Geometric Properties of Principal Curves in the Plane

Tom Duchamp
Department of Mathematics

Werner Stuetzle
Department of Statistics

University of Washington
Seattle, WA 98195

August 22, 1995

This research was supported by DOE grant DE-FG06-85-ER25006 and NSF grants CCR-8957323 and DMS-9103002

1991 Mathematics Subject Classification. Primary 62G07; secondary 62J02, 62H25, 62H30

Key words and phrases. Principal curves, least-squares, curve fitting, nonlinear regression.

Abstract

Principal curves were introduced to formalize the notion of "a curve passing through the middle of a dataset". Vaguely speaking, a curve is said to pass through the middle of a dataset if every point on the curve is the average of the observations projecting onto it. This idea can be made precise by defining principal curves for probability densities. Principal curves can be regarded as a generalization of linear principal components — if a principal curve happens to be a straight line, then it is a principal component. In this paper we study principal curves in the plane. We show that principal curves are solutions of a differential equation. By solving this differential equation, we find principal curves for uniform densities on rectangles and annuli. There are oscillating solutions besides the obvious straight and circular ones, indicating that principal curves in general will not be unique. If a density has several principal curves, they have to cross, a property somewhat analogous to the orthogonality of principal components. Finally, we study principal curves for spherical and elliptical distributions.

1 Introduction

The problem of fitting one or two-dimensional manifolds to point sets in two, respectively three, dimensions occurs in a variety of contexts, such as modeling of object boundaries in two or three-dimensional images (Banfield and Raftery (1992), Brinkley (1985), Martin et al (1993), Schudy and Ballard (1978, 1979), Sheehan et al (1992)), and reconstruction of objects from range data (Fang and Gossard (1992), Goshtasby (1992), Hoppe et al (1992, 1993), Muraki (1991), Solina and Bajcsy (1990), Vemuri et al (1986)). It is typically not reasonable to assume that the unknown manifold is of a simple parametric form, like an ellipsoid. Fitting methods need to be flexible and able to accommodate a variety of shapes.

Manifold fitting is fundamentally different from regression. It is worthwhile to contrast the respective goals for the case of two-dimensional data. In regression, we are given points $(x_1, y_1), \ldots, (x_n, y_n)$. The goal is to find a *function* $f(x)$ summarizing the dependence of the response variable Y on the predictor variable X; the two variables are thus treated asymmetrically. Under the assumption that f is linear, a common choice is the least squares straight line. There has been a large amount of research on nonparametric regression methods that make only very general assumptions about the nature of f.

In manifold fitting, we are also given points $\mathbf{x}_1, \ldots \mathbf{x}_n \in \mathbf{R}^2$. The goal is to find a *one-dimensional manifold* Γ summarizing the association between the variables X_1 and X_2. The two variables are treated symmetrically. It is usually clear from the context whether the manifold should be topologically a circle or a closed interval. Under the assumption that Γ is a straight line, a common choice is the largest principal component. Nonparametric methods not relying on the linearity assumption have typically been crafted on an ad-hoc basis. This is not satisfactory, and a theoretical underpinning would be desirable.

To statistically analyze the behavior of fitting methods we need a stochastic model that is thought to give rise to the data. A simple model is to assume that the data points are i.i.d. observations of a two-dimensional random vector distributed according to some (unknown) density p. Then two questions arise: (1) Which characteristic of the density are we trying to estimate, and why? (2) How can we estimate this characteristic?

In the context of regression, answers to those questions are well understood. We usually estimate the conditional expectation $\mathrm{E}(Y \mid x)$, because it minimizes the expected squared prediction error $\mathrm{E}(Y - f(X))^2$ among all functions f. There are many approaches to estimating $\mathrm{E}(Y \mid x)$, often based on local averaging. The amount of averaging or, more generally, the complexity of the model is typically chosen to minimize an estimate of expected squared prediction error, like the cross-validated residual sum of squares.

In the context of manifold estimation the situation is not as clear cut — there are no generally accepted answers to questions (1) and (2) above. In this paper we discuss answers based on the concept of *principal curves*. Principal curves were introduced in Hastie (1984) and Hastie and Stuetzle (1989) to formalize the notion of "a curve passing through the middle of a dataset". Vaguely speaking, a curve Γ is said to pass through the middle of a dataset if every point $\mathbf{x}$ on the curve is the average of the observations projecting onto it.

To make this idea precise, Hastie and Stuetzle (1989) define principal curves for probability densities.

