Chapter 3

The Principal Curve and Surface models

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

3.1. The principal curves of a probability distribution.

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

3.1.1. One dimensional curves.

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

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

If $\|\boldsymbol{f}'(z)\| \equiv 1$ then $l = \lambda_1 - \lambda_0$. This is a rather desirable situation, since if all the coordinate variables are in the same units of measurement, then λ is also in those units. The vector $\boldsymbol{f}'(\lambda)$ is tangent to the curve at λ and is sometimes called the ***velocity vector*** at λ. A curve with $\|\boldsymbol{f}'\| \equiv 1$ is called a unit speed parametrized curve. We can always reparametrize any smooth curve to make it unit speed. If $\boldsymbol{v}$ is a unit vector, then $\boldsymbol{f}(\lambda) = \boldsymbol{v}_0 + \lambda \boldsymbol{v}$ is a unit speed *straight* curve.

The vector $\boldsymbol{f}''(\lambda)$ 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 $\boldsymbol{f}''/\|\boldsymbol{f}''\|$

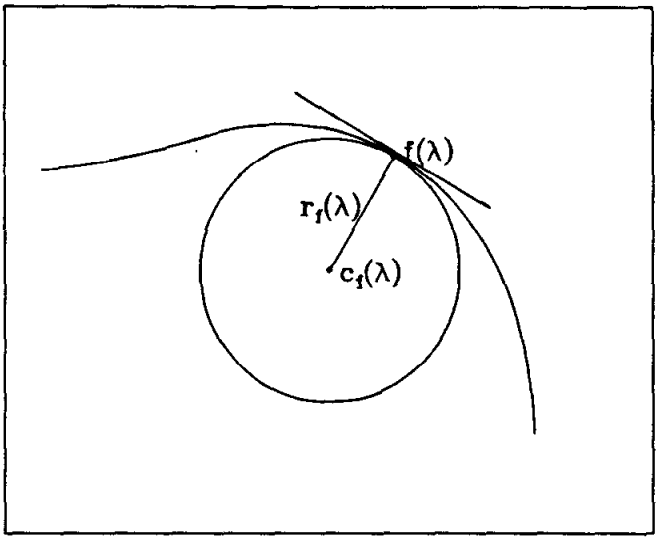

Figure (3.1) The radius of curvature is the radius of the circle tangent to the curve with the same acceleration as the curve.

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

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

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

3.1.2. Definition of principal curves.

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

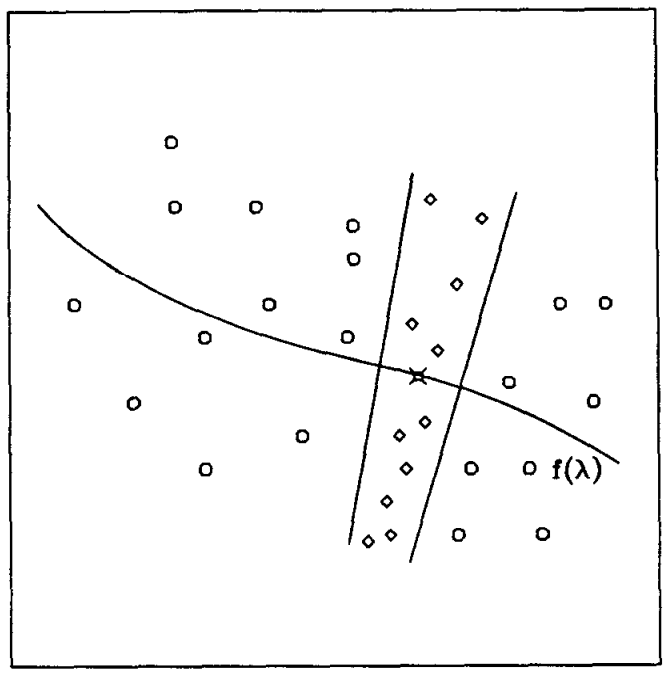

Figure (3.2) Each point on a principal curve is the average of the points that project there.

