and Stuetzle.
These all smooth a one dimensional response against a covariate. In our case, the variable to be smoothed is \( p \) dimensional; so we simply smooth each coordinate separately. We discuss the choice of the smoother bandwidth in section 4.

Finally, we estimate \( D^2(t) \) in the obvious way, by adding up the distances of each point in the sample from the current curve.

One special version of this algorithm is obtained by replacing the smoother by the least squares straight line smoother. In this case we have:

**Theorem 3**

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 in appendix).

### 2.5. A demonstration of the Principal Curve procedure.

As an illustration, we generate data from an underlying model. The series of plots in figure 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

\[
\begin{pmatrix}
  x_1 \\
  x_2
\end{pmatrix} = \begin{pmatrix}
  5 \sin(\lambda) \\
  5 \cos(\lambda)
\end{pmatrix} + \begin{pmatrix}
  e_1 \\
  e_2
\end{pmatrix}
\]

(10)

where \( \lambda \) 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 7a. Since this is nearly parallel to the horizontal axis, the 2nd iteration in figure 7b will be close to the usual scatterplot smooth. This is clearly an inappropriate summary for this constructed data set.

The curve in figure 7g does substantially better than the previous iterations. This is because the span of the smoother is reduced. This means that the size of the neighborhood used for local averaging is smaller.

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 \( \sqrt{\text{EX}_t^2} = 1 \). We will see in section 4 that for this situation, the true circle does not 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. This is due to the overfit factor — for details, see Hastie (1984).

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 8a. The new curve is a function of \( \hat{\lambda}^{(0)} \).
Figure 7a  The solid curve is the principal component line. $D(\hat{j}^{(0)}) = 3.43$

Figure 7b  $D(\hat{j}^{(1)}) = 3.34$

Figure 7c  $D(\hat{j}^{(2)}) = 3.03$

Figure 7d  $D(\hat{j}^{(3)}) = 2.64$
Figure 7e  $D(\hat{j}^{(4)}) = 2.37$

Figure 7f  $D(\hat{j}^{(6)}) = 1.91$

Figure 7g  $D(\hat{j}^{(8)}) = 1.60$

Figure 7h  $D(\hat{j}^{(10)}) = 0.96$
measured along this vector as in figure 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 as a function of the initial principal component.

![Diagram](image)

Figure 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. The Principal Surfaces Model.

3.1. Definition of Principal Surfaces.

The definitions and results outlined in the previous sections generalize naturally to higher dimensional surfaces. We discuss briefly the generalization to 2-dimensional surfaces, and refer the reader to Hastie (1984) for more details.

A smooth 2-dimensional globally parametrized surface in $\mathbb{R}^p$ is a function $f : A \mapsto \mathbb{R}^p$ for $A \subset \mathbb{R}^2$.
where \( f \) is a vector of smooth functions:

\[
f(\lambda) = \begin{pmatrix}
    f_1(\lambda_1, \lambda_2) \\
    f_2(\lambda_1, \lambda_2) \\
    \vdots \\
    f_p(\lambda_1, \lambda_2)
\end{pmatrix}
\]

(11)

Figure 9 Each point on a principal surface is the average of the points that project there.

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 \( \mathbb{R}^p \), parametrized by \( \lambda \in \Lambda_f \), a 2-dimensional parameter vector.

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

\[
\lambda_f(x) = \max_{\lambda_2} \max_{\lambda_1} \{ \lambda : \| x - f(\lambda) \| = \inf_{\mu} \| x - f(\mu) \| \}.
\]

(12)

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:

\[
E(X \mid \lambda_f(X) = \lambda) = f(\lambda) \quad \text{for a.e. } \lambda \in \Lambda_f.
\]

Figure 9 demonstrates the situation.