Definition of principal curves: Let $\mathbf{X}$ denote a two-dimensional random vector distributed according to a probability density p, and let $\Gamma \subset \mathbf{R}^2$ be a smoothly embedded closed interval (*arc*) or circle (*loop*). For each point $\mathbf{x} \in \mathbf{R}^2$, let $d(\mathbf{x}, \Gamma)$ denote the distance from $\mathbf{x}$ to Γ. Because Γ is compact, for each $\mathbf{x} \in \mathbf{R}^2$ the distance $d(\mathbf{x}, \Gamma)$ is realized by at least one point of Γ. Of course, there may be several such points; a point $\mathbf{x}$ with several closest points on the curve is called an *ambiguity point*. The *projection map*

$$\pi_\Gamma : \mathbf{R}^2 \to \Gamma \,,$$

is the map which assigns to each $\mathbf{x} \in \mathbf{R}^2$ a point $\pi_\Gamma(\mathbf{x}) \in \Gamma$ realizing the distance from $\mathbf{x}$ to Γ, i.e.

$$d(\mathbf{x}, \Gamma) = \|x - \pi_\Gamma(\mathbf{x})\| \,.$$

Notice that the map π_Γ is well-defined only on the complement of the set of ambiguity points of Γ. But the set of ambiguity points has Lebesgue measure zero (see Hastie and Stuetzle (1989)) and can be ignored in probability calculations. It is not difficult to show that π_Γ is continuous on the complement of the set of ambiguity points.

The vague notion that every point on the curve should be the average of the observations projecting onto it can now be formalized:

Definition 1 *(Hastie and Stuetzle (1989)) A curve* Γ *is called* self-consistent *or a* principal curve *of a density* p *if* $E(\mathbf{X} \mid \pi_\Gamma(\mathbf{X}) = \mathbf{x}) = \mathbf{x}$ *for almost every* $\mathbf{x} \in \Gamma$.

The notion of projection also leads to a natural definition of the distance between a random vector $\mathbf{X}$ or its associated density, and a curve Γ:

$$d^2(\mathbf{X}, \Gamma) = \mathrm{E}(\|\mathbf{X} - \pi_\Gamma(\mathbf{X})\|^2)\,.$$

Principal curves as generalizations of linear principal components: Besides formalizing the notion of "a curve passing through the middle of a dataset", principal curves share two properties with linear principal components, which make them appear as a natural generalization (Hastie and Stuetzle (1989)):

- If a principal curve happens to be a straight line, then it is a (linear) principal component.
- Principal curves are critical points of the distance function in the variational sense: let Γ be a principal curve, and let Γ_t be a smooth family of curves with $\Gamma_0 = \Gamma$, then

  $$\frac{d}{dt} d^2(\mathbf{X}, \Gamma_t)\Big|_{t=0} = 0$$

  Linear principal components share this property if Γ_t is restricted to be a smooth family of straight lines. The largest principal component minimizes the distance to $\mathbf{X}$, the smallest principal component maximizes the distance (among all lines passing through the mean), and the others are saddlepoints.

Summary of results: The goal of this paper is to further contribute to the theoretical understanding of principal curves. We now present an informal synopsis of our results:

1. Suppose that $\Gamma \subset \Omega$ is a principal arc for a density p with compact support Ω. Then Γ satisfies the following *transversality conditions*:

(1) the endpoints of Γ lie on the boundary of Ω; (2) Γ intersects the boundary orthogonally; (3) the endpoints of Γ are (weakly) convex points of Ω.

2. If Γ_1 and Γ_2 both are principal curves for a density p, they cannot be linearly separable.

3. Suppose that Γ is a principal curve for a density p. Under appropriate conditions on Γ and p, the curvature of Γ can be expressed in terms of certain conditional moments of p. Principal curves that satisfy these conditions are called *regular*.

4. Regular principal curves are solutions of a system of ordinary differential equations. These equations can be used to calculate principal curves for uniform densities on rectangles and annuli. There are oscillating solutions besides the obvious straight and circular ones, indicating that principal curves in general will not be unique.

5. Any two regular principal curves of a density intersect.

6. The only regular principal arcs for a radially symmetric density are straight lines through the center.

7. While the uniform density on a circular disk has a circular principal curve, the uniform density on an ellipse does not have an elliptical principal curve.