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

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

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

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

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

Definition

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

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

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

3.1.3. Existence of principal curves.

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

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

3.1.4. The distance property of principal curves.

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

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

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

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

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

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

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

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

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

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

3.1.4.1 A smooth subset of principal curves.

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

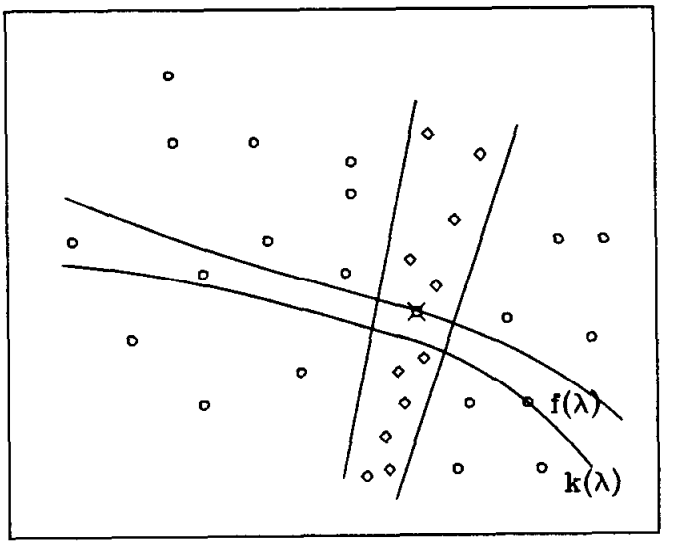

Figure 3.3 The conditional expectation curve gets at least as close to the points as the original curve.

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

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

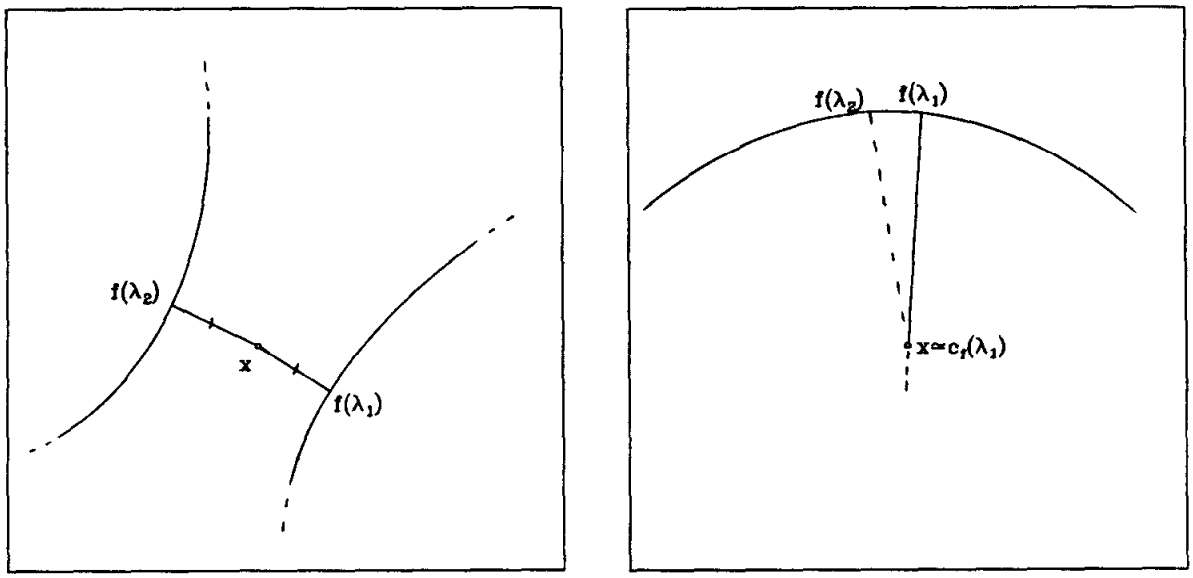

Figure 3.4 The continuity constraint avoids global ambiguities (a) and local ambiguities (b) in projection.

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

3.2. The principal surfaces of a probability distribution.

3.2.1. Two dimensional surfaces.

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

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

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

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

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

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

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

3.2.2. Definition of principal surfaces.

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

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

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

Definition

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

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

Figure (3.5) demonstrates the situation.

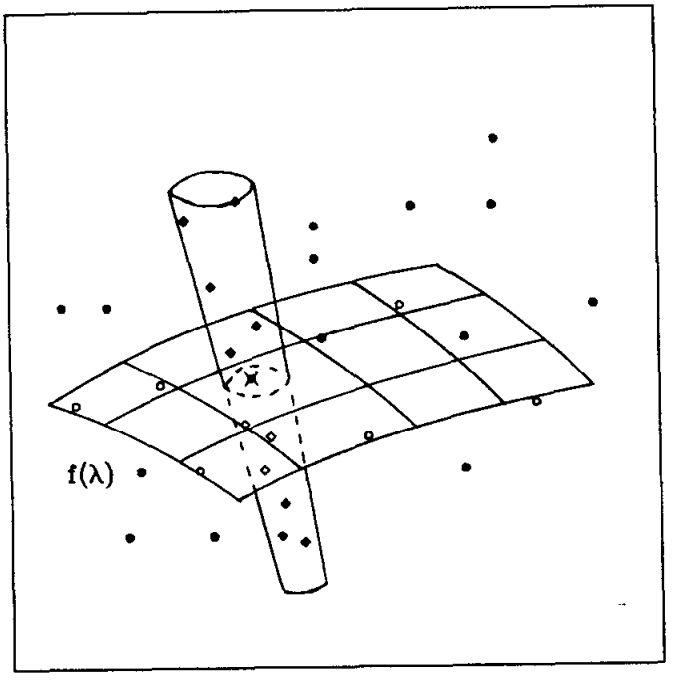

Figure 3.5 Each point on a principal surface is the average of the points that project there.

