Principal Curves

TREVOR HASTIE and WERNER STUETZLE*

Principal curves are smooth one-dimensional curves that pass through the *middle* of a p-dimensional data set, providing a nonlinear summary of the data. They are nonparametric, and their shape is suggested by the data. The algorithm for constructing principal curves starts with some prior summary, such as the usual principal-component line. The curve in each successive iteration is a *smooth* or local average of the p-dimensional points, where the definition of local is based on the distance in arc length of the projections of the points onto the curve found in the previous iteration. In this article principal curves are defined, an algorithm for their construction is given, some theoretical results are presented, and the procedure is compared to other generalizations of principal components. Two applications illustrate the use of principal curves. The first describes how the principal-curve procedure was used to align the magnets of the Stanford linear collider. The collider uses about 950 magnets in a roughly circular arrangement to bend electron and positron beams and bring them to collision. After construction, it was found that some of the magnets had ended up significantly out of place. As a result, the beams had to be bent too sharply and could not be focused. The engineers realized that the magnets did not have to be moved to their originally planned locations, but rather to a sufficiently smooth arc through the middle of the existing positions. This arc was found using the principal-curve procedure. In the second application, two different assays for gold content in several samples of computer-chip waste appear to show some systematic differences that are blurred by measurement error. The classical approach using linear errors in variables regression can detect systematic linear differences but is not able to account for nonlinearities. When the first linear principal component is replaced with a principal curve, a local "bump" is revealed, and bootstrapping is used to verify its presence.

KEY WORDS: Errors in variables; Principal components; Self-consistency; Smoother; Symmetric.

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 pattern exhibited by the points in the scatterplot. The type of summary we choose depends on the goal of our analysis; a trivial summary is the mean vector that simply locates the center of the cloud but conveys no information about the joint behavior 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. Hence the aim of the analysis is to seek a rule for predicting the 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 Fig. 1a).

In many situations we do not have a preferred variable that we wish to label "response," but would still like to summarize the joint behavior 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 with the reduced cost and increased speed of computing nonparametric scatterplot smoothers have gained popularity. These include kernel smoothers (Watson 1964), nearest-neighbor smoothers (Cleveland 1979), and spline smoothers (Silverman 1985). In general, scatterplot smoothers produce a curve that attempts to minimize the vertical deviations (as depicted in Fig. 1c), subject to some form of smoothness constraint. The nonparametric versions referred to before allow the data to dictate the form of the nonlinear dependency.

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 pass through the *middle* of the data in a smooth way, whether or not the middle of the data is a straight line. This situation is depicted in Figure 1d. These curves, like linear principal components, focus on the orthogonal or shortest distance to the points. 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 of the data that project onto this point, and average them, then this average coincides with the point on the curve.

The algorithm for finding principal curves is equally intuitive. Starting with any smooth curve (usually the largest principal component), it checks if this curve is self-consistent by projecting and averaging. If it is not, the procedure is repeated, using the new curve obtained by

* Trevor Hastie is Member of Technical Staff, AT&T Bell Laboratories, Murray Hill, NJ 07974. Werner Stuetzle is Associate Professor, Department of Statistics, University of Washington, Seattle, WA 98195. This work was developed for the most part at Stanford University, with partial support from U.S. Department of Energy Contracts DE-AC03-76SF and DE-AT03-81-ER10843, U.S. Office of Naval Research Contract N00014-81-K-0340, and U.S. Army Research Office Contract DAAG29-82-K-0056. The authors thank Andreas Buja, Tom Duchamp, Iain Johnstone, and Larry Shepp for their theoretical support, Robert Tibshirani, Brad Efron, and Jerry Friedman for many helpful discussions and suggestions, Horst Friedsam and Will Oren for supplying the Stanford linear collider example and their help with the analysis, and both referees for their constructive criticism of earlier drafts.

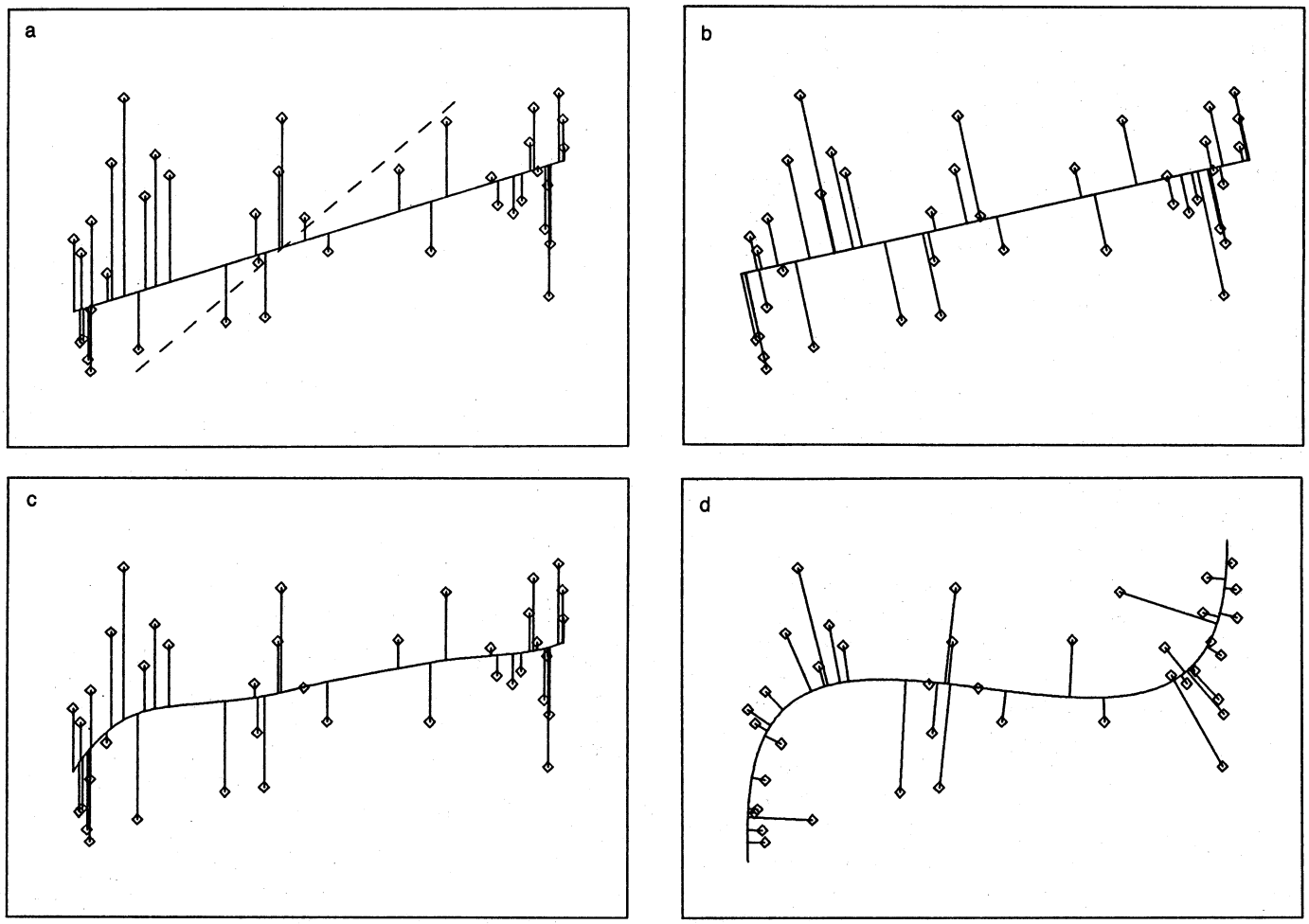

Figure 1. (a) The linear regression line minimizes the sum of squared deviations in the response variable. (b) The principal-component line minimizes the sum of squared deviations in all of the variables. (c) The smooth regression curve minimizes the sum of squared deviations in the response variable, subject to smoothness constraints. (d) The principal curve minimizes the sum of squared deviations in all of the variables, subject to smoothness constraints.

averaging as a starting guess. This is iterated until (hopefully) convergence.

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

1. In errors-in-variables regression it is assumed that there is randomness in the predictors as well as the response. This can occur in practice when the predictors 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. The errors-in-variables regression technique models the expectation of y as a linear function of the systematic component of x. In the case of a single predictor, the model is estimated by the principal-component line. This is also the total least squares method of Golub and van Loan (1979). More details are given in an example in Section 8.
2. Often we want to replace several highly correlated variables with 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.
3. In factor analysis we model the systematic component of the data by linear functions of a small set of unobservable variables called factors. Often the models are estimated using linear principal components; in the case of one factor [Eq. (1), as follows] one could use the largest principal component. Many variations of this model have appeared in the literature.

In all the previous situations the model can be written as

$$\mathbf{x}_i = \mathbf{u}_0 + \mathbf{a}\lambda_i + \mathbf{e}_i, \tag{1}$$

where $\mathbf{u}_0 + \mathbf{a}\lambda_i$ is the systematic component and $\mathbf{e}_i$ is the random component. If we assume that $\text{cov}(\mathbf{e}_i) = \sigma^2 I$, then the least squares estimate of $\mathbf{a}$ is the first linear principal component.

A natural generalization of (1) is the nonlinear model

$$\mathbf{x}_i = \mathbf{f}(\lambda_i) + \mathbf{e}_i. \tag{2}$$

This might then be a factor analysis or structural model, and for two variables and some restrictions an errors-in-variables regression model. In the same spirit as before, where we used the first linear principal component to estimate (1), the techniques described in this article can be used to estimate the systematic component in (2).

We focus on the definition of principal curves and an

algorithm for finding them. We also present some theoretical results, although many open questions remain.

2. THE PRINCIPAL CURVES OF A PROBABILITY DISTRIBUTION

We first give a brief introduction to one-dimensional curves, and then define the principal curves of smooth probability distributions in p space. Subsequent sections give algorithms for finding the curves, both for distributions and finite realizations. This is analogous to motivating a scatterplot smoother, such as a moving average or kernel smoother, as an estimator for the conditional expectation of the underlying distribution. We also briefly discuss an alternative approach via regularization using smoothing splines.

2.1 One-Dimensional Curves

A one-dimensional curve in p-dimensional space is a vector $\mathbf{f}(\lambda)$ of p functions of a single variable λ. These functions are called the coordinate functions, and λ provides an ordering along the curve. If the coordinate functions are smooth, then $\mathbf{f}$ is by definition a smooth curve. We can apply any monotone transformation to λ, and by modifying the coordinate functions appropriately the curve remains unchanged. The parameterization, however, is different. There is a natural parameterization for curves in terms of the arc length. The arc length of a curve $\mathbf{f}$ from λ_0 to λ_1 is given by $l = \int_{\lambda_0}^{\lambda_1} \|\mathbf{f}'(z)\| dz$. If $\|\mathbf{f}'(z)\| \equiv 1$, then $l = \lambda_1 - \lambda_0$. This is a desirable situation, since if all of the coordinate variables are in the same units of measurement, then λ is also in those units.

The vector $\mathbf{f}'(\lambda)$ is tangent to the curve at λ and is sometimes called the velocity vector at λ. A curve with $\|\mathbf{f}'\| \equiv 1$ is called a unit-speed parameterized curve. We can always reparameterize any smooth curve with $\|\mathbf{f}'\| > 0$ to make it unit speed. In addition, our intuitive concept of smoothness relates more naturally to unit-speed curves. For a unit-speed curve, smoothness of the coordinate functions translates directly into smooth visual appearance of the point set $\{\mathbf{f}(\lambda), \lambda \in \Lambda\}$ (absence of sharp bends). If $\mathbf{v}$ is a unit vector, then $\mathbf{f}(\lambda) = \mathbf{v}_0 + \lambda\mathbf{v}$ is a unit-speed straight line. This parameterization is not unique: $l^*(\lambda) = \mathbf{u} + a\mathbf{v} + \lambda\mathbf{v}$ is another unit-speed parameterization for the same line. In the following we always assume that $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.

The vector $\mathbf{f}''(\lambda)$ 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 $\mathbf{f}''/\|\mathbf{f}''\|$ is called the principal normal to the curve at λ. The vectors $\mathbf{f}'(\lambda)$ and $\mathbf{f}''(\lambda)$ span a plane. There is a unique unit-speed circle in the plane that goes through $\mathbf{f}(\lambda)$ and has the same velocity and acceleration at $\mathbf{f}(\lambda)$ as the curve itself (see Fig. 2). The radius $r_{\mathbf{f}}(\lambda)$ of this circle is called the radius of curvature of the curve $\mathbf{f}$ at λ; it is easy to see that $r_{\mathbf{f}}(\lambda) = 1/\|\mathbf{f}''(\lambda)\|$. The center $\mathbf{c}_{\mathbf{f}}(\lambda)$ of the circle is called the center of curvature of $\mathbf{f}$ at λ. Thorpe (1979) gave a clear introduction to these and related ideas in differential geometry.