Notation and conventions: The following notation is used throughout the paper: L_Γ denotes the length of Γ; Λ denotes either the closed interval $[0, L_\Gamma]$ when Γ is an arc, or the circle of circumference L_Γ when Γ is a loop; $\mathbf{x} = \mathbf{x}(s)$ denotes an arc length parameterization of Γ. The unit tangent and normal vector fields to Γ are written $\mathbf{T}(s)$ and $\mathbf{N}(s)$, respectively, and oriented so that the pair $(\mathbf{T}(s), \mathbf{N}(s))$ is consistent with the standard orientation of $\mathbf{R}^2$. The angle between the positive x-axis and $\mathbf{T}(s)$ is denoted by $\theta(s)$.

The map

$$\lambda : \begin{cases} \mathbf{R}^2 & \to \Lambda \\ \mathbf{x} & \mapsto s = \lambda(\mathbf{x}) \end{cases}$$

defined by the formula $\pi_\Gamma(\mathbf{x}) = \mathbf{x}(\lambda(\mathbf{x}))$ is called the *projection index* (see Hastie and Stuetzle (1989)).

Finally, throughout this paper $\Omega \subset \mathbf{R}^2$ denotes a compact, connected region with smooth boundary $\partial\Omega$. The density p is assumed to be supported on Ω, strictly positive on the interior of Ω, and smooth on all of Ω.

2 Transversality Conditions

The goal of this section is to study the endpoints of principal arcs.

Lemma 1. *If $\Gamma \subset \Omega$ is a principal curve of p, then the line connecting any point $\mathbf{x} \in \Omega$ to its projection $\pi_\Gamma(\mathbf{x})$ intersects Γ orthogonally:*

$$\langle (\mathbf{x} - \pi_\Gamma(\mathbf{x})), \mathbf{T}(\pi_\Gamma(\mathbf{x})) \rangle = 0$$

for all $\mathbf{x} \in \Omega$.

Proof. Choose a point $\mathbf{x} \in \Omega$. If $\pi_\Gamma(\mathbf{x})$ is an interior point of Γ then the identity follows from the fact that $\pi_\Gamma(\mathbf{x})$ is a point of Γ realizing the distance between $\mathbf{x}$ and Γ.

Suppose that $\mathbf{x}_0 = \pi_\Gamma(\mathbf{x})$ is an endpoint of Γ. Without loss of generality, we may assume that the orientation of Γ has been chosen so that $\mathbf{T}_0 \equiv \mathbf{T}(\mathbf{x}_0)$ is outward-pointing. Note that the inequality $\langle (\mathbf{y} - \mathbf{x}_0), \mathbf{T}_0 \rangle \geq 0$ is satisfied for all $\mathbf{y} \in \Omega$ such that $\pi_\Gamma(\mathbf{y}) = \mathbf{x}_0$. For otherwise, the distance from $\mathbf{y}$ to Γ would be strictly less than $\|\mathbf{y} - \mathbf{x}_0\|$. Consequently, the subset $\pi_\Gamma^{-1}(\mathbf{x}_0) \cap \Omega$ is contained in the half-plane $H = \{\mathbf{x} \in \mathbf{R}^2 \mid \langle (\mathbf{x} - \mathbf{x}_0), \mathbf{T}_0 \rangle \geq 0\}$.

Suppose that the proposition is false. Then $\langle (\mathbf{y} - \mathbf{x}_0), \mathbf{T}_0 \rangle > 0$ for some point $\mathbf{y} \in \Omega$ with $\pi(\mathbf{y}) = \mathbf{x}_0$. Since p is continuous on Ω and strictly positive on the interior of Ω, there is an open set $Q \subset H$, such that (i) $\pi_\Gamma(Q) = \mathbf{x}_0$, (ii) $p > 0$ on Q and (iii) $\langle (\mathbf{y} - \mathbf{x}_0), \mathbf{T}_0 \rangle > 0$ for all $\mathbf{y} \in Q$. But these conditions together imply that the point $\mathrm{E}(\mathbf{X} \mid \pi_\Gamma(\mathbf{X}))$ is contained in the interior of H, violating the self-consistency condition $\mathrm{E}(\mathbf{X} \mid \pi_\Gamma(\mathbf{X}) = \mathbf{x}_0) = \mathbf{x}_0$ ($\mathbf{x}_0$ is on the boundary of H). ▌

The following proposition is an immediate consequence of Lemma 1:

Proposition 1. *If $\Gamma \subset \Omega$ is a principal arc then Γ satisfies the following transversality conditions:*

1. *the endpoints of Γ lie on the boundary $\partial\Omega$;*
2. *Γ intersects $\partial\Omega$ orthogonally;*
3. *the endpoints of Γ are (weakly) convex points of $\partial\Omega$.*

3 Normal coordinates

Suppose that Γ is a principal curve for p and that there are no ambiguity points in the support of p. To interpret self-consistency as a curvature condition we need the notion of *normal coordinates*.