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

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

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

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

3.3. An algorithm for finding principal curves and surfaces.

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

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

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

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

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

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

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

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

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

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

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

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

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

Principal surface algorithm

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

repeat: over iteration counter j

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

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

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

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

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

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

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

3.4. Principal curves and surfaces for data sets.

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

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

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

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

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

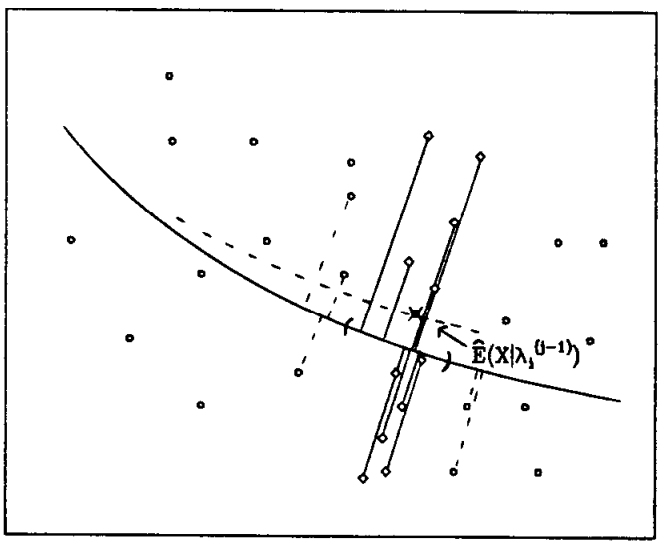

Figure 3.6 We estimate the conditional expectation $\mathbf{E}(X \mid \lambda^{(j-1)} = \hat{\lambda}_i^{(j-1)})$ by averaging the observations x_k for which $\hat{\lambda}_k^{(j-1)}$ is close to $\hat{\lambda}_i^{(j-1)}$.

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

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

3.5. Demonstrations of the procedures.

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

3.5.1. The circle in two-space.

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

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

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

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

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

The square root of the average squared orthogonal distance is displayed at each iteration. If the true curve was linear the expected orthogonal distance for any point would be $\sqrt{\mathbf{E}\chi_1^2} = 1$. We will see in chapter 4 that for this situation, the true circle does not