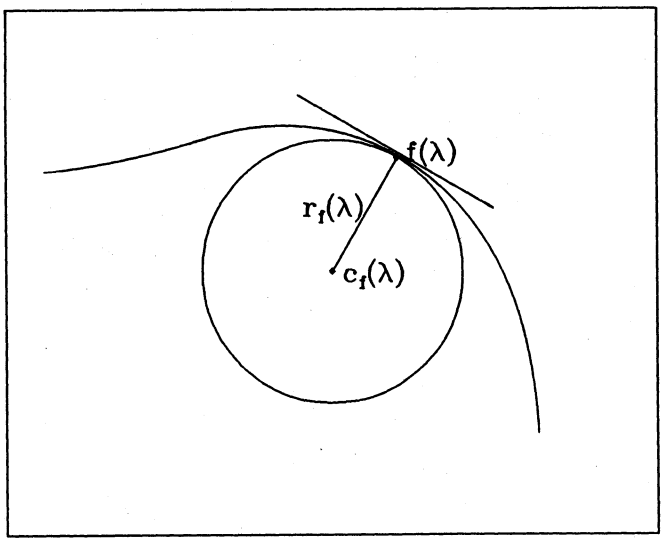

Figure 2. The radius of curvature is the radius of the circle tangent to the curve with the same acceleration as the curve.

2.2 Definition of Principal Curves

Denote by $\mathbf{X}$ a random vector in $\mathbf{R}^p$ with density h and finite second moments. Without loss of generality, assume $E(\mathbf{X}) = \mathbf{0}$. Let $\mathbf{f}$ denote a smooth (C^∞) unit-speed curve in $\mathbf{R}^p$ parameterized over $\Lambda \subseteq \mathbf{R}^1$, a closed (possibly infinite) interval, that does not intersect itself ($\lambda_1 \neq \lambda_2 \Rightarrow \mathbf{f}(\lambda_1) \neq \mathbf{f}(\lambda_2)$) and has finite length inside any finite ball in $\mathbf{R}^p$.

We define the projection index $\lambda_{\mathbf{f}}: \mathbf{R}^p \to \mathbf{R}^1$ as

$$\lambda_{\mathbf{f}}(\mathbf{x}) = \sup_{\lambda}\{\lambda : \|\mathbf{x} - \mathbf{f}(\lambda)\| = \inf_{\mu}\|\mathbf{x} - \mathbf{f}(\mu)\|\}. \quad (3)$$

The projection index $\lambda_{\mathbf{f}}(\mathbf{x})$ of $\mathbf{x}$ is the value of λ for which $\mathbf{f}(\lambda)$ is closest to $\mathbf{x}$. If there are several such values, we pick the largest one. We show in the Appendix that $\lambda_{\mathbf{f}}(\mathbf{x})$ is well defined and measurable.

Definition 1. The curve $\mathbf{f}$ is called self-consistent or a principal curve of h if $E(\mathbf{X} \mid \lambda_{\mathbf{f}}(\mathbf{X}) = \lambda) = \mathbf{f}(\lambda)$ for a.e. λ.

Figure 3 illustrates the intuitive motivation behind our definition of a principal curve. For any particular parameter value λ we collect all of the observations that have $\mathbf{f}(\lambda)$ as their closest point on the curve. If $\mathbf{f}(\lambda)$ is the average of those observations, and if this holds for all λ, then $\mathbf{f}$ is called a principal curve. In the figure we have actually averaged observations projecting into a neighborhood on the curve. This gives the flavor of our data algorithms to come; we need to do some kind of local averaging to estimate conditional expectations.

The definition of principal curves immediately gives rise to several interesting questions: For what kinds of distributions do principal curves exist, how many different principal curves are there for a given distribution, and what are their properties? We are unable to answer those questions in general. We can, however, show that the definition is not vacuous, and that there are densities that do have principal curves.

It is easy to check that for ellipsoidal distributions the

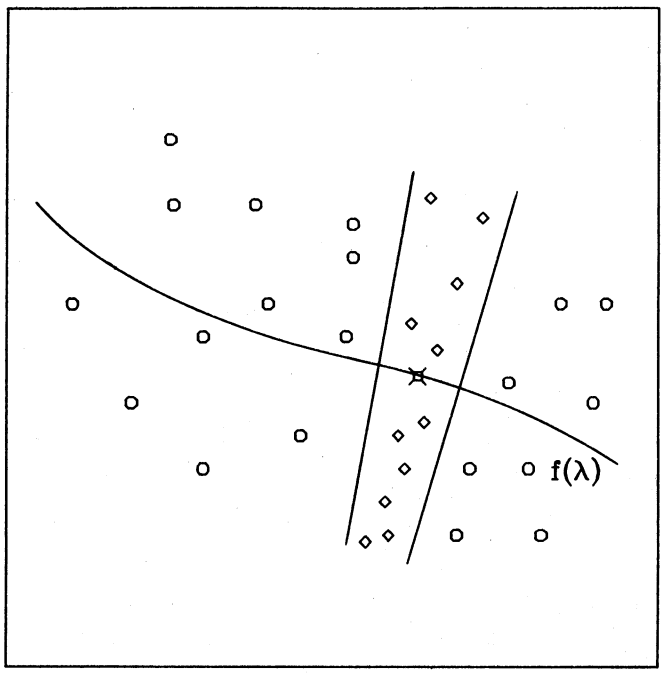

Figure 3. Each point on a principal curve is the average of the points that project there.

principal components are principal curves. For a spherically symmetric distribution, any line through the mean vector is a principal curve. For any two-dimensional spherically symmetric distribution, a circle with the center at the origin and radius $E\|\mathbf{X}\|$ is a principal curve. (Strictly speaking, a circle does not fit our definition, because it does intersect itself. Nevertheless, see our note at the beginning of the Appendix, and Sec. 5.6, for more details.)

We show in the Appendix that for compact Λ it is always possible to construct densities with the carrier in a thin tube around $\mathbf{f}$, which have $\mathbf{f}$ as a principal curve.

What about data generated from the model $\mathbf{X} = \mathbf{f}(\lambda) + \boldsymbol{\epsilon}$, with $\mathbf{f}$ smooth and $E(\boldsymbol{\epsilon}) = 0$? Is $\mathbf{f}$ a principal curve for this distribution? The answer generally seems to be no. We show in Section 7 in the more restrictive setting of data scattered around the arc of a circle that the mean of the conditional distribution of $\mathbf{x}$, given $\lambda(\mathbf{x}) = \lambda_0$, lies outside the circle of curvature at λ_0; this implies that $\mathbf{f}$ cannot be a principal curve. So in this situation the principal curve is biased for the functional model. We have some evidence that this bias is small, and it decreases to 0 as the variance of the errors gets small relative to the radius of curvature. We discuss this bias as well as estimation bias (which fortunately appears to operate in the opposite direction) in Section 7.

3. CONNECTIONS BETWEEN PRINCIPAL CURVES AND PRINCIPAL COMPONENTS

In this section we establish some facts that make principal curves appear as a reasonable generalization of linear principal components.

Proposition 1. If a straight line $l(\lambda) = \mathbf{u}_0 + \lambda\mathbf{v}_0$ is self-consistent, then it is a principal component.

Proof. The line has to pass through the origin, because

$$\begin{aligned}\mathbf{0} = E(\mathbf{X}) &= E_\lambda E(\mathbf{X} \mid \lambda_{\mathbf{f}}(\mathbf{X}) = \lambda) \\ &= E_\lambda(\mathbf{u}_0 + \lambda\mathbf{v}_0) \\ &= \mathbf{u}_0 + \bar{\lambda}\mathbf{v}_0.\end{aligned}$$

Therefore, $\mathbf{u}_0 = \mathbf{0}$ (recall that we assumed $\mathbf{u}_0 \perp \mathbf{v}_0$). It remains to show that $\mathbf{v}_0$ is an eigenvector of Σ, the covariance of $\mathbf{X}$:

$$\begin{aligned}\Sigma\mathbf{v}_0 &= E(\mathbf{X}\mathbf{X}^t)\mathbf{v}_0 \\ &= E_\lambda E(\mathbf{X}\mathbf{X}^t\mathbf{v}_0 \mid \lambda_{\mathbf{f}}(\mathbf{X}) = \lambda) \\ &= E_\lambda E(\mathbf{X}\mathbf{X}^t\mathbf{v}_0 \mid \mathbf{X}^t\mathbf{v}_0 = \lambda) \\ &= E_\lambda E(\lambda\mathbf{X} \mid \mathbf{X}^t\mathbf{v}_0 = \lambda) \\ &= E_\lambda\lambda^2\mathbf{v}_0.\end{aligned}$$

Principal components need not be self-consistent in the sense of the definition; however, they are self-consistent with respect to linear regression.

Proposition 2. Suppose that $\mathbf{l}(\lambda)$ is a straight line, and that we linearly regress the p components X_j of $\mathbf{X}$ on the projection $\lambda\mathbf{l}(\mathbf{X})$ resulting in linear functions $f_j(\lambda)$. Then, $\mathbf{f} = \mathbf{l}$ iff $\mathbf{v}_0$ is an eigenvector of Σ and $\mathbf{u}_0 = 0$.

The proof of this requires only elementary linear algebra and is omitted.

A Distance Property of Principal Curves

An important property of principal components is that they are critical points of the distance from the observations.

Let $d(\mathbf{x}, \mathbf{f})$ denote the usual euclidean distance from a point $\mathbf{x}$ to its projection on $\mathbf{f}$: $d(\mathbf{x}, \mathbf{f}) = \|\mathbf{x} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{x}))\|$, and define $D^2(h, \mathbf{f}) = E_h d^2(\mathbf{X}, \mathbf{f})$. Consider a straight line $\mathbf{l}(\lambda) = \mathbf{u} + \lambda\mathbf{v}$. The distance $D^2(h, \mathbf{f})$ in this case may be regarded as a function of $\mathbf{u}$ and $\mathbf{v}$: $D^2(h, l) = D^2(h, \mathbf{u}, \mathbf{v})$. It is well known that $\text{grad}_{u,v} D^2(h, \mathbf{u}, \mathbf{v}) = 0$ iff $\mathbf{u} = \mathbf{0}$ and $\mathbf{v}$ is an eigenvector of Σ, that is, the line l is a principal-component line.

We now restate this fact in a variational setting and extend it to principal curves. Let $\mathcal{G}$ denote a class of curves parameterized over Λ. For $\mathbf{g} \in \mathcal{G}$ define $\mathbf{f}_t = \mathbf{f} + t\mathbf{g}$. This creates a perturbed version of $\mathbf{f}$.

Definition 2. The curve $\mathbf{f}$ is called a critical point of the distance function for variations in the class $\mathcal{G}$ iff

$$\left.\frac{dD^2(h, \mathbf{f}_t)}{dt}\right|_{t=0} = 0 \quad \forall\, \mathbf{g} \in \mathcal{G}.$$

Proposition 3. Let $\mathcal{G}_l$ denote the class of straight lines $\mathbf{g}(\lambda) = \mathbf{a} + \lambda\mathbf{b}$. A straight line $l_0(\lambda) = \mathbf{a}_0 + \lambda\mathbf{b}_0$ is a critical point of the distance function for variations in $\mathcal{G}_l$ iff $\mathbf{b}_0$ is an eigenvector of $\text{cov}(\mathbf{X})$ and $\mathbf{a}_0 = \mathbf{0}$.

The proof involves straightforward linear algebra and is omitted. A result analogous to Proposition 3 holds for principal curves.

Proposition 4. Let $\mathcal{G}_B$ denote the class of smooth (C^∞) curves parameterized over Λ, with $\|\mathbf{g}\| \leq 1$ and $\|\mathbf{g}'\| \leq 1$. Then $\mathbf{f}$ is a principal curve of h iff $\mathbf{f}$ is a critical point of the distance function for perturbations in $\mathcal{G}_B$.

A proof of Proposition 4 is given in the Appendix. The condition that $\|\mathbf{g}\|$ is bounded guarantees that $\mathbf{f}_t$ lies in a thin tube around $\mathbf{f}$ and that the tubes shrink uniformly, as $t \to 0$. The boundedness of $\|\mathbf{g}'\|$ ensures that for t small enough, $\mathbf{f}_t'$ is well behaved and, in particular, bounded away from 0 for $t < 1$. Both conditions together guarantee that, for small enough t, λ_{f_t} is well defined.