3.2. Estimation of Principal Surfaces.

The algorithms for principal surfaces are very similar to those for curves. As a starting surface we use some smooth prior surface such as the plane spanned by the first two principal components. The algorithm then iterates over the following two steps:

1.) The Projection Step. The points are projected onto the surface of the previous iteration. This amounts to finding the parameter vector \( \lambda \) that corresponds to the closest point on the surface to the point in question. For finite data sets we use the \( p - d \) tree to find the closest fitted surface point to given point.

2.) The Conditional Expectation Step.

Here we find the conditional expectation of the random vector \( X \), given its point of projection on the surface of the previous iteration. We are conditioning on a 2-dimensional random vector \( \lambda \). The data analogue amounts to using a two dimensional smoother. We used a home brew locally weighted local planar smoother, which is a natural analogue of the Cleveland (1979) smoother. The word local here suggests finding nearest neighbors in the 2-dimensional space of the parameter \( \lambda \). We once again used the 2-d tree for this purpose (see Hastie, 1984 for details).

3.3. A Demonstration of the Principal Surface Procedure.

The data is artificially constructed, with no noise, by generating points uniformly on 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 section 5).

There are 61 data points, as shown in figure 10a. One point is placed at each intersection of 5 equally spaced parallels and 12 equally spaced meridians. The extra point is placed at the north pole. (If we placed a point at the south pole the solution would never move from the starting plane, which would, in fact, be a principal surface.) Figures 10b to 10d show various stages in the iterative procedure, and figure 10e shows another view of the final surface. Figure 10f is a parameter map of the two dimensional \( \lambda \). 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
**Figure 10a** The data points are placed in a uniform pattern on the surface of a sphere. The south pole is missing.

**Figure 10b** The second iteration of the principal surface procedure finds a surface that is a function of the first iteration.

**Figure 10c** An intermediate stage in the iterations.

**Figure 10d** The final surface produced by the principal surface routine.
Figure 10e  Another view of the final principal surface.

Figure 10f  The λ map is a two dimensional summary of the data. It resembles a stereographic map of the world.

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.

Ideally we would display this example on a motion graphics workstation in order to see the 3 dimensions.*


4.1. Model 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 to the 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.

4.1.1 A simple model for investigating bias.

The scenario we shall consider is a circle in 2-space. This can be parametrized by a unit speed curve

* The authors have prepared a motion graphics movie called Principal Curves and Surfaces. The ball is one of 4 examples demonstrated in the movie.
Figure 11 The data is generated from the arc of a circle with radius $\rho$ and with iid $\mathcal{N}(0, \sigma^2 I)$ errors. The location on the circle is selected uniformly.

$f(\lambda)$ with constant curvature $1/\rho$, where $\rho$ is the radius of the circle:

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

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

The points $z$ are generated as follows: First a $\lambda$ is selected uniformly from $\Lambda_\pi$. Given this value of $\lambda$ we pick the point $z$ from some smooth symmetric distribution with first two moments $(f(\lambda), \sigma^2 I)$ where $\sigma$ 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 $\rho$. * Consider the points that project onto a small arc $\Lambda_\theta$ of the circle (see figure 11). 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.

* I thank Art Owen for suggesting this result.
The following result is proved in Hastie (1984):

\[ E(x \mid \lambda_f(x) \in \Lambda_{\theta}) = \begin{pmatrix} r^*_{\theta} \\ 0 \end{pmatrix} \]

where

\[ r^*_{\theta} = r^* \frac{\sin(\theta/2)}{\theta/2}, \]

and

\[ r^* = E \sqrt{(\rho + e_1)^2 + e_2^2} \]

\[ \approx \rho + \frac{\sigma^2}{2\rho} . \]

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

Although the situation is artificially constructed, it separates nicely the two components of bias. One could argue that locally any curve looks like the arc of a circle; this argument would suffice if the errors where shrinking in some asymptotic sense, such as if each point was an average. Efron (1984) proves a result similar to (14) where this is in fact the case. The arc-angle \( \theta \) here represents the span of the smoother. There is clearly an optimal span at which the two bias components cancel exactly. In practice this is not much help since we require knowledge of the radius of curvature and the error variance to determine it. Typically these will be changing as we move along the curve.

