Chapter 4

Theory for principal curves and surfaces

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

4.1. The projection index is measureable.

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

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

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

where

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

the usual euclidean distance between two vectors. Now set

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

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

$$\lambda_{\boldsymbol{f}}(\boldsymbol{x}) = \sup M(\boldsymbol{x})$$

$\lambda_{\boldsymbol{f}}(\boldsymbol{x})$ is attained because $M(\boldsymbol{x})$ is closed, and we have avoided ambiguities.

Theorem 4.1

$\lambda_{\boldsymbol{f}}(\boldsymbol{x})$ is a measureable function of $\boldsymbol{x}$.

---

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

Proof

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

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

$$\begin{aligned}\{x \mid \lambda_f(x) \le c\} &= \bigcap_{\lambda \in (c,\lambda_1]} \bigcup_{\lambda' \in [\lambda_0, c]} \{x \mid d(x, f(\lambda)) > d(x, f(\lambda'))\} \\ &\stackrel{\text{def}}{=} A_c\end{aligned}$$

The first step in the proof is to show that

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

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

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

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

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

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

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

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

The second step is to show that

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

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

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

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

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

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

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

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

4.2. The stationarity property of principal curves.

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

We first establish some more notation. Suppose $f(\lambda) : \Lambda \mapsto \mathcal{G}$ is a unit speed continuously differentiable parametrized curve in $\mathbb{R}^p$, where Λ is an interval in $\mathbb{R}^1$. Let $g(\lambda)$ be defined similarly, without the unit speed restriction. An ϵ perturbed version of f is $f_\epsilon \stackrel{\text{def}}{=} f(\lambda) + \epsilon g(\lambda)$. Suppose X has a continuous density in $\mathbb{R}^p$ which we denote by h, and let $D^2(h, f_\epsilon)$ be defined as before by

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

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

Definition

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

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

(We have to show that this derivative exists.)

Theorem 4.2

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

Note:

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

Proof

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

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

Then

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

Upon taking expected values we get

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

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

We split $\boldsymbol{v}$ into a component $\boldsymbol{v}_e = c\boldsymbol{e}_1$ along $\boldsymbol{e}_1$ and an orthogonal component $\boldsymbol{v}^*$. Thus $\boldsymbol{v} = c\boldsymbol{v}_e + \boldsymbol{v}^*$ where $\boldsymbol{e}_1'\boldsymbol{v}^* = 0$. So $\boldsymbol{f}_\epsilon = \lambda((1+c\epsilon)\boldsymbol{e}_1 + \epsilon\boldsymbol{v}^*)$. We now plug this into (4.1) to get

$$\begin{aligned} D^2(h, \boldsymbol{f}_\epsilon) &= \text{tr}\ \Sigma - \frac{((1+c\epsilon)\boldsymbol{e}_1 + \epsilon\boldsymbol{v}^*)'\Sigma((1+c\epsilon)\boldsymbol{e}_1 + \epsilon\boldsymbol{v}^*)}{(1+c\epsilon)^2 + \epsilon^2} \\ &= \text{tr}\ \Sigma - \frac{(1+c\epsilon)^2\boldsymbol{e}_1'\Sigma\boldsymbol{e}_1 + 2\epsilon(1+c\epsilon)\boldsymbol{e}_1'\Sigma\boldsymbol{v}^* + \epsilon^2\boldsymbol{v}^{*\prime}\Sigma\boldsymbol{v}^*}{(1+c\epsilon)^2 + \epsilon^2} \end{aligned} \tag{4.2}$$

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

$$\left.\frac{dD^2(h, \boldsymbol{f}_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = -2\boldsymbol{e}_1'\Sigma\boldsymbol{v}^*.$$

If $\boldsymbol{e}_1$ is a principal component of Σ then this term is zero for all $\boldsymbol{v}^*$ and hence for all $\boldsymbol{v}$. Alternatively, if this term, and hence the derivative, is zero for all $\boldsymbol{v}$ and hence all $\boldsymbol{v}^{*\prime}\boldsymbol{e}_1 = 0$, we have

$$\begin{aligned} &\boldsymbol{v}^{*\prime}\Sigma\boldsymbol{e}_1 = 0 \ \ \forall\ \boldsymbol{v}^{*\prime}\boldsymbol{e}_1 = 0 \\ \Rightarrow &\Sigma\boldsymbol{e}_1 = c\boldsymbol{e}_1 \\ \Rightarrow &\boldsymbol{e}_1 \text{ is an eigenvector of } \Sigma \end{aligned}$$

∎

Note:

Suppose $\boldsymbol{v}$ is in fact another eigenvector of Σ, with eigenvalue d, then

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

This shows that $\boldsymbol{f}$ might be a maximum, a minimum or a saddle point.