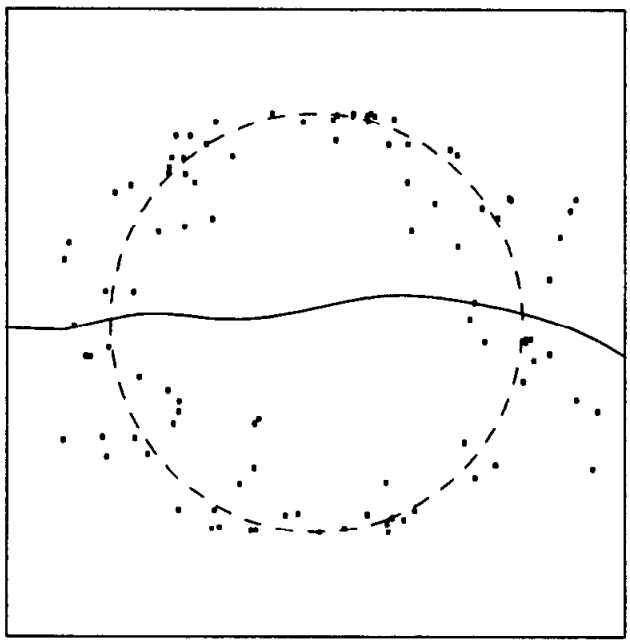

Figure 3.7a The dashed curve is the usual scatterplot smooth. $D(S) = 3.35$

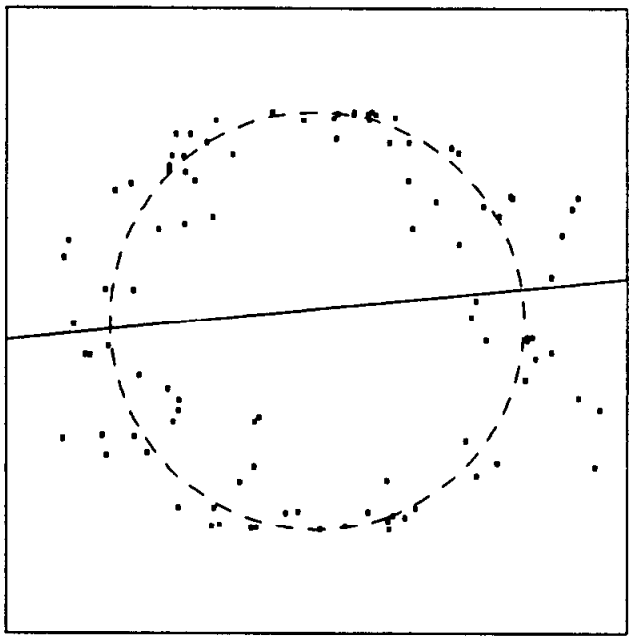

Figure 3.7b The dashed curve is the principal component line. $D(\hat{f}^{(0)}) = 3.43$