4.1.2 A less simple bias demonstration.

This example demonstrates the bias components in a situation where the curvature is not constant, namely the sine wave in \( \mathbb{R}^2 \).

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 \( X = f(\lambda) + \epsilon \) where \( \lambda \sim U[0, 2] \) and \( \epsilon \sim N(0, I/4) \) independent of \( \lambda \). Figure 12 shows 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 \( E(X \mid \lambda_f(X) = \lambda) \). In this situation 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 \( \sin(\lambda \pi)/0.88 \). There are two dashed curves in the figure. They represent \( E(X \mid \lambda_f(X) \in \Lambda_s(\lambda)) \), where \( \Lambda_s(\lambda) \) represents a symmetric interval or span of length \( s \Lambda \) about \( \lambda \) (Boundary effects were eliminated by cyclically extending the range of \( \lambda \)). 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.
Figure 12  The thick curve is 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%.

4.2. A variance decomposition for principal curves.

Usually linear principal components are approached via variance considerations. The first component is that linear combination of the variables with the largest variance. If $\lambda = (\lambda_1, \ldots, \lambda_n)'$ are the projections of $z_i$ onto the first principal component $a$ of $X$, a centered $n \times p$ data matrix, then the following variance decomposition is easily derived:

$$\sum_{j=1}^{p} \text{Var}(z_j) = \text{Var}(\lambda) + E \|z - a\lambda\|^2$$  (15)

where $\text{Var}(\cdot)$ 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 $\text{Var}(\cdot)$ and $E(\cdot)$ have their usual meaning. The second term on the right of (15) 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:

\[
\sum_{j=1}^{p} \text{Var}(x_j) = \sum_{j=1}^{p} \mathbb{E}(x_j - \mathbb{E}(x_j \mid w))^2 + \sum_{j=1}^{p} \text{Var}(\mathbb{E}(x_j \mid w)).
\]  

(16)

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

\[
\sum_{j=1}^{p} \text{Var}(x_j) = \mathbb{E}\left\| x - f(\lambda_f(x)) \right\|^2 + \sum_{j=1}^{p} \text{Var}(f_j(\lambda_f(x))).
\]  

(17)

This gives us an analogous result to (15) 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 (17) holds only approximately:

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

(18)

The reason for this is that most practical scatterplot smoothers are not projections, whereas conditional expectations are. The usual orthogonality between the fit and the residuals no longer holds.

We make the following observations:

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

\[
\sum_{j=1}^{p} \text{Var}(f_j(\lambda_f(x))) = \sum_{j=1}^{p} a_j^2 \text{Var}(\lambda_a(x)) = \text{Var}(\lambda)
\]

since \( a \) has length 1. Here we have written \( \lambda \) for the function \( \lambda_a(z) = a'z \).

- if the \( f_j \) are approximately linear we can use the Delta method to obtain

\[
\sum_{j=1}^{p} \text{Var}(f_j(\lambda_f(x))) \approx \sum_{j=1}^{p} (f_j'_{\lambda}(\lambda_f(x)))^2 \text{Var}(\lambda_f(x)) = \text{Var}(\lambda_f(x))
\]

if \( f \) is unit speed, in which case \( \|f'\| = 1 \).
4.3. Cross-validation and Span selection.

4.3.1 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 $0.5n$ observations in each neighborhood, 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.

4.3.2 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 non-parametric regression as well. In regression we can use cross-validation (Stone 1977) to select the span.

This amounts to choosing a span to minimize an estimate of the expected prediction error, if we were to use the estimated regression to predict future observations. This is achieved by predicting each observation in the sample with a smooth estimated from the sample with that observation omitted.