4. AN ALGORITHM FOR FINDING PRINCIPAL CURVES

By analogy to linear principal-component analysis, we are particularly interested in finding smooth curves corresponding to local minima of the distance function. Our strategy is to start with a smooth curve, such as the largest linear principal component, and check if it is a principal curve. This involves projecting the data onto the curve and then evaluating their expectation conditional on where they project. Either this conditional expectation coincides with the curve, or we get a new curve as a by-product. We then check if the new curve is self-consistent, and so on. If the self-consistency condition is met, we have found a principal curve. It is easy to show that both of the operations of projection and conditional expectation reduce the expected distance from the points to the curve.

The Principal-Curve Algorithm

The previous discussion motivates the following iterative algorithm.

Initialization: Set $\mathbf{f}^{(0)}(\lambda) = \bar{\mathbf{x}} + \mathbf{a}\lambda$, where $\mathbf{a}$ is the first linear principal component of h. Set $\lambda^{(0)}(\mathbf{x}) = \lambda_{\mathbf{f}^{(0)}}(\mathbf{x})$.

Repeat: Over iteration counter j

1. Set $\mathbf{f}^{(j)}(\cdot) = E(\mathbf{X} \mid \lambda_{\mathbf{f}^{(j-1)}}(\mathbf{X}) = \cdot)$.
2. Define $\lambda^{(j)}(\mathbf{x}) = \lambda_{\mathbf{f}^{(j)}}(\mathbf{x})\ \forall\ \mathbf{x} \in h$; transform $\lambda^{(j)}$ so that $\mathbf{f}^{(j)}$ is unit speed.
3. Evaluate $D^2(h, \mathbf{f}^{(j)}) = E_{\lambda^{(j)}} E[\|\mathbf{X} - \mathbf{f}(\lambda^{(j)}(\mathbf{X}))\|^2 \mid \lambda^{(j)}(\mathbf{X})]$.

Until: The change in $D^2(h, \mathbf{f}^{(j)})$ is below some threshold.

There are potential problems with this algorithm. Although principal curves are by definition differentiable, there is no guarantee that the curves produced by the conditional-expectation step of the algorithm have this property. Discontinuities can certainly occur at the endpoints of a curve. The problem is illustrated in Figure 4, where the expected values of the observations projecting onto $\mathbf{f}(\lambda_{\min})$ and $\mathbf{f}(\lambda_{\max})$ are disjoint from the new curve. If this occurs, we have to join $\mathbf{f}(\lambda_{\min})$ and $\mathbf{f}(\lambda_{\max})$ to the rest of the curve in a differentiable fashion. In light of the previous discussion, we cannot prove that the algorithm converges. All we have is some evidence in its favor:

1. By definition, principal curves are fixed points of the algorithm.
2. Assuming that each iteration is well defined and produces a differentiable curve, we can show that the expected distance $D^2(h, \mathbf{f}^{(j)})$ converges.

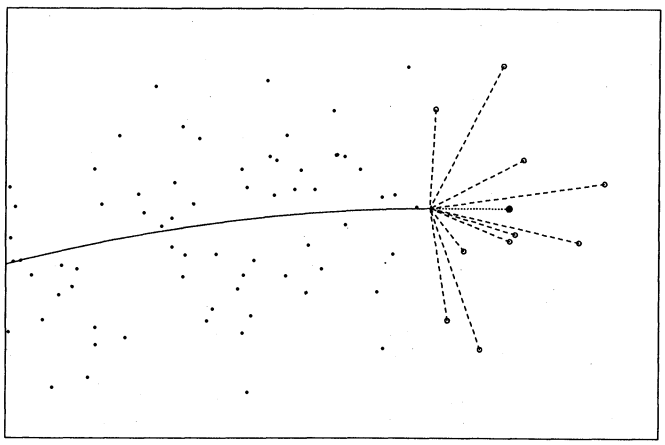

Figure 4. The mean of the observations projecting onto an endpoint of the curve can be disjoint from the rest of the curve.

3. If the conditional-expectation operation in the principal-curve algorithm is replaced by fitting a least squares straight line, then the procedure converges to the largest principal component.

5. PRINCIPAL CURVES FOR DATA SETS

So far, we have considered principal curves of a multivariate probability distribution. In reality, however, we usually work with finite multivariate data sets. Suppose that $\mathbf{X}$ is an $n \times p$ matrix of n observations on p variables. We regard the data set as a sample from an underlying probability distribution.

A curve $\mathbf{f}(\lambda)$ is represented by n tuples $(\lambda_i, \mathbf{f}_i)$, joined up in increasing order of λ to form a polygon. Clearly, the geometric shape of the polygon depends only on the order, not on the actual values of the λ_i. We always assume that the tuples are sorted in increasing order of λ, and we use the arc-length parameterization, for which $\lambda_1 = 0$ and λ_i is the arc length along the polygon from $\mathbf{f}_1$ to $\mathbf{f}_i$. This is the discrete version of the unit-speed parameterization.

As in the distribution case, the algorithm alternates between a projection step and an expectation step. In the absence of prior information we use the first principal-component line as a starting curve; the $\mathbf{f}_i$ are taken to be the projections of the n observations onto the line.

We iterate until the relative change in the distance $|D^2(h, \mathbf{f}^{(j-1)}) - D^2(h, \mathbf{f}^{(j)})|/D^2(h, \mathbf{f}^{(j-1)})$ is below some threshold (we use .001). The distance is estimated in the obvious way, adding up the squared distances of the points in the sample to their closest points on the current curve. We are unable to prove that the algorithm converges, or that each step guarantees a decrease in the criterion. In practice, we have had no convergence problems with more than 40 real and simulated examples.

5.1 The Projection Step

For fixed $\mathbf{f}^{(j)}(\cdot)$ we wish to find for each $\mathbf{x}_i$ in the sample the value $\lambda_i = \lambda_{\mathbf{f}^{(j)}}(\mathbf{x}_i)$.

Define d_{ik} as the distance between $\mathbf{x}_i$ and its closest point on the line segment joining each pair $(\mathbf{f}^{(j)}(\lambda_k^{(j)}), \mathbf{f}^{(j)}(\lambda_{k+1}^{(j)}))$. Corresponding to each d_{ik} is a value $\lambda_{ik} \in [\lambda_k^{(j)}, \lambda_{k+1}^{(j)}]$. We then set λ_i to the λ_{ik} corresponding to the smallest value

of d_{ik}:

$$\lambda_i = \lambda_{ik^*} \quad \text{if } d_{ik^*} = \min_{k=1}^{n-1} d_{ik}. \tag{4}$$

Corresponding to each λ_i is an interpolated $\mathbf{f}_i^{(j)}$; using these values to represent the curve, we replace λ_i by the arc length from $\mathbf{f}_1^{(j)}$ to $\mathbf{f}_i^{(j)}$.

5.2 The Conditional-Expectation Step: Scatterplot Smoothing

The goal of this step is to estimate $\mathbf{f}^{(j+1)}(\lambda) = E(\mathbf{X} \mid \lambda_{\mathbf{f}^{(j)}} = \lambda)$. We restrict ourselves to estimating this quantity at n values of λ, namely $\lambda_1, \ldots, \lambda_n$ found in the projection step. A natural way of estimating $E(\mathbf{X} \mid \lambda_{\mathbf{f}^{(j)}} = \lambda_i)$ would be to gather all of the observations that project onto $\mathbf{f}^{(j)}$ at λ_i and find their mean. Unfortunately, there is generally only one such observation, $\mathbf{x}_i$. It is at this stage that we introduce the *scatterplot smoother*, a fundamental building block in the principal-curve procedure for finite data sets. We estimate the conditional expectation at λ_i by averaging all of the observations $\mathbf{x}_k$ in the sample for which λ_k is *close* to λ_i. As long as these observations are close enough and the underlying conditional expectation is smooth, the bias introduced in approximating the conditional expectation is small. On the other hand, the variance of the estimate decreases as we include more observations in the neighborhood.

Scatterplot Smoothing. Local averaging is not a new idea. In the more common regression context, scatterplot smoothers are used to estimate the regression function $E(Y \mid x)$ by local averaging. Some commonly used smoothers are kernel smoothers (e.g., Watson 1964), spline smoothers (Silverman 1985; Wahba and Wold 1975), and the locally weighted running-line smoother of Cleveland (1979). All of these 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. Our current implementation of the algorithm is an *S* function (Becker, Chambers, and Wilks 1988) that allows any scatterplot smoother to be used. We have experience with all of those previously mentioned, although all of the examples were fitted using locally weighted running lines. We give a brief description; for details see Cleveland (1979).

Locally Weighted Running-Lines Smoother. Consider the estimation of $E(x \mid \lambda)$, that is, a single coordinate function based on a sample of pairs $(\lambda_1, x_1), \ldots, (\lambda_n, x_n)$, and assume the λ_i are ordered. To estimate $E(x \mid \lambda)$, the smoother fits a straight line to the wn observations $\{x_j\}$ closest in λ_j to λ_i. The estimate is taken to be the fitted value of the line at λ_i. The fraction w of points in the neighborhood is called the span. In fitting the line, weighted least squares regression is used. The weights are derived from a symmetric kernel centered at λ_i that dies smoothly to 0 within the neighborhood. Specifically, if h_i is the distance to the wnth nearest neighbor, then the points λ_j in the neighborhood get weights $w_{ij} = (1 - |(\lambda_j - \lambda_i)/h_i|^3)^3$.

5.3 A Demonstration of the Algorithm

To illustrate the principal-curve procedure, we generated a set of 100 data points from a circle in two dimensions with independent Gaussian errors in both coordinates:

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

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

Figure 5 shows the data, the circle (dashed line), and the estimated curve (solid line) for selected steps of the iteration. The starting curve is the first principal component (Fig. 5a). Any line through the origin is a principal curve for the population model (5), but this is not generally the case for data. Here the algorithm converges to an estimate for another population principal curve, the circle. This example is admittedly artificial, but it presents the principal-curve procedure with a particularly tough job. The starting guess is wholly inappropriate and the projection of the points onto this line does not nearly represent the final ordering of the points when projected onto the solution curve. Points project in a certain order on the starting vector (as depicted in Fig. 6). The new curve is a function of $\lambda^{(0)}$ obtained by averaging the coordinates of points close in $\lambda^{(0)}$. The new $\lambda^{(1)}$ 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 from the ordering along the previous curve. This enables the successive curves to bend to shapes that could not be parameterized as a function of the linear principal component.

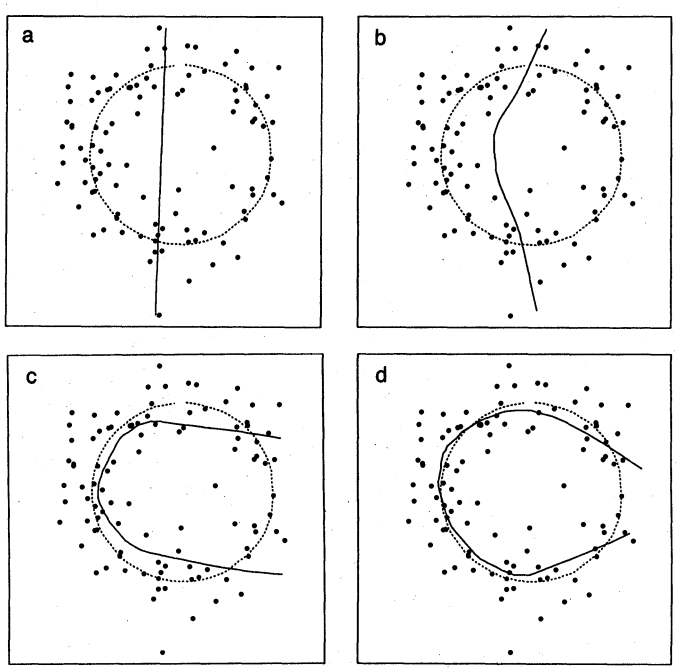

Figure 5. Selected Iterates of the Principal-Curve Procedure for the Circle Data. In all of the figures we see the data, the circle from which the data are generated, and the current estimate produced by the algorithm: (a) the starting curve is the principal-component line, with average squared distance $D^2(\hat{f}^{(0)}) = 12.91$; (b) iteration 2: $D^2(\hat{f}^{(2)}) = 10.43$; (c) iteration 4: $D^2(\hat{f}^{(4)}) = 2.58$; (d) final iteration 8: $D^2(\hat{f}^{(8)}) = 1.55$.