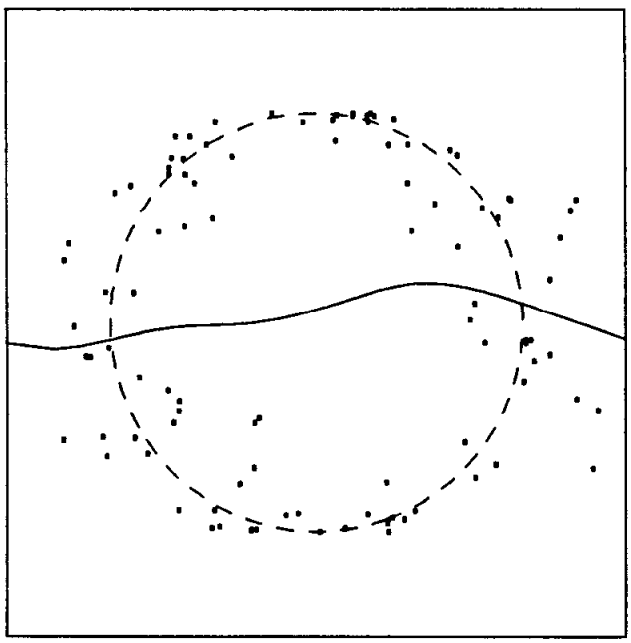

Figure 3.7c $D(\hat{f}^{(1)}) = 3.34$

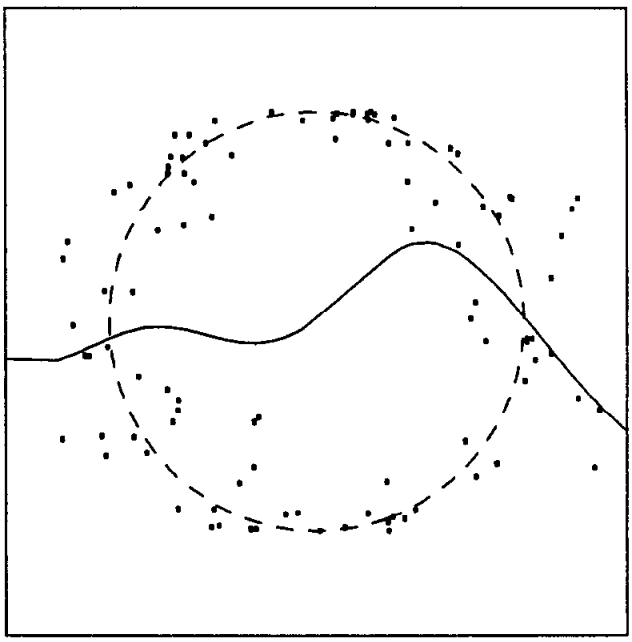

Figure 3.7d $D(\hat{f}^{(2)}) = 3.03$

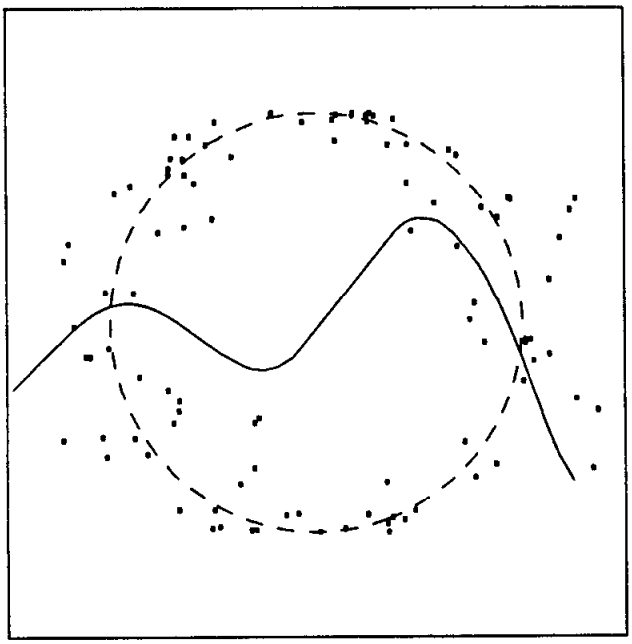

Figure 3.7e $D(\hat{f}^{(3)}) = 2.64$

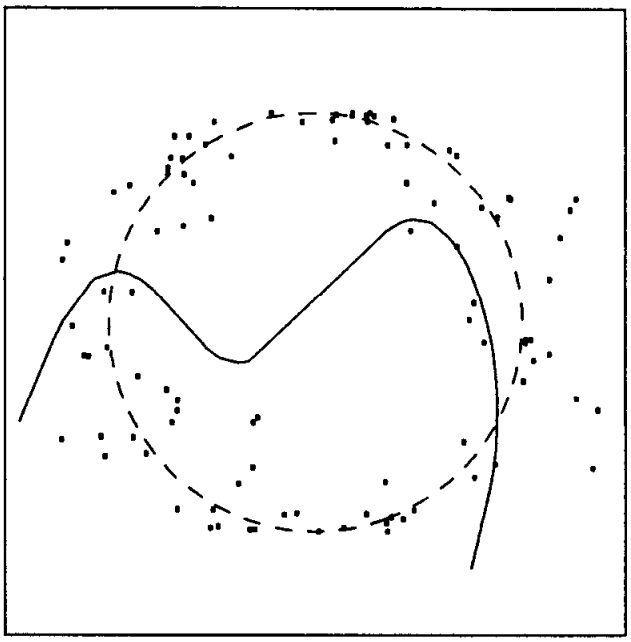

Figure 3.7f $D(\hat{f}^{(4)}) = 2.37$

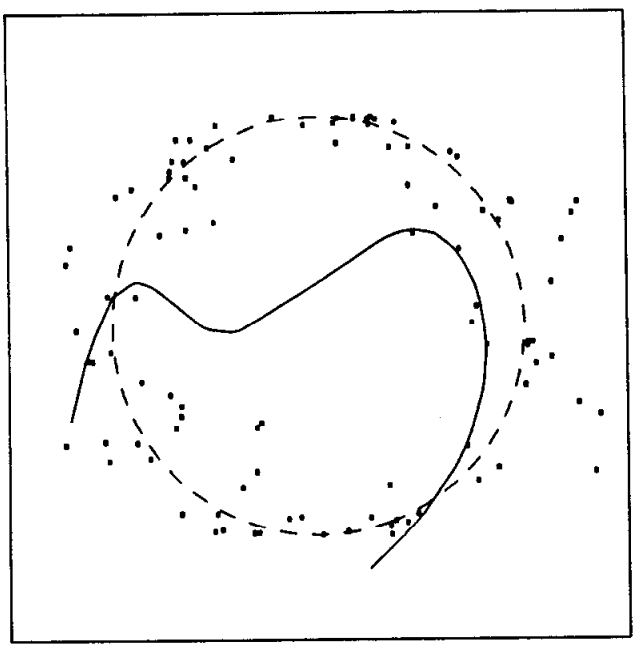

Figure 3.7g $D(\hat{f}^{(5)}) = 2.25$