Theorem 4.3

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

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

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

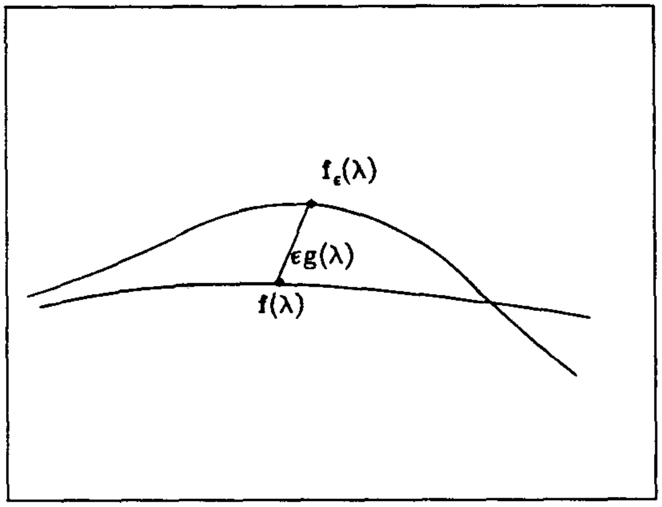

Figure (4.1) $f_\epsilon(\lambda)$ depicted as a function of $f(\lambda)$.

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

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

Proof of theorem 4.3.

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

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

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

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

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

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

if this limit exists. Now

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

Expanding the first norm we get

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

and thus

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

where Y_1 is some bounded random variable.

Similarly we have

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

We expand the first norm again, and get

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

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

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

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

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

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

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

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

Alternatively, suppose that

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

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

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

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

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

Corollary

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

Proof

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

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

Lemma 4.3.1

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

Proof

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

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

Lemma 4.3.2

The set of ambiguity points has probability measure zero.

Proof

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

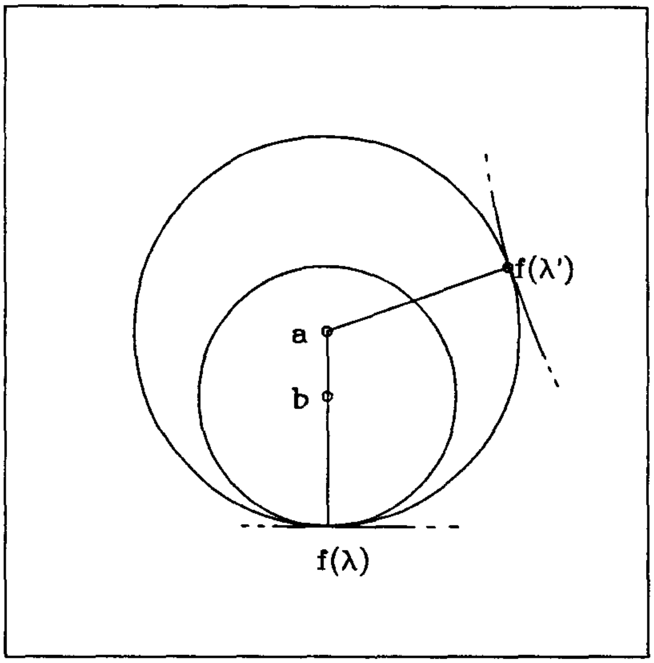

Figure 4.2 There are at most two ambiguity points on the normal to the curve; one on either side of the curve.