One thought in the case of Principal curves 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 coordinates separately. This expresses the belief that each estimated coordinate function is an estimate for a true underlying function. As such it will have bias and variance, and we trade them off. For details and a formal motivation, see Hastie (1984).

5. Examples.

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

5.1. Example: Gold assay pairs.

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
Figure 13a  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 13b  Estimated coordinate functions. The dashed curve is the outside lab, the solid curve the inhouse lab.

is in their interest to know which laboratory produces on average lower gold content assays for a given sample.

The data in figure 13a consists of 250 pairs of gold assays. Each point is represented by the vector $\mathbf{z}_i$ where $z_{ji} = \log(1 + \text{assay yield for } i\text{th assay pair for lab } j)$ 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)).

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 section 6.

The model presented here for the above data is

\[
\begin{pmatrix}
    x_{1i} \\
    x_{2i}
\end{pmatrix}
= 
\begin{pmatrix}
    f_1(r_i) \\
    f_2(r_i)
\end{pmatrix}
+ 
\begin{pmatrix}
    e_{1i} \\
    e_{2i}
\end{pmatrix}
\]  

(19)
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:

\[
\begin{pmatrix}
x_{1i} \\
x_{2i}
\end{pmatrix} = \begin{pmatrix}
\alpha + \beta z_i \\
z_i
\end{pmatrix} + \begin{pmatrix}
e_{1i} \\
e_{2i}
\end{pmatrix}
\]

(20)

where \( f_2(r_i) = z_i \) and

\[
f_1(r_i) = f_1 \circ f_2^{-1}(z_i) \quad \text{(assuming \( f_2 \) is monotone)}
\]

\[
= \alpha + \beta z_i
\]

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 *

\[
\text{Var}(e_{1i}) = \text{Var}(e_{2i}) \quad \forall \ i.
\]

The model is estimated using the principal curve estimate for the data and is represented by the solid curve in figure 13a. 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 to 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 13b.

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 \( z_i \). Each such sample, which samples the points \( z_i \) with replacement, is called a bootstrap sample.

Figure 13c 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.

* In the linear model one usually requires that \( \text{Var}(e_{ji}) = \text{constant}_j \). This assumption can be relaxed here.
Figure 13c 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.

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 \( \hat{\tau}_i \), the true gold level for sample \( i \). However, \( \hat{\tau}_i \) need not be the true gold level at all. It may be any variable that orders the pairs \( f(\hat{\tau}_i) \) 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 \( \tau_i \) do provide us with a good summary variable for each of the pairs, if that is required:

\[
\hat{\tau}_i = h(x_i)
\]

since we obtain \( \hat{\tau}_i \) 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. to tackle the problem using standard regression techniques.
5.2. Example: Geological data.

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 $X_1, \ldots, X_{10}$, and analyze the first three.

![Figure 14a](image)

**Figure 14a** 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 14a 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 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 observations 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 inversely proportional to their distance from the curve. In the smoothing step, a weighted smooth is used, and if the weight is below a certain threshold, it is set to 0. Four points were identified as outliers, and are labelled differently in figure 14a. We would
really consider them model outliers, since in that region of the curve the model does not appear to fit very well.

Figure 14b 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 coordinate plots would then represent the mean level of the particular mineral at different depths (see figure 14c). Usually one would have to use these coordinate plots to identify the factors, just as one uses the factor loadings in the linear case.

5.3. Example: Lipoprotein data.

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 \( S_f 3 - 4 \)), LDL 7-8, and HDL 3 (High Density Lipoprotein) in the sample of 81 men. Figures 15a,b are different views of the principal surface found for the data. Quantitatively this surface explains 97.4% of the variability in the data, and accounts for 80% of the residual variance unexplained by the principal
Figure 14c  The estimated coordinate functions or factor loading curves for the three minerals.