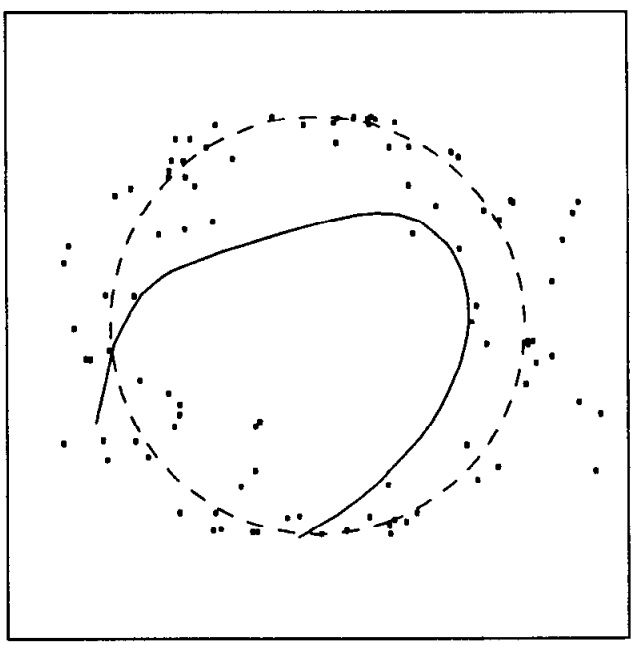

Figure 3.7h $D(\hat{f}^{(6)}) = 1.91$

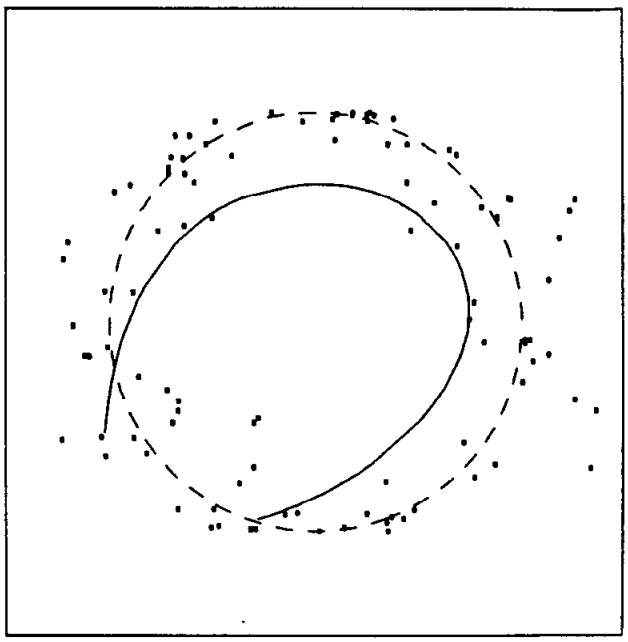

Figure 3.7i $D(\hat{f}^{(7)}) = 1.64$

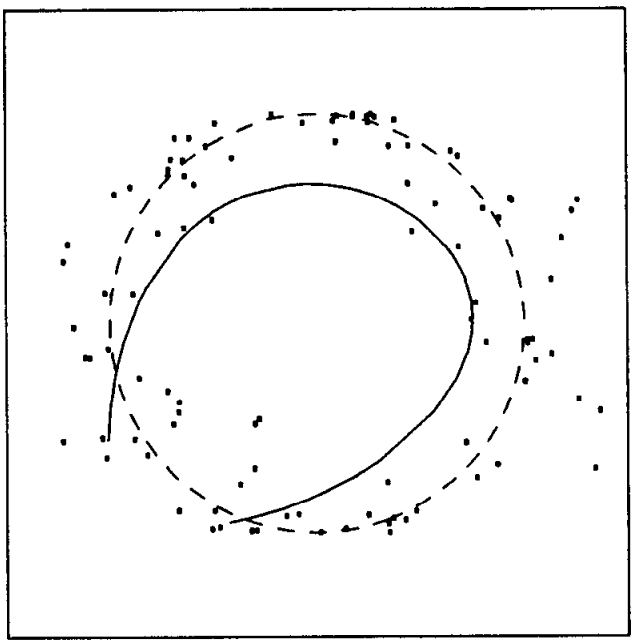

Figure 3.7j $D(\hat{f}^{(8)}) = 1.60$

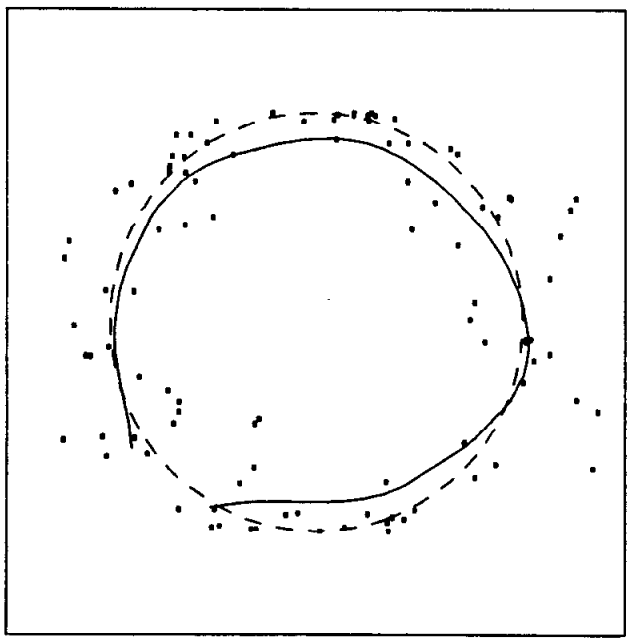

Figure 3.7k $D(\hat{f}^{(9)}) = 0.97$. The span is automatically reduced at this stage.