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

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

Corollary

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

Proof

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

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

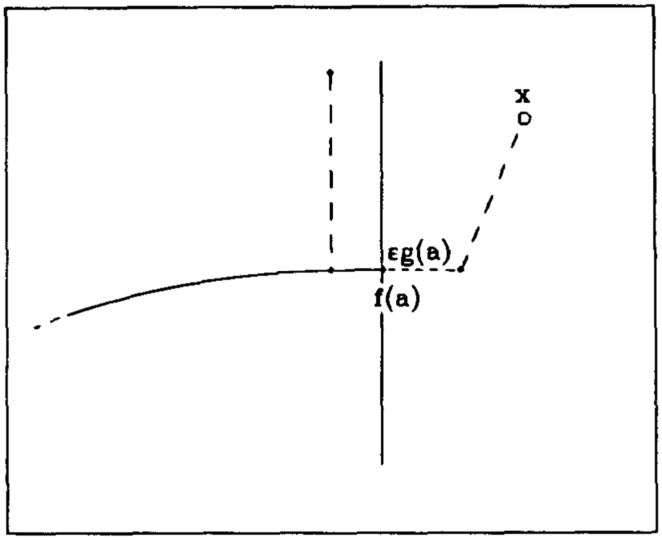

Figure 4.3 The set of points to the right of $f(a)$ that project there has measure zero.

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

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

Lemma 4.3.3

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

Proof

If $\lambda_f(x) \in (a,b)$ the proof is immediate. Suppose then that $\lambda_f(x) = a$. Rotate the co-ordinates so that $f'(a) = e_1$. No points to the left of $f(a)$ project there. Suppose f is a principal curve. This then implies that the set of points that are to the right of $f(a)$ and project at $f(a)$ has conditional measure zero, else the conditional expectation would be to the right. Thus they also have unconditional measure zero.

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

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

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

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

Theorem 4.4

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

Proof

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

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

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

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

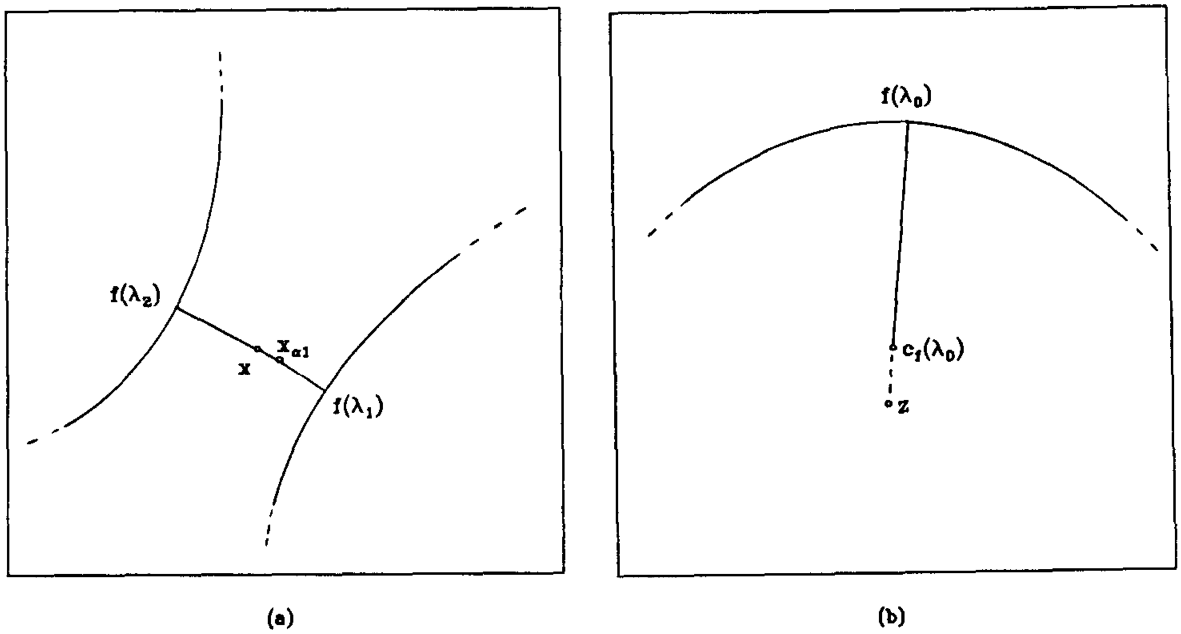

Figure 4.4 The continuity constraint avoids global ambiguities (a) and local ambiguities (b) in projection.