Definition 2 . *The* normal coordinate map *of* Γ *is the map* $\nu_\Gamma : \Lambda \times \mathbf{R} \to \mathbf{R}^2$ *defined by the formula*

$$\nu_\Gamma(s, v) = \mathbf{x}(s) + v\mathbf{N}(s)$$

and the normal coordinate transformation *is the map* $\mu_\Gamma : \Omega \to \Lambda \times \mathbf{R}$ *defined by the formula*

$$\mu_\Gamma(\mathbf{x}) = (\lambda(\mathbf{x}), \langle \mathbf{x} - \mathbf{x}(\lambda(\mathbf{x})), \mathbf{N}(\lambda(\mathbf{x}))\rangle) .$$

The components (s, v) *of* $\mu_\Gamma(\mathbf{x})$ *are called the* normal coordinates *of* $\mathbf{x}$.

By virtue of our assumption that Ω does not contain ambiguity points of Γ, the normal map is a left inverse of the normal coordinate transformation μ_Γ:

$$\nu_\Gamma \circ \mu_\Gamma = id_\Omega.$$

We can now state a formal definition of regularity:

Definition 3 . *A smooth curve* $\Gamma \subset \Omega$ *is called* regular *if the following conditions are satisfied:*

1. Ω *contains no ambiguity points of* Γ.
2. *The map* $\mu_\Gamma : \Omega \to \Lambda \times \mathbf{R}$ *is a diffeomorphism onto its image.*
3. *The image* $\mu_\Gamma(\Omega)$ *is of the form*

   $$\mu_\Gamma(\Omega) = \{(s, v) \in \Lambda \times \mathbf{R} \mid v_-(s) \leq v \leq v_+(s)\},$$

   where v_- *and* v_+ *are smooth and* $v_-(s) < 0 < v_+(s)$ *on the interior of* Λ.

Remark 1 *Regularity condition (3) implies that for all* $\mathbf{x}$ *on the boundary of* Ω *the line segment joining* $\mathbf{x}$ *to* $\pi_\Gamma(\mathbf{x})$ *is not tangential to the boundary at* $\mathbf{x}$.

Figure 1 shows a regular arc and a regular loop. Regularity implies that for any $s \in \Lambda$ the set $\mathcal{V}(s) = \{v \mid (s,v) \in \mu_\Gamma(\Omega)\}$ of points in Ω projecting onto $\mathbf{x}(s)$ is an interval:

$$\mathcal{V}(s) = \{\mathbf{x}(s) + v\mathbf{N}(s) \mid v_-(s) < v < v_+(s)\}.$$

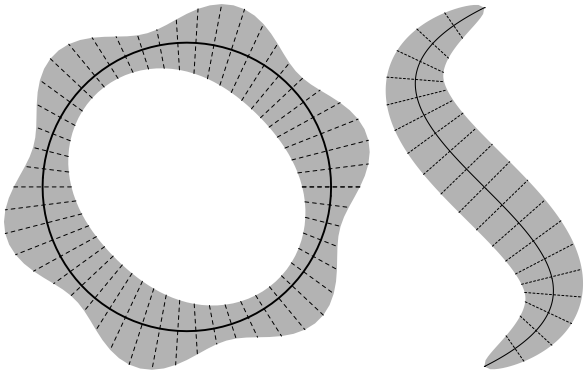

Figure 1: A regular loop and a regular arc.

For later reference, we now calculate the Jacobian determinant of the normal map. Recall that the *curvature function* $\kappa = \kappa(s)$ is given by the formula

$$\kappa = \frac{d\theta}{ds}.$$

Let $\imath$ and $\jmath$ denote the standard unit vectors in $\mathbf{R}^2$. Differentiation with respect to s of the identities

$$\mathbf{T} = \cos(\theta)\imath + \sin(\theta)\jmath \text{ and } \mathbf{N} = -\sin(\theta)\imath + \cos(\theta)\jmath$$

yields the *Frenet formulas*

$$\frac{d\mathbf{T}}{ds} = \kappa\mathbf{N} \quad \text{and} \quad \frac{d\mathbf{N}}{ds} = -\kappa\mathbf{T}.$$