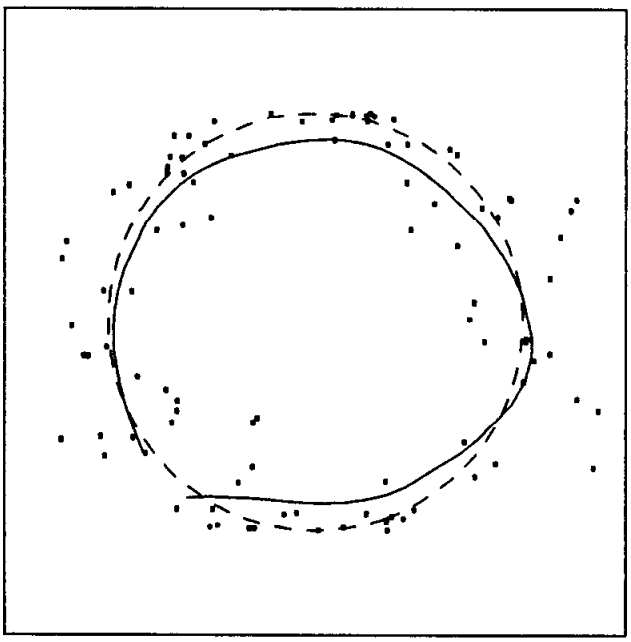

Figure 3.7l $D(\hat{f}^{(10)}) = 0.96$

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

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

3.5.2. The half-sphere in three-space.

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

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

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

---

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

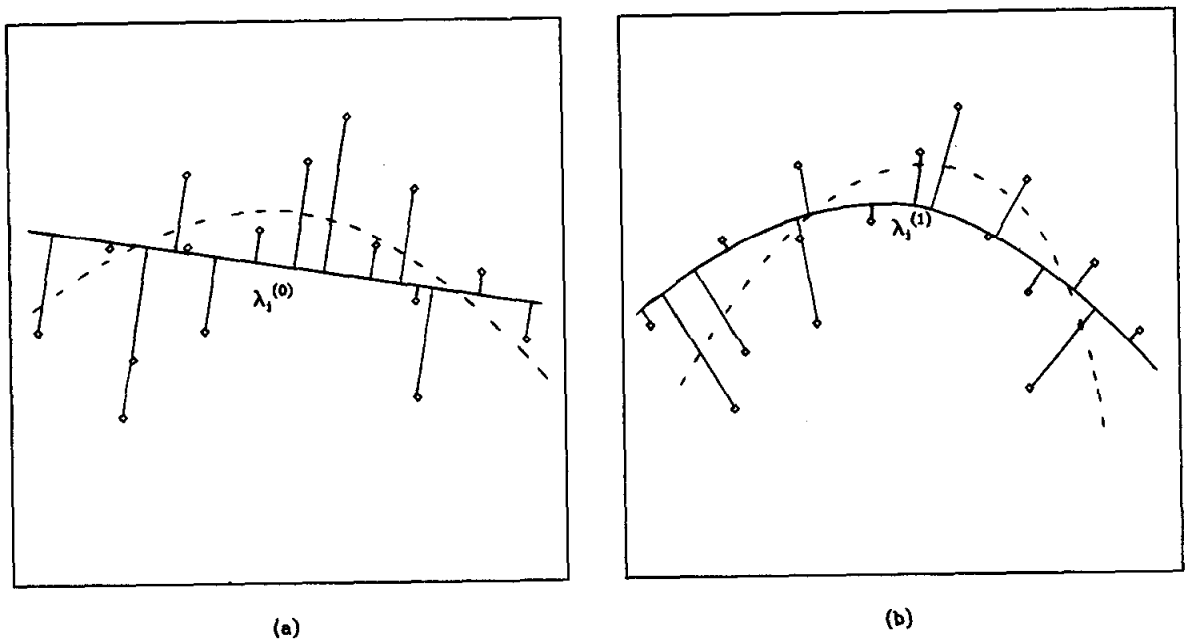

Figure 3.8 The curve of the the first iteration is a function of $\hat{\lambda}^{(0)}$ measured along the starting vector (a). The curve of the the second iteration is a function of $\hat{\lambda}^{(1)}$ measured along the curve of the first iteration (b).