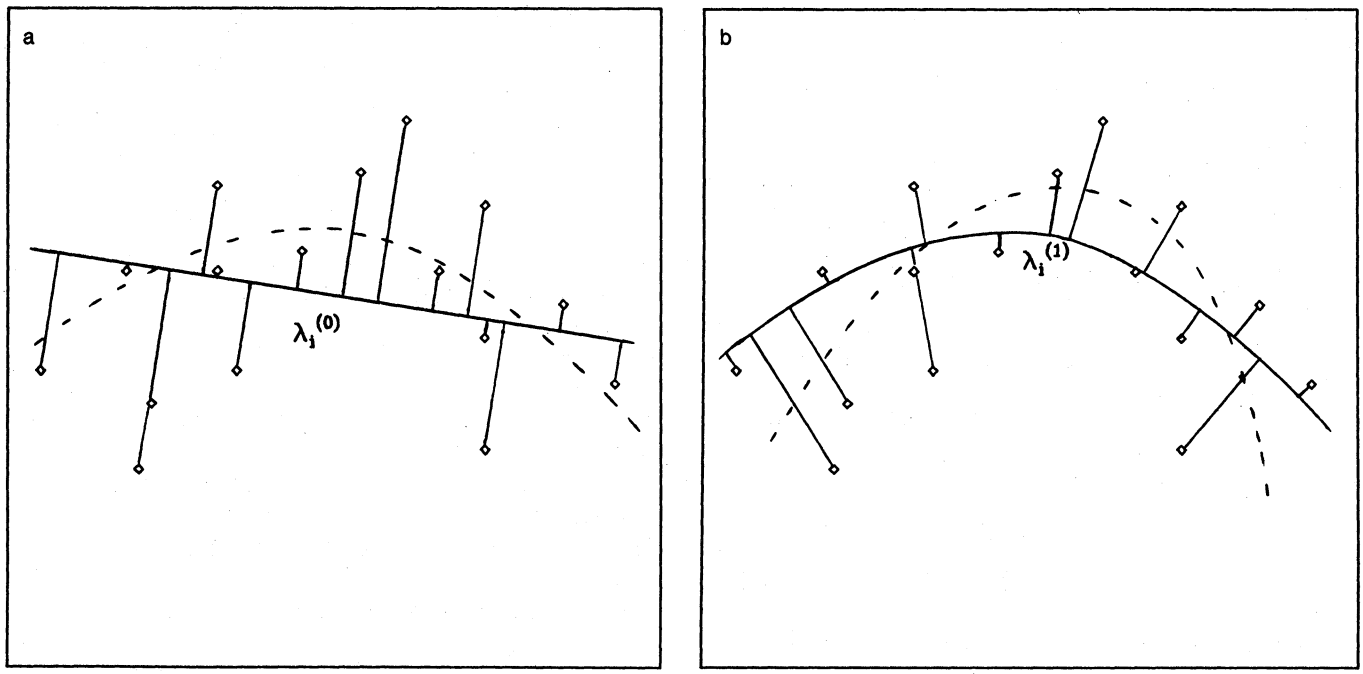

Figure 6. Schematics Emphasizing the Iterative Nature of the Algorithm. The curve of the first iteration is a function of $\lambda^{(0)}$ measured along the starting vector (a). The curve of the second iteration is a function of $\lambda^{(1)}$ measured along the curve of the first iteration (b).

5.4 Span Selection for the Scatterplot Smoother

The crucial parameter of any local averaging smoother is the size of the neighborhood over which averaging takes place. We discuss the choice of the span w for the locally weighted running-line smoother.

A Fixed-Span Strategy. The common first guess for $\mathbf{f}$ is a straight line. In many interesting situations, the final curve is not a function of the arc length of this initial curve (see Fig. 6). It is reached by successively bending the original curve. We have found that if the initial span of the smoother is too small, the curve may bend too fast, and follow the data too closely. Our most successful strategy has been to initially use a large span, and then to decrease it gradually. In particular, we start with a span of $.6n$ observations in each neighborhood, and let the algorithm converge (according to the criterion outlined previously). We then drop the span to $.5n$ and iterate till convergence. Finally, the same is done at $.4n$, by which time the procedure has found the general shape of the curve. The curves in Figure 5 were found using this strategy.

Spans of this magnitude have frequently been found appropriate for scatterplot smoothing in the regression context. In some applications, especially the two-dimensional ones, we can plot the curve and the points and select a span that seems appropriate for the data. Other applications, such as the collider-ring example in Section 8, have a natural criterion for selecting the span.

Automatic Span Selection by Cross-Validation. Assume the procedure has converged to a self-consistent (with respect to the smoother) curve for the span last used. We do not want the fitted curve to be too wiggly relative to the density of the data. As we reduce the span, the average distance decreases and the curve follows the data more closely. The human eye is skilled at making trade-offs between smoothness and fidelity to the data; we would like a procedure that makes this judgment automatically.

A similar situation arises in nonparametric regression, where we have a response y and a covariate x. One rationale for making the smoothness judgment automatically is to ensure that the fitted function of x does a good job in predicting future responses. Cross-validation (Stone 1974) is an approximate method for achieving this goal, and proceeds as follows. We predict each response y_i in the sample using a smooth estimated from the sample with the ith observation omitted; let $\hat{y}_{(i)}$ be this predicted value, and define the cross-validated residual sum of squares as $\text{CVRSS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2$. CVRSS/n is an approximately unbiased estimate of the expected squared prediction error. If the span is too large, the curve will miss features in the data, and the bias component of the prediction error will dominate. If the span is too small, the curve begins to fit the noise in the data, and the variance component of the prediction error will increase. We pick the span that corresponds to the minimum CVRSS.

In the principal-curve algorithm, we can use the same procedure for estimating the spans for each coordinate function separately, as a final smoothing step. Since most smoothers have this feature built in as an option, cross-validation in this manner is trivial to implement. Figure 7a shows the final curve after one more smoothing step, using cross-validation to select the span—nothing much has changed.

On the other hand, Figure 7b shows what happens if we continue iterating with the cross-validated smoothers. The spans get successively smaller, until the curve almost interpolates the data. In some situations, such as the Stanford linear collider example in Section 8, this may be exactly what we want. It is unlikely, however, that in this

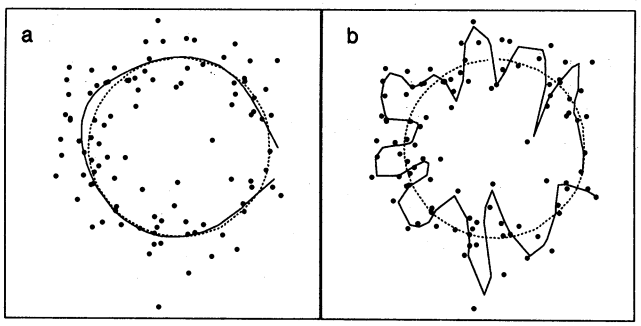

Figure 7. (a) The Final Curve in Figure 6 With One More Smoothing Step, Using Cross-Validation Separately for Each of the Coordinates—$D^2(\hat{f}^{(9)}) = 1.28$. (b) The Curve Obtained by Continuing the Iterations (−.12), Using Cross-Validation at Every Step.

event cross-validation would be used to pick the span. A possible explanation for this behavior is that the errors in the coordinate functions are autocorrelated; cross-validation in this situation tends to pick spans that are too small (Hart and Wehrly 1986).

5.5 Principal Curves and Splines

Our algorithm for estimating principal curves from samples is motivated by the algorithm for finding principal curves of densities, which in turn is motivated by the definition of principal curves. This is analogous to the motivation for kernel smoothers and locally weighted running-line smoothers. They estimate a conditional expectation, a population quantity that minimizes a population criterion. They do not minimize a data-dependent criterion.

On the other hand, smoothing splines do minimize data-dependent criteria. The cubic smoothing spline for a set of n pairs $(\lambda_1, x_1), \ldots, (\lambda_n, x_n)$ and penalty (smoothing parameter) μ minimizes

$$D^2(f) = \sum_{i=1}^{n}(x_i - f(\lambda_i))^2 + \mu \int (f''(\lambda))^2 \, d\lambda, \quad (6)$$

among all functions f with f' absolutely continuous and $f'' \in L_2$ (e.g., see Silverman 1985). We suggest the following criterion for defining principal curves in this context: Find $\mathbf{f}(\lambda)$ and $\lambda_i \in [0, 1]$ $(i = 1, \ldots, n)$ so that

$$D^2(\mathbf{f}, \lambda) = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{f}(\lambda_i)\|^2 + \mu \int_0^1 \|\mathbf{f}''(\lambda)\|^2 \, d\lambda \quad (7)$$

is minimized over all $\mathbf{f}$ with $f_j \in S_2[0, 1]$. Notice that we have confined the functions to the unit interval and thus do not use the unit-speed parameterization. Intuitively, for a fixed smoothing parameter μ, functions defined over an arbitrarily large interval can satisfy the second-derivative smoothness criterion and visit every point. It is easy to make this argument rigorous.

We now apply our alternating algorithm to these criteria:

1. Given $\mathbf{f}$, minimizing $D^2(\mathbf{f}, \lambda)$ over λ_i only involves the first part of (7) and is our usual projection step. The λ_i are rescaled to lie in [0, 1].

2. Given λ_i, (7) splits up into p expressions of the form (6), one for each coordinate function. These are optimized by smoothing the p coordinates against λ_i using a cubic spline smoother with parameter μ.

The usual penalized least squares arguments show that if a minimum exists, it must be a cubic spline in each coordinate. We make no claims about its existence, or about global convergence properties of this algorithm.

An advantage of the spline-smoothing algorithm is that it can be computed in $O(n)$ operations, and thus is a strong competitor for the kernel-type smoothers that take $O(n^2)$ unless approximations are used. Although it is difficult to guess the smoothing parameter μ, alternative methods such as using the approximate degrees of freedom (see Cleveland 1979) are available for assessing the amount of smoothing and thus selecting the parameter.

Our current implementation of the algorithm allows a choice of smoothing splines or locally weighted running lines, and we have found it difficult to distinguish their performance in practice.

5.6 Further Illustrations and Discussion of the Algorithm

The procedure worked well on the circle example and several other artificial examples. Nevertheless, sometimes its behavior is surprising, at least at first glance. Consider a data set from a spherically symmetric unimodal distribution centered at the origin. A circle with radius $E\|\mathbf{x}\|$ is a principal curve, as are all straight lines passing through the origin. The circle, however, has smaller expected squared distance from the observations than the lines.

The 150 points in Figure 8 were sampled independently from a bivariate spherical Gaussian distribution. When the principal-curve procedure is started from the circle, it does not move much, except at the endpoints (as depicted in Fig. 8a). This is a consequence of the smoothers' endpoint behavior in that it is not constrained to be periodic. Figure 8b shows what happens when we use a periodic version of the smoother, and also start at a circle. Nevertheless, starting from the linear principal component (where theoretically it should stay), and using the nonperiodic smoother, the algorithm iterates to a curve that, apart from the endpoints, appears to be attempting to model the circle. (See Fig. 8c; this behavior occurred repeatedly over several simulations of this example. The ends of the curve are stuck and further iterations do not free them.)

The example illustrates the fact that the algorithm tends to find curves that are minima of the distance function. This is not surprising; after all, the principal-curve algorithm is a generalization of the power method for finding eigenvectors, which exhibits exactly the same behavior. The power method tends to converge to an eigenvector for the largest eigenvalue, unless special precautions are taken.

Interestingly, the algorithm using the periodic smoother and starting from the linear principal component finds a circle identical to that in Figure 8b.

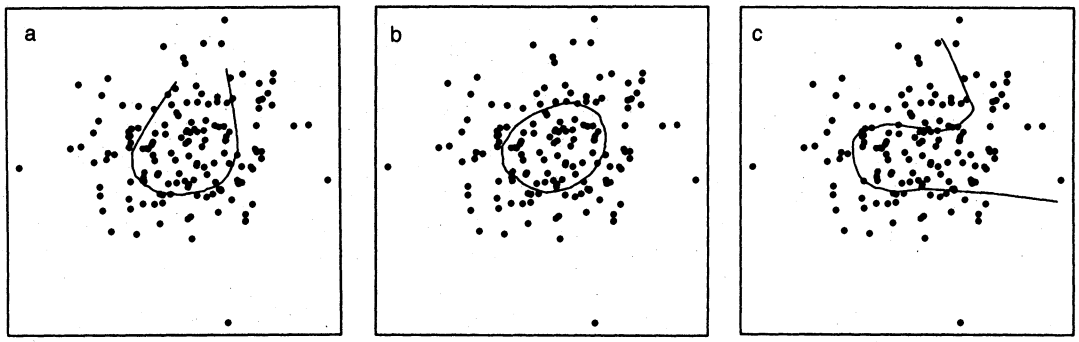

Figure 8. Some Curves Produced by the Algorithm Applied to Bivariate Spherical Gaussian Data: (a) The Curve Found When the Algorithm Is Started at a Circle Centered at the Mean; (b) The Circle Found Starting With Either a Circle or a Line but Using a Periodic Smoother; (c) The Curve Found Using the Regular Smoother, but Starting at a Line. A periodic smoother ensures that the curve found is closed.