component plane. Qualitatively, we see that the surface has interesting structure in only two of the coordinates, 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 coordinate plot (figure 15c) of the estimated coordinate function for the variables LDL 7-8 which helps confirm this claim. The relationship between LDL 7-8 and $(\lambda_1, \lambda_2)$ depends mainly on the level of $\lambda_1$. Similar information is conveyed by the other coordinate plots, or can be seen from the estimated surface directly. This suggests a model of the form

$$
\begin{pmatrix}
\text{LDL 3-4} \\
\text{LDL 7-8} \\
\text{HDL 3}
\end{pmatrix} = 
\begin{pmatrix}
f_1(\lambda_1) \\
f_2(\lambda_1) \\
f_3(\lambda_2)
\end{pmatrix} + 
\begin{pmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\varepsilon_3
\end{pmatrix}.
$$

As specified $\lambda_2$ 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 15d 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 the next section. Their results are qualitatively the same as ours for the LDL pair.
**Figure 15a** 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 15b** The principal surface as in figure 15a from a slightly oblique perspective.

**Figure 15c** The estimated coordinate function for LDL 7-8 versus $\lambda$. $\lambda_2$ has little effect.

**Figure 15d** The principal curve for the serum concentrations LDL 7-8 and LDL 3-4.
6. Discussion and conclusions.

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 introduced by transforming the space or by transforming the model.

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


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 coordinates. One then seeks a subspace of this non-linear coordinate 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 (1969), and a good description can be found in Gnanadesikan (1977). They suggested using polynomial functions of the original $p$ coordinates. The resulting linear combinations are then of the form (for $p = 2$ and quadratic polynomials)

$$
\lambda_j = a_{1j} x_1 + a_{2j} x_2 + a_{3j} x_1 x_2 + a_{4j} x_1^2 + a_{5j} x_2^2
$$

(21)

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 (21) 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 coordinates.

The model has been generalized further to include more general transformations of the coordinates 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 coordinates. The problem can be formulated as follows: Find $a$ and $s'(z) = (s_1(x_1), \ldots, s_p(x_p))$ such that

$$
E \| s(z) - ad'(z) \|^2 = \min
$$

(22)

or alternatively such that

$$
\text{Var}[a'd'(z)] = \max
$$

(23)

where $E s_j(z_j) = 0$, $a'a = 1$ and $E s_j^2(z_j) = 1$. The idea is to transform the coordinates suitably and then find the linear principal components. If in (23) we replaced $\max$ by $\min$ then we would be estimating the constraint in the transformed space. Hastie (1984) outlines a procedure for estimating the functions. 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

\[
\begin{pmatrix}
\hat{s}_1(z_1) \\
\vdots \\
\hat{s}_p(z_p)
\end{pmatrix} = a a' s(z).
\]

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

\[
\begin{pmatrix}
\hat{x}_1 \\
\vdots \\
\hat{x}_p
\end{pmatrix} = \begin{pmatrix}
s_1^{-1}(a_1 z) \\
\vdots \\
s_p^{-1}(a_p z)
\end{pmatrix}, \tag{24}
\]

where \( z = a' s(z) \). Equation (24) defines a parametrized curve. The curve is not completely general since the coordinate functions are monotone. For the same reason, Gnanadesikan (1977) expressed the desirability of having procedures for estimating models of the type proposed in this paper.

• We are estimating manifolds that are close to the data in the transformed coordinates. 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 \( p \) coordinates. Of course, if the functions are linear, both approaches are identical.

An advantage of the technique is that it can easily be generalized 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 \( AA' s(z) \). 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.


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). Gnanadesikan (1977) gives a concise description. 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 \( m \) coordinates 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 our uniform ball example. They do not, however, provide a model for dissimilarities which was the original intention of multidimensional scaling.

6.2.1 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 example. 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.

6.2.2 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 parametric setting, although their emphasis is on the factor analysis model. We introduce self-consistent curves as summaries of the data. In certain situations, these summaries can be used as estimates of the appropriate non-linear functional and factor models.