3.6. Principal surfaces and principal components.

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

3.6.1. A Variance decomposition.

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

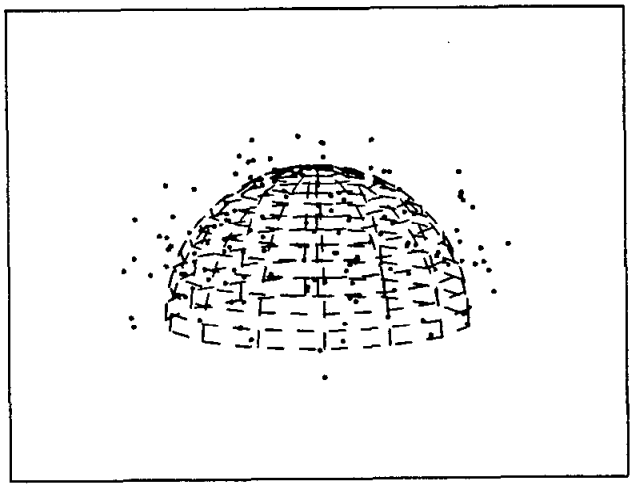

Figure 3.9a. The generating surface and the data. $D(S) = 1.0$

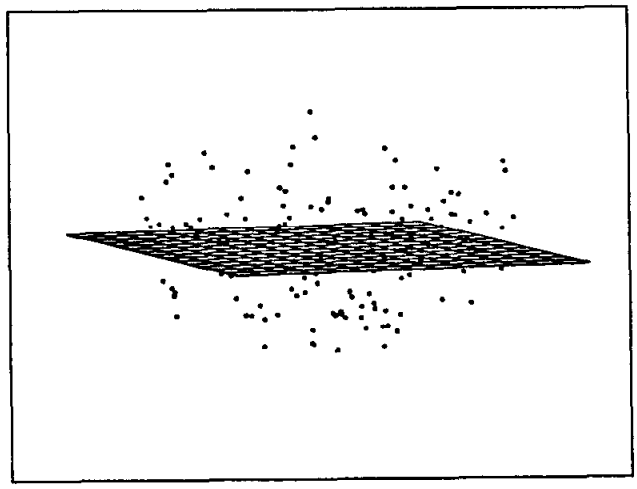

Figure 3.9b. The principal component plane. $D(\hat{f}^{(0)}) = 1.59$

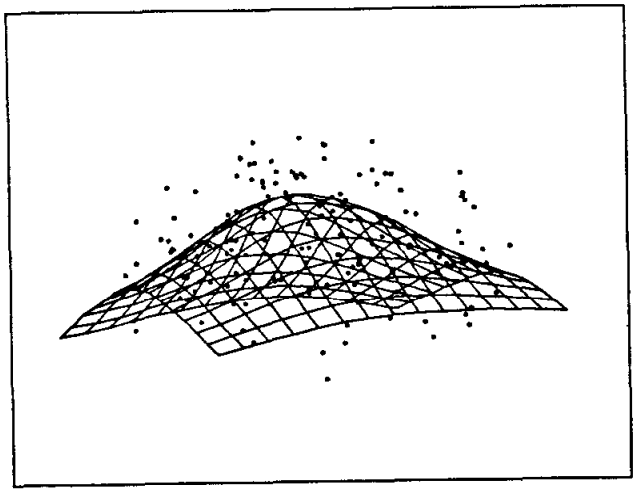

Figure 3.9c. $D(\hat{f}^{(1)}) = 1.20$

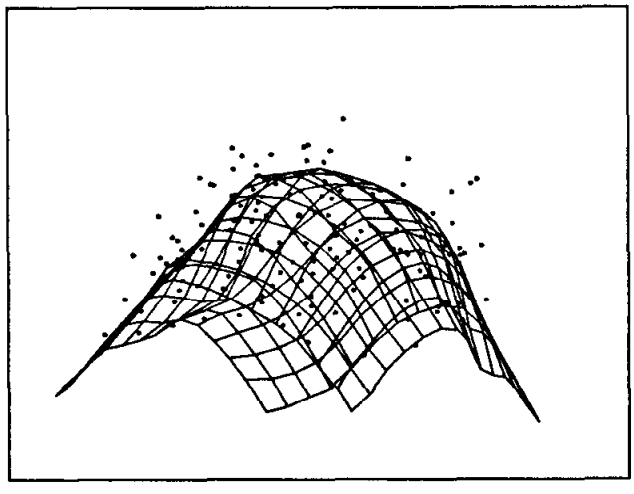

Figure 3.9d. $D(\hat{f}^{(4)}) = 0.78$

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

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

where $\mathbf{Var}(\cdot)$ 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 $\mathbf{Var}(\cdot)$ and $E(\cdot)$ have their usual meaning. The second term on the right of (3.12) is the expected squared distance of a point to its projection onto the principal direction.*

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

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

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

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

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

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

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

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

We make the following observations:

---

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

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

$$\sum_{j=1}^{p} \mathbf{Var}\left(f_j(\lambda_f(\boldsymbol{x}))\right) = \sum_{j=1}^{p} a_j^2 \mathbf{Var}\left(\lambda_{\boldsymbol{a}}(\boldsymbol{x})\right)$$
$$= \mathbf{Var}(\lambda)$$

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

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

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

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

3.6.2. The power method.

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

Theorem 3.1

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

Proof

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

$$\boldsymbol{\lambda}^{(0)} = X\boldsymbol{a}^{(0)}$$

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

$$a_j^{(1)} = \frac{\boldsymbol{\lambda}^{(0)\prime}\boldsymbol{x}_j}{\boldsymbol{\lambda}^{(0)\prime}\boldsymbol{\lambda}^{(0)}}$$

or

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

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

and $\boldsymbol{a}^{(1)}$ 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). ∎