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

Theorem 4.5

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

Proof

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

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

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

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

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

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

We evaluate these two conditions:

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

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

which contradicts (2) above. ∎

4.4. Some results on bias.

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

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

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

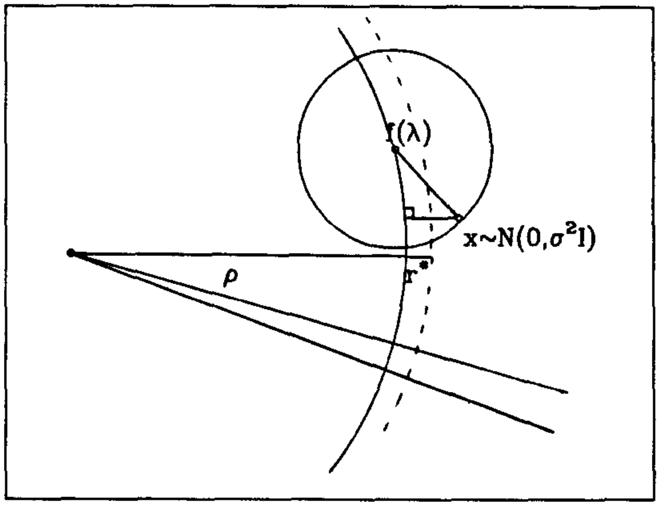

Figure 4.5 The data is generated from the arc of a circle with radius ρ and with iid $\mathcal{N}(0, \sigma^2 I)$ errors. The location on the circle is selected uniformly.

4.4.1. A simple model for investigating bias.

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

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

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

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

One would hope that the principal curve procedure, operating in distribution space and starting at the true curve, would converge to this minimizing distance circle in this idealized situation. It turns out that this is indeed the case.

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

Theorem 4.6

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

Then

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

where

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

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

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

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

Lemma 4.6.1

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

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

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

---

* I thank Art Owen for suggesting this result.

Proof of lemma 4.6.1

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

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

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

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

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

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

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

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

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

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

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

which is non-negative.

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

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

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

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

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

This yields an expected squared distance of

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

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

Proof of theorem 4.6.

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

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

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

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

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

Also, we have

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

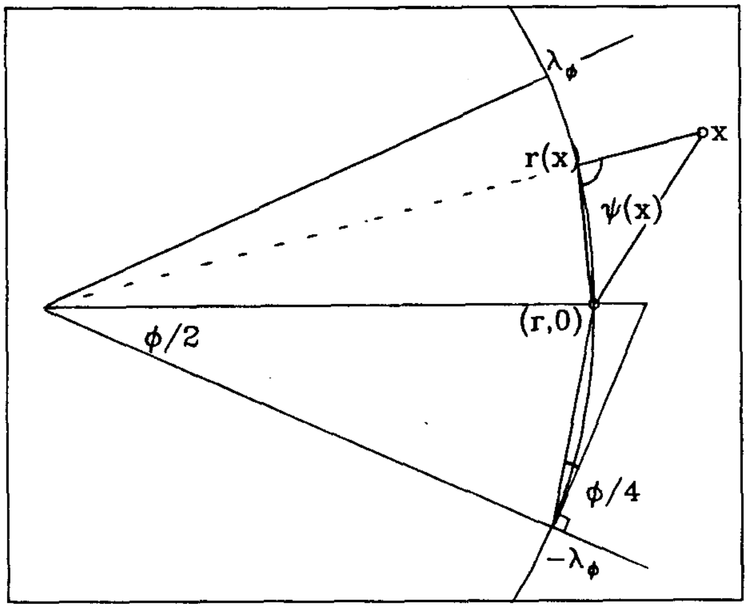

Figure 4.6 The conditional expectation of x, given $\lambda_f(x) \in \Lambda_\phi$.

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

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

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

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

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

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

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

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

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

$$\begin{aligned}\lim_{\phi \mapsto 0} \mathbf{E}\left(x \mid \lambda_f(x) \in \Lambda_\phi\right) &= \ \mathbf{E}\left(x \mid \lambda_f(x) = 0\right) \\ &= r^*\end{aligned}$$

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

$$\mathbf{E}\left(x \mid \lambda_f(x) \in \Lambda_\theta\right) = \ \mathbf{E}_{\lambda_f(x) \in \Lambda_\theta}\ \mathbf{E}\left(x \mid \lambda_f(x) = \lambda\right)$$

Suppose λ corresponds to an angle z, then

$$\mathbf{E}\left(x \mid \lambda_f(x) = \lambda\right) = \begin{pmatrix} r^* \cos(z) \\ r^* \sin(z) \end{pmatrix}$$

Thus

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

∎

Corollary

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

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

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

Corollary

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

Proof

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

4.4.2. From the circle to the helix.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.4.3. One more bias demonstration.