6. BIAS CONSIDERATIONS: MODEL AND ESTIMATION BIAS

Model bias occurs when the data are of the form $\mathbf{x} = \mathbf{f}(\lambda) + \mathbf{e}$ and we wish to recover $\mathbf{f}(\lambda)$. In general, if $\mathbf{f}(\lambda)$ has curvature, it is not a principal curve for the distribution it generates. As a consequence, the principal-curve procedure can only find a biased version of $\mathbf{f}(\lambda)$, even if it starts at the generating curve. This bias goes to 0 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 by averaging over neighborhoods, which usually has a flattening effect. We demonstrate this bias with a simple example.

A Simple Model for Investigating Bias

Suppose that the curve $\mathbf{f}$ is an arc of a circle centered at the origin and with radius ρ, and the data $\mathbf{x}$ are generated from a bivariate Gaussian, with mean chosen uniformly on the arc and variance $\sigma^2 I$. Figure 9 depicts the situation. Intuitively, it seems that more mass is put outside the circle than inside, so the circle closest to the data should have radius larger than ρ. Consider the points that project onto a small arc $\Lambda_\theta(\lambda)$ of the circle with angle θ centered at λ, as depicted in the figure. 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. This implies that the conditional expectation lies outside the circle.

We can prove (Hastie 1984) that $E(\mathbf{x} \mid \lambda_f(\mathbf{x}) \in \Lambda_\theta(\lambda)) = (r_\theta/\rho\mathbf{f})(\lambda)$, where

$$r_\theta = r^* \frac{\sin(\theta/2)}{\theta/2} \tag{8}$$

and

$$r^* = E[(\rho + e_1)^2 + e_2^2]^{1/2}$$
$$\approx \rho + (\sigma^2/2\rho).$$

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

Equation (8) nicely separates the two components of bias. Even if we had infinitely many observations and thus would not need local averaging to estimate conditional expectation, the circle with radius ρ would not be a stationary point of the algorithm; the principal curve is a circle with radius $r^* > \rho$. The factor $\sin(\theta/2)/(\theta/2)$ is attributable to local averaging. 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 is needed to determine it. Typically, these quantities will change as we move along the curve. Hastie (1984) gives a demonstration that these bias patterns persist in a situation where the curvature changes along the curve.

7. EXAMPLES

This section contains two examples that illustrate the use of the procedure.

7.1 The Stanford Linear Collider Project

This application of principal curves was implemented by a group of geodetic engineers at the Stanford Linear Accelerator Center (SLAC) in California. They used the

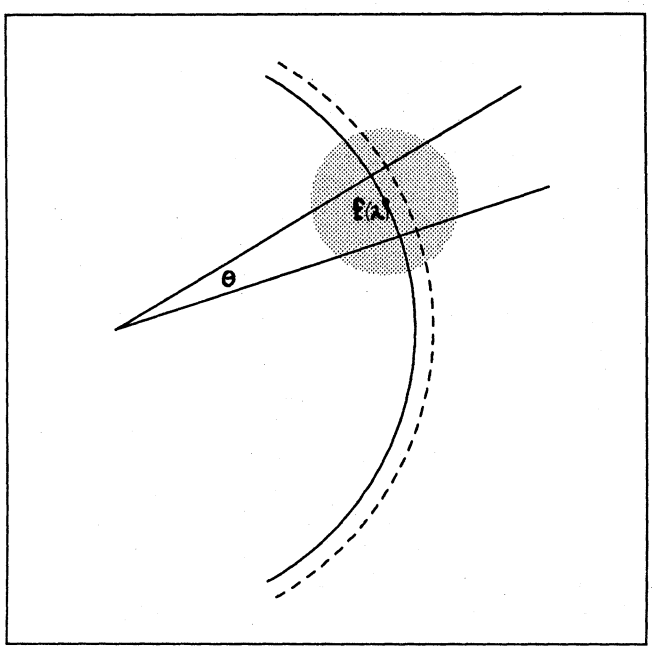

Figure 9. The data are generated from the arc of a circle with radius ρ and with iid $N(\mathbf{0}, \sigma^2 I)$ errors. The location on the circle is selected uniformly. The best fitting circle (dashed) has radius larger than the generating curve.

software developed by the authors in consultation with the first author and Jerome Friedman of SLAC.

The Stanford linear collider (SLC) collides two intense and finely focused particle beams. Details of the collision are recorded in a collision chamber and studied by particle physicists, whose major goal is to discover new subatomic particles. Since there is only one linear accelerator at SLAC, it is used to accelerate a positron and an electron bunch in a single pulse, and the collider arcs bend these beams to bring them to collision (see Fig. 10).

Each of the two collider arcs contain roughly 475 magnets (23 segments of 20 plus some extras), which guide the positron and electron beam. Ideally, these magnets lie on a smooth curve with a circumference of about 3 kilometers (km) (as depicted in the schematic). The collider has a third dimension, and actually resembles a floppy tennis racket, because the tunnel containing the magnets goes underground (whereas the accelerator is aboveground).

Measurement errors were inevitable in the procedure used to place the magnets. This resulted in the magnets lying close to the planned curve, but with errors in the range of ± 1.0 millimeters (mm). A consequence of these errors was that the beam could not be adequately focused.

The engineers realized that it was not necessary to move the magnets to the ideal curve, but rather to a curve through the existing magnet positions that was smooth enough to allow focused bending of the beam. This strategy would theoretically reduce the amount of magnet movement necessary. The principal-curve procedure was used to find this curve. The remainder of this section describes some special features of this simple but important application.

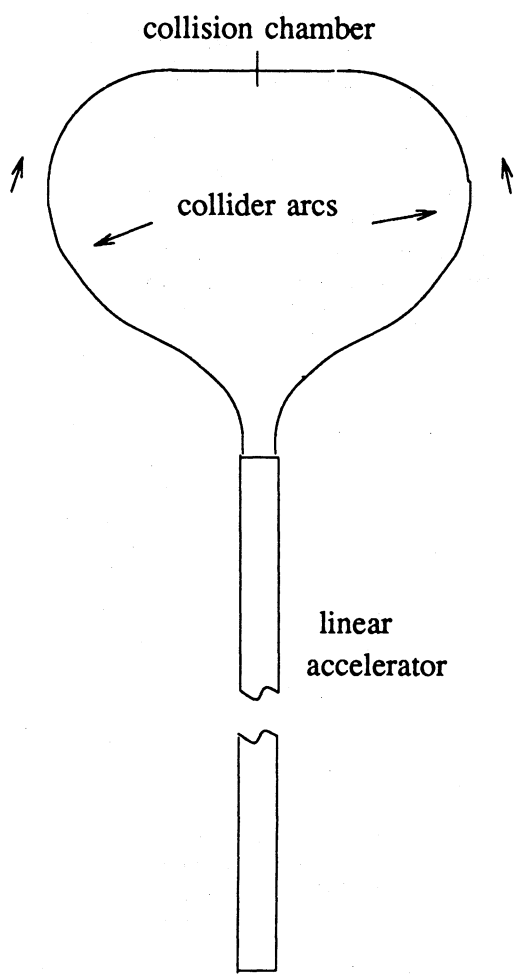

Figure 10. A Rough Schematic of the Stanford Linear Accelerator and the Linear Collider Ring.

Initial attempts at fitting curves used the data in the measured three-dimensional goedetic coordinates, but it was found that the magnet displacements were small relative to the bias induced by smoothing. The theoretical arc was then removed, and subsequent curve fitting was based on the residuals. This was achieved by replacing the three coordinates of each magnet with three new coordinates: (a) the arc length from the beginning of the arc till the point of projection onto the ideal curve (x), (b) the distance from the magnet to this projection in the horizontal plane (y), and (c) the distance in the vertical plane (z).

This technique effectively removed the major component of the bias and is an illustration of how special situations lend themselves to adaptations of the basic procedure. Of course, knowledge of the ideal curve is not usually available in other applications.

There is a natural way of choosing the smoothing parameter in this application. The fitted curve, once transformed back to the original coordinates, can be represented by a polygon with a vertex at each magnet. The angle between these segments is of vital importance, since the further it is from 180°, the harder it is to launch the particle beams into the next segment without hitting the wall of the beam pipe [diameter 1 centimeter (cm)]. In fact, if θ_i measures the departure of this angle from 180°, the operating characteristics of the magnet specify a threshold $\theta_{\max}$ of .1 milleradian. Now, no smoothing results in no magnet movement (no work), but with many magnets violating the threshold. As the amount of smoothing (span) is increased, the angles tend to decrease, and the residuals and thus the amounts of magnet movement increase. The strategy was to increase the span until no magnets violated the angle constraint. Figure 11 gives the fitted vertical and horizontal components of the chosen curve, for a section of the north arc consisting of 149 magnets. This relatively rough curve was then translated back to the original coordinates, and the appropriate adjustments for each magnet were determined. The systematic trend in these coordinate functions represents systematic departures of the magnets from the theoretical curve. Only 66% of the magnets needed to be moved, since the remaining 34% of the residuals were below 60 μm in length and thus considered negligible.

There are some natural constraints on the system. Some of the magnets were fixed by design and thus could not be moved. The beam enters the arc parallel to the accelerator, so the initial magnets do no bending. Similarly, there are junction points at which no bending is allowed. These constraints are accommodated by attaching weights to the points representing the magnets and using a

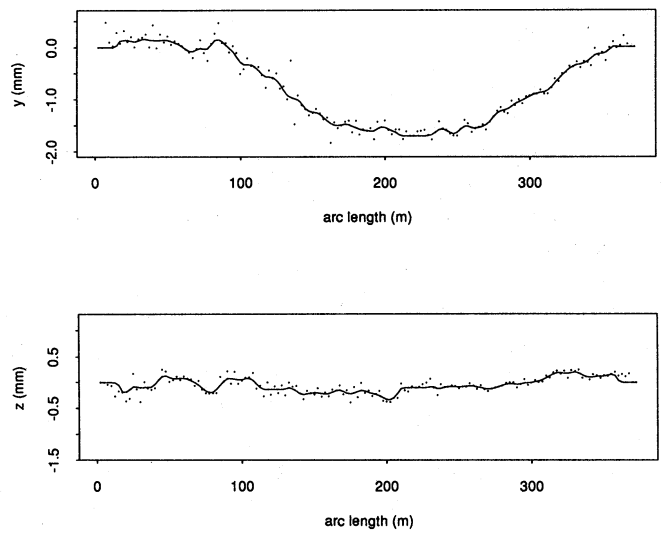

Figure 11. The Fitted Coordinate Functions for the Magnet Positions for a Section of the Standard Linear Collider. The data represent residuals from the theoretical curve. Some (35%) of the deviations from the fitted curve were small enough that these magnets were not moved.

weighted version of the smoother in the algorithm. By giving the fixed magnets sufficiently large weights, the constraints are met. Figure 11 has the parallel constraints built in at the endpoints.

Finally, since some of the magnets were way off target, we used a resistant version of the fitting procedure. Points are weighted according to their distance from the fitted curve, and deviations beyond a fixed threshold are given weight 0.

7.2 Gold Assay Pairs

A California-based company collects computer-chip waste to sell it for its content of gold and other precious metals. Before bidding for a particular cargo, the company takes a sample to estimate the gold content of the whole lot. The sample is split in two. One subsample is assayed by an outside laboratory, and the other by their own in-house laboratory. The company eventually wishes to 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 12 consist of 250 pairs of gold assays. Each point represents an observation $\mathbf{x}_i$ with x_{ji} = log(1 + assay yield for the ith assay pair for lab j), where j = 1 corresponds to the outside lab and j = 2 to the in-house lab. The log transformation stabilizes the variance and produces a more even scatter of points than the untransformed data. [There were many more small assays (1 ounce (oz) per ton) than larger ones (>10 oz per ton).]

Our model for these data is

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} f(\tau_i) \\ \tau_i \end{pmatrix} + \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}, \tag{9}$$

where τ_i is the expected gold content for sample i using the in-house lab assay, $f(\tau_i)$ is the expected assay result for the outside lab relative to the in-house lab, and e_{ji} is measurement error, assumed iid with $\text{var}(e_{1i}) = \text{var}(e_{2i})$ $\forall\, i$.

This is a generalization of the linear errors-in-variables model, the structural model (if we regard the τ_i themselves as unobservable random variables), or the functional model (if the τ_i are considered fixed):

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} \alpha + \beta\tau_i \\ \tau_i \end{pmatrix} + \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}. \tag{10}$$

Model (10) essentially looks for deviations from the 45° line, and is estimated by the first principal component.