6.2.3 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 \( x \).
• 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 4**

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

$$x^* = \frac{x + y}{\sqrt{2}},$$

$$y^* = \frac{x - y}{\sqrt{2}}. \quad (25)$$

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

$$\hat{x} = \frac{\text{Smooth}(x^* | x^*) + \text{Smooth}(y^* | x^*)}{\sqrt{2}},$$

$$\hat{y} = \frac{\text{Smooth}(x^* | x^*) - \text{Smooth}(y^* | x^*)}{\sqrt{2}}. \quad (26)$$

Since the smoother is linear, and in view of (25), (26) becomes

$$\hat{x} = \text{Smooth}(x | x^*),$$

$$\hat{y} = \text{Smooth}(y | x^*). \quad (27)$$

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

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

6.3. 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.

- 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 latter has been used successfully a number of times in practical data analysis 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.
Appendix

We supply the proofs omitted in the paper. For convenience the theorems are restated here.

**Theorem 1 (section 2.2)**

Suppose \( f(\lambda) = \mathbf{E}X + \lambda \mathbf{w}_0 \) with \( \|\mathbf{w}_0\| = 1 \), and \( g(\lambda) = \lambda \mathbf{v}, \|\mathbf{v}\| = 1 \). Thus \( \mathcal{G} = \mathcal{L} \), the class of all unit speed straight lines. Then \( f \) is a critical point of the distance function in \( \mathcal{L} \) iff \( \mathbf{w}_0 \) is an eigenvector of \( \Sigma = \text{cov}(X) \).

Note:

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

**Proof**

The closest point from \( \mathbf{z} \) to any line \( \lambda \mathbf{w} \) through the origin is found by projecting \( \mathbf{z} \) onto \( \mathbf{w} \) and has parameter value

\[
\lambda_{\mathbf{w}}(\mathbf{z}) = \frac{\mathbf{z}'\mathbf{w}}{\|\mathbf{w}\|}
\]

Then

\[
d^2(\mathbf{z}, \lambda \mathbf{w}) = \left\| \mathbf{z} - \frac{\mathbf{w}' \mathbf{z}}{\|\mathbf{w}\|^2} \mathbf{w} \right\|^2
\]

\[
= \|\mathbf{z}\|^2 - \frac{|\mathbf{w}' \mathbf{z}|^2}{\mathbf{w}' \mathbf{w}}
\]

Upon taking expected values we get

\[
D^2(h, \lambda \mathbf{w}) = \text{tr} \, \Sigma - \frac{\mathbf{w}' \Sigma \mathbf{w}}{\mathbf{w}' \mathbf{w}}.
\]

We now apply the above to \( f_\epsilon \) instead of \( \mathbf{w} \), but first make a simplifying assumption. We can assume w.l.o.g. that \( \mathbf{w}_0 = \mathbf{e}_1 \) since the problem is invariant to rotations.

We split \( \mathbf{v} \) into a component \( \mathbf{v}_\epsilon = \epsilon \mathbf{e}_1 \) along \( \mathbf{e}_1 \) and an orthogonal component \( \mathbf{v}^* \). Thus \( \mathbf{v} = \epsilon \mathbf{v}_\epsilon + \mathbf{v}^* \) where \( \mathbf{v}_\epsilon^* \mathbf{v}^* = 0 \). So \( \mathbf{f}_\epsilon = \lambda ((1 + \epsilon \epsilon) \mathbf{e}_1 + \epsilon \mathbf{v}^*) \). We now plug this into (28) to get

\[
D^2(h, \mathbf{f}_\epsilon) = \text{tr} \, \Sigma - \frac{((1 + \epsilon \epsilon) \mathbf{e}_1 + \epsilon \mathbf{v}^*)' \Sigma ((1 + \epsilon \epsilon) \mathbf{e}_1 + \epsilon \mathbf{v}^*)}{(1 + \epsilon \epsilon)^2 + \epsilon^2}
\]

\[
= \text{tr} \, \Sigma - \frac{(1 + \epsilon \epsilon)^2 \mathbf{e}_1' \Sigma \mathbf{e}_1 + 2 \epsilon (1 + \epsilon \epsilon) \mathbf{e}_1' \Sigma \mathbf{v}^* + \epsilon^2 \mathbf{v}^* \Sigma \mathbf{v}^*}{(1 + \epsilon \epsilon)^2 + \epsilon^2}
\]