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

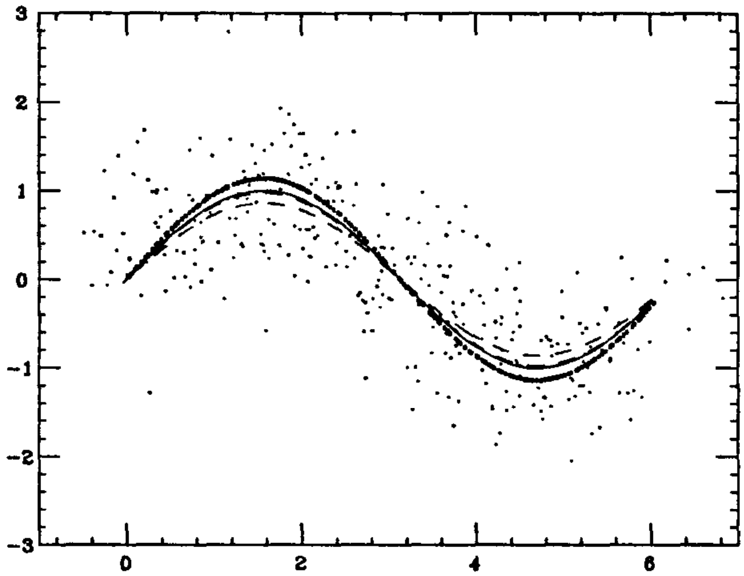

Figure 4.7 The thick curve is the the principal curve using conditional expectations at the model, and shows the *model bias*. The two dashed curves show the compounded effect of model and *estimation* bias at spans of 30% and 40%.

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

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

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

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

and achieves a minimum radius of 1 unit. The model for the data is $X = f(\lambda) + e$ where $\lambda \sim U[0, 2]$ and $e \sim \mathcal{N}(0, I/4)$ independent of λ. Figure 4.7shows the true model (solid curve), and the points are a sample from the model, included to give an idea of the error structure. The thick curve is $\mathbf{E}(X \mid \lambda_f(X) = \lambda)$. Here is a situation where the model bias results in a curve with more curvature, namely a minimum radius of 0.88 units. This curve was found by simulation, and is well approximated by $1/0.88\sin(\lambda\pi)$. There are two dashed curves in the figure. They represent $\mathbf{E}(X \mid \lambda_f(X) \in \Lambda_s(\lambda))$, where $\Lambda_s(\lambda)$ represents a symmetric interval of length $s\Lambda$ about λ (Boundary effects were eliminated by cyclically extending the range of λ.) We see that at $s = 30\%$ the estimation bias approximately cancels out the model bias, whereas at $s = 40\%$ there is a residual estimation bias.

4.5. Principal curves of elliptical distributions.

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

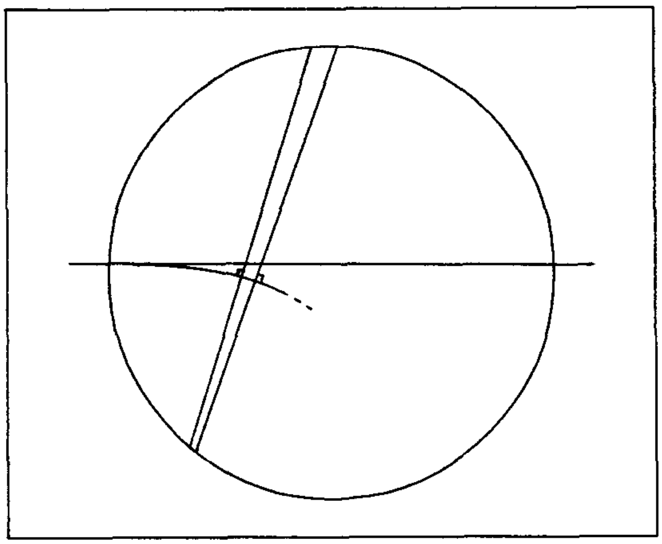

Figure (4.8) The only principal curves in $\mathcal{F}_c(h)$ of a uniform disk are the principal components.

Proposition

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

An informal proof of this claim is as follows:

- Any principal curve must enter the disk once and leave it once. This must be true since if it were to remain inside it would have to circle around. But this would violate the continuity constraint imposed by $\mathcal{F}_c(h)$ since there would have to exist points at the centers of curvature of the curve at some places. Furthermore, it cannot end inside the disk for reasons similar to those used in lemma 4.3.3.

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

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

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

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

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