Model (9) is a special case of the principal-curve model,

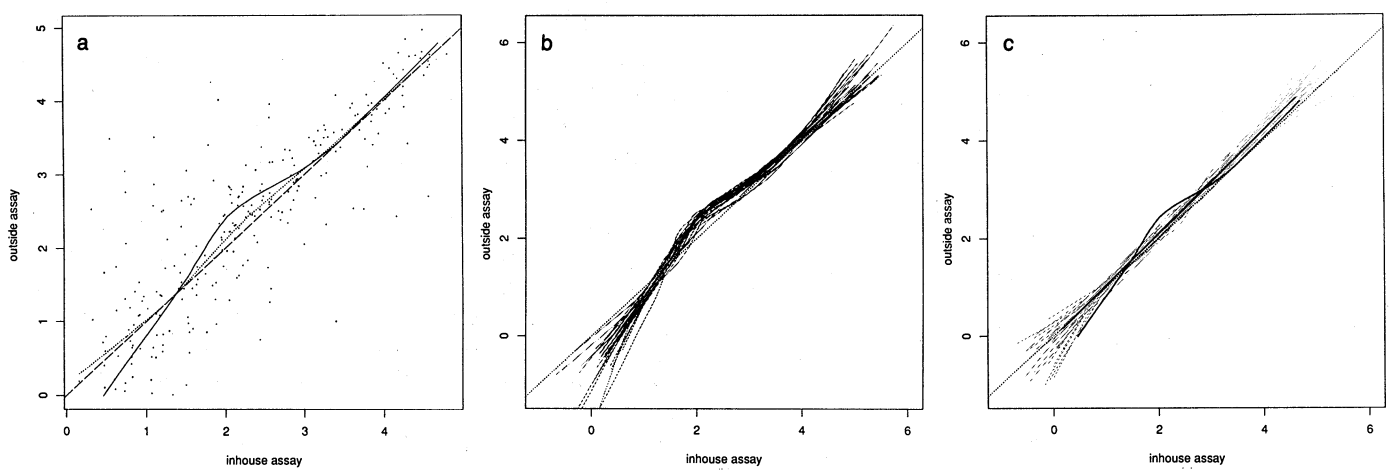

Figure 12. (a) Plot of the Log Assays for the In-House and Outside Labs. The solid curve is the principal curve, the dotted curve the scatterplot smooth, and the dashed curve the 45° line. (b) A Band of 25 Bootstrap Curves. Each curve is the principal curve of a bootstrap sample. A bootstrap sample is obtained by randomly assigning errors to the principal curve for the original data (solid curve). The band of curves appears to be centered at the solid curve, indicating small bias. The spread of the curves gives an indication of variance. (c) Another Band of 25 Bootstrap Curves. Each curve is the principal curve of a bootstrap sample, based on the linear errors-in-variables regression line (solid line). This simulation tests the null hypothesis of no kink. There is evidence that the kink is real, since the principal curve (solid curve) lies outside this band in the region of the kink.

where one of the coordinate functions is the identity. This identifies the systematic component of variable x_2 with the arc-length parameter. Similarly, we estimate (9) using a natural variant of the principal-curve algorithm. In the smoothing step we smooth only x_1 against the current value of τ, and then update τ by projecting the data onto the curve defined by $(f(\tau), \tau)$.

The dotted curve in Figure 12 is the usual scatterplot smooth of x_1 against x_2 and is clearly misleading as a scatterplot summary. The principal curve lies above the 45° line in the interval 1.4–4, which represents an untransformed assay content interval of 3–15 oz/ton. In this interval the in-house assay tends to be 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 lot is less valuable.

A natural question arising at this point is whether the bend in the curve is real, or whether the linear model (10) is adequate. If we had access to more data from the same population we could simply calculate the principal curves for the additional samples and see for how many of them the bend appeared.

In the absence of such additional samples, we use the bootstrap (Efron 1981, 1982) to simulate them. We compute the residual vectors of the observed data from the fitted curve in Figure 12a, and treating them as iid, we pool all 250 of them. Since these residuals are derived from a projection essentially onto a straight line, their expected squared length is half that of the residuals in Model (9). We therefore scale them up by a factor of $\sqrt{2}$. We then sampled with replacement from this pool, and reconstructed a bootstrap replicate by adding a sampled residual vector to each of the fitted values of the original fit. For each of these bootstrapped data sets the entire curve-fitting procedure was applied and the fitted curves were saved. This method of bootstrapping is aimed at exposing both bias and variance.

Figure 12b shows the errors-in-variables principal curves obtained for 25 bootstrap samples. The spreads of these curves give an idea of the variance of the fitted curve. The difference between their average and the original fit estimates the bias, which in this case is negligible.

Figure 12c shows the result of a different bootstrap experiment. Our null hypothesis is that the relationship is linear, and thus we sampled in the same way as before but we replaced the principal curve with the linear errors-in-variables line. The observed curve (thick solid curve) lies outside the band of curves fitted to 25 bootstrapped data sets, providing additional evidence that the bend is indeed real.

8. EXTENSION TO HIGHER DIMENSIONS: PRINCIPAL SURFACES

We have had some success in extending the definitions and algorithms for curves to two-dimensional (globally parameterized) surfaces.

A continuous two-dimensional globally parameterized surface in $\mathbf{R}^p$ is a function $\mathbf{f} : \Lambda \to \mathbf{R}^p$ for $\Lambda \subseteq \mathbf{R}^2$, where $\mathbf{f}$ is a vector of continuous functions:

$$\mathbf{f}(\boldsymbol{\lambda}) = \begin{pmatrix} f_1(\lambda_1, \lambda_2) \\ f_2(\lambda_1, \lambda_2) \\ \vdots \\ f_p(\lambda_1, \lambda_2) \end{pmatrix}. \tag{11}$$

Let $\mathbf{X}$ be defined as before, and let $\mathbf{f}$ denote a smooth two-dimensional surface in $\mathbf{R}^p$, parameterized over $\Lambda \subseteq \mathbf{R}^2$. Here the projection index $\boldsymbol{\lambda}_\mathbf{f}(\mathbf{x})$ is defined to be the parameter value corresponding to the point on the surface closest to $\mathbf{x}$.

The *principal surfaces* of h are those members of $\mathcal{G}^2$ that are self-consistent: $E(\mathbf{X} \mid \boldsymbol{\lambda}_\mathbf{f}(\mathbf{X}) = \boldsymbol{\lambda}) = \mathbf{f}(\boldsymbol{\lambda})$ for a.e. $\boldsymbol{\lambda}$. Figure 13 illustrates the situation. We do not yet have a rigorous justification for these definitions, although we have had success in implementing an algorithm.

The principal-surface algorithm is similar to the curve algorithm; two-dimensional surface smoothers are used instead of one-dimensional scatterplot smoothers. See Hastie (1984) for more details of principal surfaces, the algorithm to compute them, and examples.

9. DISCUSSION

Ours is not the first attempt at finding a method for fitting nonlinear manifolds to multivariate data. In discussing other approaches to the problem we restrict ourselves to one-dimensional manifolds (the case treated in this article).

The approach closest in spirit to ours was suggested by Carroll (1969). He fit a model of the form $\mathbf{x}_i = \mathbf{p}(\lambda_i) + \mathbf{e}_i$, where $\mathbf{p}(\lambda)$ is a vector of polynomials $p_j(\lambda) = \sum_{k=0}^{K_j} a_{jk}\lambda^k$ of prespecified degrees K_j. The goal is to find the coefficients of the polynomials and the λ_i $(i = 1, \ldots, n)$ minimizing the loss function $\sum \|\mathbf{e}_i\|^2$. The algorithm makes use of the fact that for given $\lambda_1, \ldots, \lambda_n$, the optimal

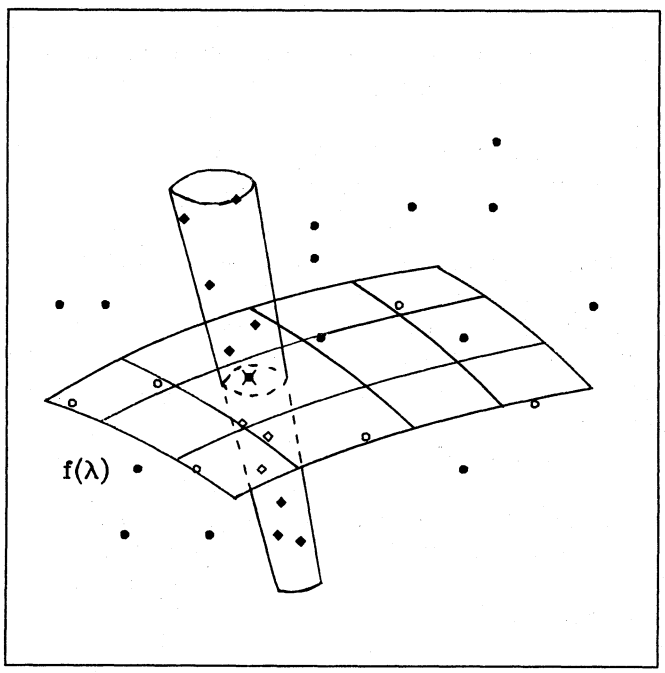

Figure 13. Each point on a principal surface is the average of the points that project there.

polynomial coefficients can be found by linear least squares, and the loss function thus can be written as a function of the λ_i only. Carroll gave an explicit formula for the gradient of the loss function, which is helpful in the n-dimensional numerical optimization required to find the optimal λ's.

The model of Etezadi-Amoli and McDonald (1983) is the same as Carroll's, but they used different goodness-of-fit measures. Their goal was to minimize the off-diagonal elements of the error covariance matrix $\Sigma = E'E$, which is in the spirit of classical linear factor analysis. Various measures for the cumulative size of the off-diagonal elements are suggested, such as $\sum_{i \neq j} \sigma_{ij}^2$. Their algorithm is similar to ours in that it alternates between improving the λ's for given polynomial coefficients and finding the optimal polynomial coefficients for given λ's. The latter is a linear least squares problem, whereas the former constitutes one step of a nonlinear optimization in n parameters.

Shepard and Carroll (1966) proceeded from the assumption that the p-dimensional observation vectors lie exactly on a smooth one-dimensional manifold. In this case, it is possible to find parameter values $\lambda_1, \ldots, \lambda_n$ such that for each one of the p coordinates, x_{ij} varies smoothly with λ_i. The basis of their method is a measure for the degree of smoothness of the dependence of x_{ij} on λ_i. This measure of smoothness, summed over the p coordinates, is then optimized with respect to the λ's: one finds those values of $\lambda_1, \ldots, \lambda_n$ that make the dependence of the coordinates on the λ's as smooth as possible.

We do not go into the definition and motivation of the smoothness measure; it is quite subtle, and we refer the interested reader to the original source. We just wish to point out that instead of optimizing smoothness, one could optimize a combination of smoothness and fidelity to the data as described in Section 5.5, which would lead to modeling the coordinate functions as spline functions and should allow the method to deal with noise in the data better.

In view of this previous work, what do we think is the contribution of the present article?

- From the operational point of view it is advantageous that there is no need to specify a parametric form for the coordinate functions. Because the curve is represented as a polygon, finding the optimal λ's for given coordinate functions is easy. This makes the alternating minimization attractive and allows fitting of principal curves to large data sets.
- From the theoretical point of view, the definition of principal curves as conditional expectations agrees with our mental image of a summary. The characterization of principal curves as critical points of the expected squared distance from the data makes them appear as a natural generalization of linear principal components. This close connection is further emphasized by the fact that linear principal curves are principal components, and that the algorithm converges to the largest principal component if conditional expectations are replaced by least squares straight lines.

APPENDIX: PROOFS OF PROPOSITIONS

We make the following assumptions: Denote by $\mathbf{X}$ a random vector in $\mathbf{R}^p$ with density h and finite second moments. Let $\mathbf{f}$ denote a smooth (C^∞) unit-speed curve in $\mathbf{R}^p$ parameterized over a closed, possibly infinite interval $\Lambda \subseteq \mathbf{R}^1$. We assume that $\mathbf{f}$ does not intersect itself [$\lambda_1 \neq \lambda_2 \Rightarrow \mathbf{f}(\lambda_1) \neq \mathbf{f}(\lambda_2)$] and has finite length inside any finite ball. Under these conditions, the set $\{\mathbf{f}(\lambda), \lambda \in \Lambda\}$ forms a smooth, connected one-dimensional manifold diffeomorphic to the interval Λ. Any smooth, connected one-dimensional manifold is diffeomorphic either to an interval or a circle (Milnor 1965). The results and proofs following could be slightly modified to cover the latter case (closed curves).

Existence of the Projection Index

Existence of the projection index is a consequence of the following two lemmas.