Differentiating w.r.t. \( \epsilon \) and setting \( \epsilon = 0 \) we get

\[
\frac{dD^2(h, \mathbf{f}_\epsilon)}{d\epsilon} \bigg|_{\epsilon=0} = -2 \mathbf{e}_1' \Sigma \mathbf{v}^*.
\]
If $e_1$ is a principal component of $\Sigma$ 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 $v^*e_1 = 0$, we have

$$v^*\Sigma e_1 = 0 \quad \forall v^*e_1 = 0$$

$$\Rightarrow \Sigma e_1 = ce_1$$

$$\Rightarrow e_1 \text{ is an eigenvector of } \Sigma$$

Theorem 2 (section 2.2)

Let $G$ be the class of unit speed differentiable curves defined on $\Lambda$, 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 $G$.

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 (Hastie, 1984). In practice this is not a restriction.

Proof

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

$$\frac{d}{d\epsilon}D^2(h, f_\epsilon) = \frac{d}{d\epsilon} \mathbb{E}_h \left\| X - f_\epsilon(\lambda f_\epsilon(X)) \right\|^2.$$  \hspace{1cm} (30)

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

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

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
\[ Z_\epsilon \leq -2 \left( X - f(\lambda_f(X)) \right) \cdot g(\lambda_f(X)) + \epsilon \left\| g(\lambda_f(X)) \right\|^2 \leq Y_1 \]
where \( Y_1 \) is some bounded random variable.

Similarly we have
\[ Z_\epsilon \geq \frac{\left\| X - f_\epsilon(\lambda_{f_\epsilon}(X)) \right\|^2 - \left\| X - f(\lambda_f(X)) \right\|^2}{\epsilon} \]
We expand the first norm again, and get
\[ Z_\epsilon \geq -2 \left( X - f(\lambda_f(X)) \right) \cdot g(\lambda_f(X)) + \epsilon \left\| g(\lambda_f(X)) \right\|^2 \geq Y_2 \]
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 \( \lambda_f(X) \) is continuous in \( \epsilon \) at \( \epsilon = 0 \). Moreover, this limit is given by
\[ \lim_{\epsilon \to 0} Z_\epsilon = \frac{d}{d\epsilon} \left| \left.X - f(\lambda_{f_\epsilon}(X)) \right\|_\epsilon = 0 \right| \]
\[ = -2 \left( X - f(\lambda_f(X)) \right) \cdot g(\lambda_f(X)). \]

We show in lemma 2.1 that this continuity condition is met almost surely.

We denote the distribution function of \( \lambda_f(X) \) by \( h_\lambda \), and get
\[ \frac{d}{d\epsilon} D^2(h, f_\epsilon) \bigg|_{\epsilon = 0} = -2E_{h_\lambda} \left( E(X \mid \lambda_f(X) = \lambda) - f(\lambda) \right) \cdot g(\lambda). \tag{31} \]

If \( f(\lambda) \) is a principal curve of \( h \), then \( E(X \mid \lambda_f(X) = \lambda) = f(\lambda) \) for all \( \lambda \) in the support of \( h_\lambda \), and thus
\[ \frac{d}{d\epsilon} D^2(h, f_\epsilon) \bigg|_{\epsilon = 0} = 0 \quad \forall \text{ differentiable } g. \]