Lemma 5.1. For every $\mathbf{x} \in \mathbf{R}^p$ and for any $r > 0$, the set $Q = \{\lambda \mid \|\mathbf{x} - \mathbf{f}(\lambda)\| \leq r\}$ is compact.

Proof. Q is closed, because $\|\mathbf{x} - \mathbf{f}(\lambda)\|$ is a continuous function of λ. It remains to show that Q is bounded. Suppose that it were not. Then, there would exist an unbounded monotone sequence $\lambda_1, \lambda_2, \ldots$, with $\|\mathbf{x} - \mathbf{f}(\lambda_i)\| \leq r$. Let B denote the ball around $\mathbf{x}$ with radius $2r$. Consider the segment of the curve between $\mathbf{f}(\lambda_i)$ and $\mathbf{f}(\lambda_{i+1})$. The segment either leaves and reenters B, or it stays entirely inside. This means that it contributes at least $\min(2r, |\lambda_{i+1} - \lambda_i|)$ to the length of the curve inside B. As there are infinitely many such segments, and the sequence $\{\lambda_i\}$ is unbounded, $\mathbf{f}$ would have infinite length in B, which is a contradiction.

Lemma 5.2. For every $\mathbf{x} \in \mathbf{R}^p$, there exists $\lambda \in \Lambda$ for which $\|\mathbf{x} - \mathbf{f}(\lambda)\| = \inf_{\mu \in \Lambda} \|\mathbf{x} - \mathbf{f}(\mu)\|$.

Proof. Define $r = \inf_{\mu \in \Lambda} \|\mathbf{x} - \mathbf{f}(\mu)\|$. Set $B = \{\mu \mid \|\mathbf{x} - \mathbf{f}(\mu)\| \leq 2r\}$. Obviously, $\inf_{\mu \in \Lambda} \|\mathbf{x} - \mathbf{f}(\mu)\| = \inf_{\mu \in B} \|\mathbf{x} - \mathbf{f}(\mu)\|$. Since B is nonempty and compact (Lemma 5.1), the infimum on the right side is attained.

Define $d(\mathbf{x}, \mathbf{f}) = \inf_{\mu \in \Lambda} \|\mathbf{x} - \mathbf{f}(\mu)\|$.

Proposition 5. The projection index $\lambda_{\mathbf{f}}(\mathbf{x}) = \sup_\lambda \{\lambda \mid \|\mathbf{x} - \mathbf{f}(\lambda)\| = d(\mathbf{x}, \mathbf{f})\}$ is well defined.

Proof. The set $\{\lambda \mid \|\mathbf{x} - \mathbf{f}(\lambda)\| = d(\mathbf{x}, \mathbf{f})\}$ is nonempty (Lemma 5.2) and compact (Lemma 5.1), and therefore has the largest element.

It is not hard to show that $\lambda_{\mathbf{f}}(\mathbf{x})$ is measurable; a proof is available on request.

Stationarity of the Distance Function

We first establish some simple facts that are of interest in themselves.

Lemma 6.1. If $\mathbf{f}(\lambda_0)$ is a closest point to $\mathbf{x}$ and $\lambda_0 \in \Lambda^0$, the interior of the parameter interval, then $\mathbf{x}$ is in the normal hyperplane to $\mathbf{f}$ at $\mathbf{f}(\lambda_0)$: $\langle \mathbf{x} - \mathbf{f}(\lambda_0), \mathbf{f}'(\lambda_0) \rangle = 0$.

Proof. $d\|\mathbf{x} - \mathbf{f}(\lambda)\|^2 / d\lambda = 2\langle \mathbf{x} - \mathbf{f}(\lambda), \mathbf{f}'(\lambda) \rangle$. If $\mathbf{f}(\lambda_0)$ is a closest point and the derivative is defined ($\lambda_0 \in \Lambda^0$), then it has to vanish.

Definition. A point $\mathbf{x} \in \mathbf{R}^p$ is called an ambiguity point for a curve $\mathbf{f}$ if it has more than one closest point on the curve: $\text{card}\{\lambda \mid \|\mathbf{x} - \mathbf{f}(\lambda)\| = d(\mathbf{x}, \mathbf{f})\} > 1$.

Let A denote the set of ambiguity points. Our next goal is to show that A is measurable and has measure 0.

Define M_λ, the orthogonal hyperplane to $\mathbf{f}$ at λ, by $M_\lambda = \{\mathbf{x} \mid \langle \mathbf{x} - \mathbf{f}(\lambda), \mathbf{f}'(\lambda) \rangle = 0\}$. Now, we know that if $\mathbf{f}(\lambda)$ is a closest point to $\mathbf{x}$ on the curve and $\lambda \in \Lambda^0$, then $\mathbf{x} \in M_\lambda$. It is useful to

define a mapping that maps $\Lambda \times \mathbf{R}^{p-1}$ into $\bigcup_\lambda M_\lambda$. Choose $p-1$ smooth vector fields $\mathbf{n}_1(\lambda), \ldots, \mathbf{n}_{p-1}(\lambda)$ such that for every λ the vectors $\mathbf{f}'(\lambda)$ and $\mathbf{n}_1(\lambda), \ldots, \mathbf{n}_{p-1}(\lambda)$ are orthogonal. It is well known that such vector fields do exist. Define $\chi : \Lambda \times \mathbf{R}^{p-1} \to \mathbf{R}^p$ by $\chi(\lambda, \mathbf{v}) = \mathbf{f}(\lambda) + \sum_{i=1}^{p-1} v_i \mathbf{n}_i(\lambda)$, and set $M = \chi(\Lambda, \mathbf{R}^{p-1})$, the set of all points in $\mathbf{R}^p$ lying in some hyperplane for some point on the curve. The mapping χ is smooth, because $\mathbf{f}$ and $\mathbf{n}_1, \ldots, \mathbf{n}_{p-1}$ are assumed to be smooth.

We now present a few observations that simplify showing that A has measure 0.

Lemma 6.2. $\mu(A \cap M^c) = 0$.

Proof. Suppose that $\mathbf{x} \in A \cap M^c$. According to Lemma 6.1, this is only possible if Λ is a finite closed interval $[\lambda_{\min}, \lambda_{\max}]$ and $\mathbf{x}$ is equidistant from the endpoints $\mathbf{f}(\lambda_{\min})$ and $\mathbf{f}(\lambda_{\max})$. The set of all such points forms a hyperplane that has measure 0. Therefore $A \cap M^c$, as a subset of this measure-0 set, is measurable and has measure 0.

Lemma 6.3. Let E be a measure-0 set. It is sufficient to show that for every $\mathbf{x} \in \mathbf{R}^p \backslash E$ there exists an open neighborhood $N(\mathbf{x})$ with $\mu(A \cap N(\mathbf{x})) = 0$.

Proof. The open covering $\{N(\mathbf{x}) \mid \mathbf{x} \in \mathbf{R}^p \backslash E\}$ of $\mathbf{R}^p \backslash E$ contains a countable covering $\{N_i\}$, because the topology of $\mathbf{R}^p$ has a countable base.

Lemma 6.4. We can restrict ourselves to the case of compact Λ.

Proof. Set $\Lambda_n = \Lambda \cap [-n, n]$, $\mathbf{f}_n = \mathbf{f}/\Lambda_n$, and A_n as the set of ambiguity points of $\mathbf{f}_n$. Suppose that $\mathbf{x}$ is an ambiguity point of $\mathbf{f}$; then $\{\lambda \mid \|\mathbf{x} - \mathbf{f}(\lambda)\| = d(\mathbf{x}, \mathbf{f})\}$ is compact (Lemma 5.1). Therefore, $\mathbf{x} \in A_n$ for some n, and $A \subset \bigcup_1^\infty A_n$.

We are now ready to prove Proposition 6.

Proposition 6. The set of ambiguity points has measure 0.

Proof. We can restrict ourselves to the case of compact Λ (Lemma 6.4). As $\mu(A \cap M^c) = 0$ (Lemma 6.2) it is sufficient to show that for every $\mathbf{x} \in M$, with the possible exception of a set C of measure 0, there exists a neighborhood $N(\mathbf{x})$ with $\mu(A \cap N(\mathbf{x})) = 0$.

We choose C to be the set of critical values of χ. [A point $\mathbf{y} \in M$ is called a regular value if $\operatorname{rank}(\chi'(\mathbf{x})) = p$ for all $\mathbf{x} \in \chi^{-1}(\mathbf{y})$; otherwise $\mathbf{y}$ is called a critical value.] By Sard's theorem (Milnor 1965), C has measure 0.

Pick $\mathbf{x} \in M \cap C^c$. We first show that $\chi^{-1}(\mathbf{x})$ is a finite set $\{(\lambda_1, \mathbf{v}_1), \ldots, (\lambda_k, \mathbf{v}_k)\}$. Suppose that on the contrary there was an infinite set $\{(\xi_1, \mathbf{w}_1), (\xi_2, \mathbf{w}_2), \ldots\}$ with $\chi(\xi_i, \mathbf{w}_i) = \mathbf{x}$. By compactness of Λ and continuity of χ, there would exist a cluster point ξ_0 of $\{\xi_1, \xi_2, \ldots\}$ and a corresponding $\mathbf{w}_0$ with $\chi(\xi_0, \mathbf{w}_0) = \mathbf{x}$. On the other hand, $\mathbf{x}$ was assumed to be a regular value of χ, and thus χ would be a diffeomorphism between a neighborhood of $(\lambda_0, \mathbf{w}_0)$ and a neighborhood of $\mathbf{x}$. This is a contradiction.

Because $\mathbf{x}$ is a regular value, there are neighborhoods $L_i(\lambda_i, \mathbf{v}_i)$ and a neighborhood $N(\mathbf{x})$ such that χ is a diffeomorphism between L_i and N. Actually, a stronger statement holds. We can find $\tilde{N}(\mathbf{x}) \subset N(\mathbf{x})$, for which $\chi^{-1}(\tilde{N}) \subset \bigcup_1^k L_i$. Suppose that this were not the case. Then, there would exist a sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots \to \mathbf{x}$ and corresponding $(\xi_i, \mathbf{w}_i) \notin \bigcup_{i=1}^k L_i$ with $\chi(\xi_i, \mathbf{w}_i) = \mathbf{x}_i$. The set $\{\xi_1, \xi_2, \ldots\}$ has a cluster point $\xi_0 \notin \bigcup L_i$, and by continuity $\chi(\xi_0, \mathbf{w}_0) = \mathbf{x}$, which is a contradiction.

We have now shown that for $\mathbf{y} \in \tilde{N}(\mathbf{x})$ there exists exactly one pair $(\lambda_i(\mathbf{y}), \mathbf{v}_i(\mathbf{y})) \in L_i$, with $\chi(\lambda_i(\mathbf{y}), \mathbf{v}_i(\mathbf{y})) = \mathbf{y}$, and $\lambda_i(\mathbf{y})$ is a smooth function of $\mathbf{y}$. Define $\lambda_0(\mathbf{y}) = \lambda_{\min}$ and $\lambda_{k+1}(\mathbf{y}) = \lambda_{\max}$. Set $d_i(\mathbf{y}) = \|\mathbf{y} = \mathbf{f}(\lambda_i(\mathbf{y}))\|^2$. A simple calculation using the chain rule and the fact that $\langle \mathbf{y} - \mathbf{f}(\lambda_i(\mathbf{y})), \mathbf{f}'(\lambda_i(\mathbf{y})) \rangle = 0$ (Lemma 6.1) shows that $\operatorname{grad}(d_i(\mathbf{y})) = 2(\mathbf{y} - \mathbf{f}(\lambda_i(\mathbf{y})))$. A point $\mathbf{y} \in \tilde{N}(\mathbf{x})$ can be an ambiguity point only if $\mathbf{y} \in A_{ij}$ for some $i \neq j$, where $A_{ij} = \{\mathbf{z} \in \tilde{N}(\mathbf{x}) \mid d_i(\mathbf{z}) = d_j(\mathbf{z}), \lambda_i(\mathbf{z}) \neq \lambda_j(\mathbf{z})\}$. Nevertheless, for $\lambda_i(\mathbf{z}) \neq \lambda_j(\mathbf{z})$, $\operatorname{grad}(d_i(\mathbf{z}) - d_j(\mathbf{z})) \neq 0$, because the curve $\mathbf{f}(\lambda)$ was assumed not to intersect itself. Thus A_{ij} is a smooth, possibly not connected manifold of dimension $p-1$, which has measure 0, and $\mu(A \cap \tilde{N}(\mathbf{x})) \leq \sum_{i,j} \mu(A_{ij}) = 0$.

We have glossed over a technicality: Sard's theorem requires h to be defined on an open set. Nevertheless, we can always extend f in a smooth way beyond the boundaries of the interval.