Alternatively, suppose that
\[ E_{h_\lambda} \left( E(X - f(\lambda) \mid \lambda_f(X) = \lambda) \cdot g(\lambda) \right) = 0 \tag{32} \]
for all differentiable \( g \). In particular we could pick \( g(\lambda) = E(X \mid \lambda_f(X) = \lambda) - f(\lambda) \). Then
\[ E_{\lambda} \left\| E(X \mid \lambda_f(X) = \lambda) - f(\lambda) \right\|^2 = 0 \]
and consequently \( f \) is a principal curve. This choice of \( g \), however, might not be differentiable, so some approximation is needed.
Since (32) holds for all differentiable $g$ we can use different $g$'s to knock off different pieces of $E(X \mid \lambda_f(X) = \lambda) = f(\lambda)$. In fact we can do it one coordinate at a time. For example, suppose $E(X_1 \mid \lambda_f(X) = \lambda)$ is positive for almost every $\lambda \in (\lambda_0, \lambda_1)$. We suggest why such an interval will always exist. It can be shown (Hastie, 1984) that $\lambda_f(z)$ is continuous at almost every $z$. The set \{ $X_1 \mid \lambda_f(X) = \lambda \in (\lambda_0, \lambda_1)$ \} is the set of $X$ which exist in an open connected set in the normal plane at $\lambda$, 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 $E(X_1 \mid \lambda_f(X) = \lambda) = f_1(\lambda)$ for $\lambda \in (\lambda_0, \lambda_1)$. We can do this for all such intervals, and for each coordinate, 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 2 is true for all $g$, in particular for $g(\lambda) = \lambda v$. We then invoke theorem 1.

In order to complete the proof of theorem 2, we need to prove the following

Lemma 2.1

The projection function $\lambda_{f_\epsilon}(z)$ is continuous at $\epsilon = 0$ for almost every $z$ 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 $z$, and no other points on $f$ are as close to $z$. Thus $\exists \lambda_0 > \lambda_1$, $\lambda_f(z) = \lambda_0$ and $\|z - f(\lambda_0)\| = \|z - f(\lambda_1)\|$. It is easy to pick $g$ in this situation such that $\lambda_{f_\epsilon}(z)$ is not continuous at $\epsilon = 0$. We call such points ambiguous. However, it can be shown (Hastie, 1984) 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 $z$ as $f(\lambda_f(z)) = f(\lambda_0)$. Thus $\|z - f(\lambda_0)\| < \|z - f(\lambda_1)\| \forall \lambda_1 \in [a, b] \cap (\lambda_0 - \omega, \lambda_0 + \omega)^\epsilon$. (Notice that at the boundaries the $\omega$ interval can be suitably redefined.) Since this interval is compact, and the distance functions are differentiable, we can find a $\delta > 0$ such that $\|z - f(\lambda_0)\| \leq \|z - f(\lambda_1)\| - \delta$. Let $M = \sup_{\lambda \in [a, b]} \|g(\lambda)\|$ and $\epsilon_0 = \delta/(2M)$. Then $\|z - f_\epsilon(\lambda_0)\| < \|z - f_\epsilon(\lambda_1)\| \forall \lambda_1 \in [a, b] \cap (\lambda_0 - \omega, \lambda_0 + \omega)^\epsilon$ and $\forall \epsilon \leq \epsilon_0$. This implies that $\lambda_{f_\epsilon}(z) \in (\lambda_0 - \omega, \lambda_0 + \omega)$, and the continuity is established.

Theorem 3 (section 2.4)

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 $\lambda^{(0)}_i$ by projecting $z_i$ onto $a^{(0)}$ which we denote collectively by

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

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

$$a^{(1)}_j = \frac{\lambda^{(0)'}z_j}{\lambda^{(0)'\lambda^{(0)}}}$$

or

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

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).

Acknowledgements
We thank Andreas Buja and Robert Tibshirani for their many helpful discussions and suggestions. Thanks also go to Brad Efron and Jerome Friedman for their insightful comments, and to Iain Johnstone for some theoretical support.
References


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


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


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