In the following, let $\mathcal{G}_B$ denote the class of smooth curves parameterized over Λ, with $\|\mathbf{g}(\lambda)\| \leq 1$ and $\|\mathbf{g}'(\lambda)\| \leq 1$. For $\mathbf{g} \in \mathcal{G}_B$, define $\mathbf{f}_t(\lambda) = \mathbf{f}(\lambda) + t\mathbf{g}(\lambda)$. It is easy to see that $\mathbf{f}_t$ has finite length inside any finite ball and for $t < 1$, $\lambda_{\mathbf{f}_t}$ is well defined. Moreover, we have the following lemma.

Lemma 4.1. If $\mathbf{x}$ is not an ambiguity point for $\mathbf{f}$, then $\lim_{t \downarrow 0} \lambda_{\mathbf{f}_t}(\mathbf{x}) = \lambda_{\mathbf{f}}(\mathbf{x})$.

Proof. We have to show that for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $t < \delta$, $|\lambda_{\mathbf{f}_t}(\mathbf{x}) - \lambda_{\mathbf{f}}(\mathbf{x})| < \varepsilon$. Set $C = \Lambda \cap (\lambda_{\mathbf{f}}(\mathbf{x}) - \varepsilon, \lambda_{\mathbf{f}}(\mathbf{x}) + \varepsilon)^c$ and $d_c = \inf_{\lambda \in C} \|\mathbf{x} - \mathbf{f}(\lambda)\|$. The infimum is attained and $d_c > \|\mathbf{x} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{x}))\|$, because $\mathbf{x}$ is not an ambiguity point. Set $\delta = \frac{1}{3}(d_c - \|\mathbf{x} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{x}))\|)$. Now, $\lambda_{\mathbf{f}_t}(\mathbf{x}) \in (\lambda_{\mathbf{f}}(\mathbf{x}) - \varepsilon$ and $\lambda_{\mathbf{f}}(\mathbf{x}) + \varepsilon)$ $\forall\, t < \delta$, because

$$\begin{aligned} &\inf_{\lambda \in C} (\|\mathbf{x} - \mathbf{f}_t(\lambda)\| - \|\mathbf{x} - \mathbf{f}_t(\lambda_{\mathbf{f}}(\mathbf{x}))\|) \\ &\quad \geq d_c - \delta - \|\mathbf{x} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{x}))\| - \delta \\ &\quad = \delta > 0. \end{aligned}$$

Proof of Proposition 4. The curve $\mathbf{f}$ is a principal curve of h iff

$$\left.\frac{dD^2(h, \mathbf{f}_t)}{dt}\right|_{t=0} = 0 \quad \forall\, \mathbf{g} \in \mathcal{G}_B.$$

We use the dominated convergence theorem to show that we can interchange the orders of integration and differentiation in the expression

$$\frac{d}{dt} D^2(h, \mathbf{f}_t) = \frac{d}{dt} E_h \|\mathbf{X} - \mathbf{f}_t(\lambda_{\mathbf{f}_t}(\mathbf{X}))\|^2. \tag{A.1}$$

We need to find a pair of integrable random variables that almost surely bound

$$Z_t = \frac{\|\mathbf{X} - \mathbf{f}_t(\lambda_{\mathbf{f}_t}(\mathbf{X}))\|^2 - \|\mathbf{X} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{X}))\|^2}{t}$$

for all sufficiently small $t > 0$.

Now,

$$Z_t \leq \frac{\|\mathbf{X} - \mathbf{f}_t(\lambda_{\mathbf{f}}(\mathbf{X}))\|^2 - \|\mathbf{X} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{X}))\|^2}{t}.$$

Expanding the first norm we get

$$\begin{aligned} \|\mathbf{X} - \mathbf{f}_t(\lambda_{\mathbf{f}}(\mathbf{X}))\|^2 = \|\mathbf{X} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{X}))\|^2 + t^2 \|\mathbf{g}(\lambda_{\mathbf{f}}(\mathbf{X}))\|^2 \\ - 2t(\mathbf{X} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{X}))) \cdot \mathbf{g}(\lambda_{\mathbf{f}}(\mathbf{X})), \end{aligned}$$

and thus

$$Z_t \leq -2(\mathbf{X} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{X}))) \cdot \mathbf{g}(\lambda_{\mathbf{f}}(\mathbf{X})) + t\|\mathbf{g}(\lambda_{\mathbf{f}}(\mathbf{X}))\|^2. \tag{A.2}$$

Using the Cauchy–Schwarz inequality and the assumption that $\|\mathbf{g}\| \leq 1$, $Z_t \leq 2\|\mathbf{X} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{X}))\| + 1 \leq 2\|\mathbf{X} - \mathbf{f}(\lambda_0)\| + 1$ $\forall\, t < 1$ and arbitrary λ_0. As $\|\mathbf{X}\|$ was assumed to be integrable, so is $\|\mathbf{X} - \mathbf{f}(\lambda_0)\|$, and therefore Z_t.

Similarly, we have

$$Z_t \geq \frac{\|\mathbf{X} - \mathbf{f}_t(\lambda_{\mathbf{f}_t}(\mathbf{X}))\|^2 - \|\mathbf{X} - \mathbf{f}(\lambda_{\mathbf{f}_t}(\mathbf{X}))\|^2}{t}.$$

Expanding the first norm as before, we get

$$\begin{aligned} Z_t &\geq -2(\mathbf{X} - \mathbf{f}(\lambda_{\mathbf{f}_t}(\mathbf{X}))) \cdot \mathbf{g}(\lambda_{\mathbf{f}_t}(\mathbf{X})) \\ &\geq -2\|\mathbf{X} - \mathbf{f}(\lambda_{\mathbf{f}_t}(\mathbf{X}))\| \\ &\geq -2\|\mathbf{X} - \mathbf{f}(\lambda_0)\|, \end{aligned} \tag{A.3}$$

which is once again integrable. By the dominated convergence theorem, the interchange is justified. From (A.1) and (A.2), however, and because $\mathbf{f}$ and $\mathbf{g}$ are continuous functions, we see that the limit $\lim_{t\to 0} Z_t$ exists whenever $\lambda_{\mathbf{f}_t}(\mathbf{X})$ is continuous in t at $t = 0$. We have proved this continuity for a.e. $\mathbf{x}$ in Lemma 4.1. Moreover, this limit is given by $\lim_{t\to 0} Z_t = -2[\mathbf{X} - \mathbf{f}(\lambda_{\mathbf{f}}(\mathbf{X}))] \cdot \mathbf{g}(\lambda_{\mathbf{f}}(\mathbf{X}))$, by (A.1) and (A.2).

Denoting the distribution of $\lambda_{\mathbf{f}}(\mathbf{X})$ by h_λ, we get

$$\frac{d}{dt} D^2(h, \mathbf{f}_t)|_{t=0} = -2E_{h_\lambda}[(E(\mathbf{X} \mid \lambda_{\mathbf{f}}(\mathbf{X}) = \lambda) - \mathbf{f}(\lambda)) \cdot \mathbf{g}(\lambda)]. \tag{A.4}$$

If $\mathbf{f}(\lambda)$ is a principal curve of h, then by definition $E(\mathbf{X} \mid \lambda_{\mathbf{f}}(\mathbf{X}) = \lambda) = \mathbf{f}(\lambda)$ for a.e. λ, and thus

$$\frac{d}{dt} D^2(h, \mathbf{f}_t)|_{t=0} = 0 \quad \forall\, \mathbf{g} \in \mathcal{G}_B.$$

Conversely, suppose that

$$E_{h_\lambda}[E(\mathbf{X} - \mathbf{f}(\lambda) \mid \lambda_{\mathbf{f}}(\mathbf{X}) = \lambda) \cdot \mathbf{g}(\lambda)] = 0, \tag{A.5}$$

for all $\mathbf{g} \in \mathcal{G}_B$. Consider each coordinate separately, and reexpress (A.5) as

$$E_{h_\lambda} k(\lambda) g(\lambda) = 0 \quad \forall\, g \in \mathcal{G}_B. \tag{A.6}$$

This implies that $k(\lambda) = 0$ a.s.

Construction of Densities With Known Principal Curves

Let $\mathbf{f}$ be parameterized over a compact interval Λ. It is easy to construct densities with a carrier in a tube around $\mathbf{f}$, for which $\mathbf{f}$ is a principal curve.

Denote by B_r the ball in $\mathbf{R}^{p-1}$ with radius r and center at the origin. The construction is based on the following proposition.

Proposition 7. If Λ is compact, there exists $r > 0$ such that $\chi \mid \Lambda \times B_r$ is a diffeomorphism.

Proof. Suppose that the result were not true. Pick a sequence $r_i \to 0$. There would exist sequences $(\lambda_i, \mathbf{v}_i) \neq (\xi_i, \mathbf{w}_i)$, with $\|\mathbf{v}_i\| \leq r_i$, $\|\mathbf{w}_i\| \leq r_i$, and $\chi(\lambda_i, \mathbf{c}_i) = \chi(\xi_i, \mathbf{w}_i)$.

The sequences λ_i and ξ_i have cluster points λ_0 and ξ_0. We must have $\lambda_0 = \xi_0$, because

$$\begin{aligned} \mathbf{f}(\xi_0) &= \chi(\xi_0, 0) \\ &= \lim_{i\to\infty} \chi(\xi_i, \mathbf{v}_i) \\ &= \chi(\lambda_0, 0) \\ &= \mathbf{f}(\lambda_0), \end{aligned}$$

and by assumption $\mathbf{f}$ does not intersect itself. So there would be sequences $(\lambda_i, \mathbf{v}_i)$ and $(\xi_i, \mathbf{w}_i)$ converging to $(\lambda_0, 0)$, with $\chi(\lambda_i, \mathbf{v}_i) = \chi(\xi_i, \mathbf{w}_i)$. Nevertheless, it is easy to see that $(\lambda_0, \mathbf{0})$ is a regular point of χ and thus maps a neighborhood of $(\lambda_0, \mathbf{0})$ diffeomorphically into a neighborhood of $\mathbf{f}(\lambda_0)$, which is a contradiction.

Define $T(\mathbf{f}, r) = \chi(\Lambda \times B_r)$. Proposition 7 assures that there are no ambiguity points in $T(\mathbf{f}, r)$ and $\lambda_{\mathbf{f}}(\mathbf{x}) = \lambda$ for $\mathbf{x} \in \chi(\lambda, B_r)$.

Pick a density $\zeta(\lambda)$ on Λ and a density $\psi(\mathbf{v})$ on B_r, with $\int_{B_r} \mathbf{v}\psi(\mathbf{v}) = 0$. The mapping χ carries the product density $\zeta(\lambda) \cdot \psi(\mathbf{v})$ on $\Lambda \times B_r$ into a density $h(\mathbf{x})$ on $T(\mathbf{f}, r)$. It is easy to verify that $\mathbf{f}$ is a principal curve for h.

[*Received December 1984. Revised December 1988.*]

REFERENCES

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

Becker, R., Chambers, J., and Wilks, A. (1988), *The New S Language,* New York: Wadsworth.

Carroll, D. J. (1969), "Polynomial Factor Analysis," in *Proceedings of the 77th Annual Convention,* Arlington, VA: American Psychological Association, pp. 103–104.

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

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

——— (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans* (CBMS-MSF Regional Conference Service in Applied Mathematics, No. 38), Philadelphia: Society for Industrial and Applied Mathematics.

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

Golub, G. H., and Van Loan, C. (1979), "Total Least Squares," in *Smoothing Techniques for Curve Estimation,* Heidelberg: Springer-Verlag, pp. 69–76.

Hart, J., and Wehrly, T. (1986), "Kernel Regression Estimation Using Repeated Measurement Data," *Journal of the American Statistical Association,* 81, 1080–1088.

Hastie, T. J. (1984), "Principal Curves and Surfaces," Laboratory for Computational Statistics Technical Report 11, Stanford University, Dept. of Statistics.

Milnor, J. W. (1965), *Topology From the Differentiable Viewpoint,* Charlottesville: University of Virginia Press.

Shepard, R. N., and Carroll, D. J. (1966), "Parametric Representations of Non-linear Data Structures," in *Multivariate Analysis,* ed. P. R. Krishnaiah, New York: Academic Press, pp. 561–592.

Silverman, B. W. (1985), "Some Aspects of Spline Smoothing Approaches to Non-parametric Regression Curve Fitting," *Journal of the Royal Statistical Society,* Ser. B, 47, 1–52.

Stone, M. (1974), "Cross-validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 36, 111–147.

Thorpe, J. A. (1979), *Elementary Topics in Differential Geometry,* New York: Springer-Verlag.

Wahba, G., and Wold, S. (1975), "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation," *Communications in Statistics,* 4, 1–7.

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhyā,* Ser. A, 26, 359–372.