Then $\dfrac{\partial \nu_\Gamma(s,v)}{\partial s} = \mathbf{x}'(s) + v\mathbf{N}'(s) = (1 - v\kappa(s))\,\mathbf{T}(s)$ and $\dfrac{\partial \nu_\Gamma(s,v)}{\partial v} = \mathbf{N}(s)$. The Jacobian determinant $\dfrac{\partial(x,y)}{\partial(s,v)}$ of the normal coordinate map is now

easily computed:

$$\frac{\partial(x,y)}{\partial(s,v)} = \left| \frac{\partial \nu_\Gamma(s,v)}{\partial s} \times \frac{\partial \nu_\Gamma(s,v)}{\partial v} \right| = |(1 - v\kappa(s))\mathbf{T}(s) \times \mathbf{N}(s)| = 1 - v\kappa(s)\,. \tag{1}$$

4 Self-consistency as a Curvature Condition

Suppose that Γ is a regular principal curve for a density p. We will now derive a relation between the curvature of Γ and certain conditional moments of p. We do this by rewriting the self-consistency condition in normal coordinates. Recall that Γ is called self-consistent if $\mathrm{E}(\mathbf{X} \,|\, \pi_\Gamma(\mathbf{X}) = \mathbf{x}) = \mathbf{x}$ for almost all $\mathbf{x} \in \Gamma$. By definition of conditional expectation, this is equivalent to the condition that

$$\int_{\pi_\Gamma^{-1}(A)} \mathbf{x}\, p(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int_{\pi_\Gamma^{-1}(A)} \pi_\Gamma(\mathbf{x})\, p(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

or

$$\int_{\pi_\Gamma^{-1}(A)} (\mathbf{x} - \pi_\Gamma(\mathbf{x}))\, p(\mathbf{x})\, \mathrm{d}\mathbf{x} = 0 \tag{2}$$

for all measurable $A \subset \Gamma$. Because Γ is assumed to be regular, Equation (2) can be rewritten in normal coordinates:

$$\int_{\{(s,v) \in \mu_\Gamma(\Omega) \,|\, s \in A\}} v\, p(\mathbf{x}(s) + v\mathbf{N}(s))\, \frac{\partial(x,y)}{\partial(s,v)}\, \mathrm{d}v\, \mathrm{d}s = 0\,,$$

where A now denotes a measurable subset of Λ. This implies that

$$\int_{\mathcal{V}(s)} v\, p(\mathbf{x}(s) + v\mathbf{N}(s))\, \frac{\partial(x,y)}{\partial(s,v)}\, \mathrm{d}v = 0 \qquad s - \text{a.e.}$$

Substituting (1) for the Jacobian of the normal map gives

$$\int_{\mathcal{V}(s)} v\, p(\mathbf{x}(s) + v\mathbf{N}(s))\, \mathrm{d}v - \kappa(s) \int_{\mathcal{V}(s)} v^2\, p(\mathbf{x}(s) + v\mathbf{N}(s))\, \mathrm{d}v = 0\,, \tag{3}$$

for all $s \in \Lambda$.

Let

$$\mu_\perp(s) = \frac{\int_{\mathcal{V}(s)} v\, p(\mathbf{x}(s) + v\mathbf{N}(s))\, \mathrm{d}v}{\int_{\mathcal{V}(s)} p(\mathbf{x}(s) + v\mathbf{N}(s))\, \mathrm{d}v}$$

denote the mean of the *transverse density* on $\mathcal{V}(s)$ induced by p, and let $\sigma_\perp^2(s)$ denote its variance. Equation (3) can then be rewritten as

$$\kappa(s) = \frac{\mu_\perp(s)}{\mu_\perp^2(s) + \sigma_\perp^2(s)}. \tag{4}$$

Equation (4) relates the curvature of Γ to first and second moments of the density induced on the normal line to the curve at s. Consider the case where $p(\mathbf{x})$ is uniform. If $\mathbf{x}(s)$ happens to coincide with the transverse mean, i.e. the center of the normal line segment, then the curvature $\kappa(s)$ has to vanish. Otherwise, the center of curvature is in the direction of the transverse mean. This makes intuitive sense: Consider an infinitesimal segment of the curve centered at $\mathbf{x}(s)$. As shown in Figure 2, the set of points projecting onto this segment is wedge shaped. If the center of the normal line segment falls below $\mathbf{x}(s)$, then the part of the wedge below the curve is longer than the part above the curve. In order for the mean of the segment to fall on the curve, the segment thus has to open up as we go upwards, meaning the curvature of Γ has to be negative.

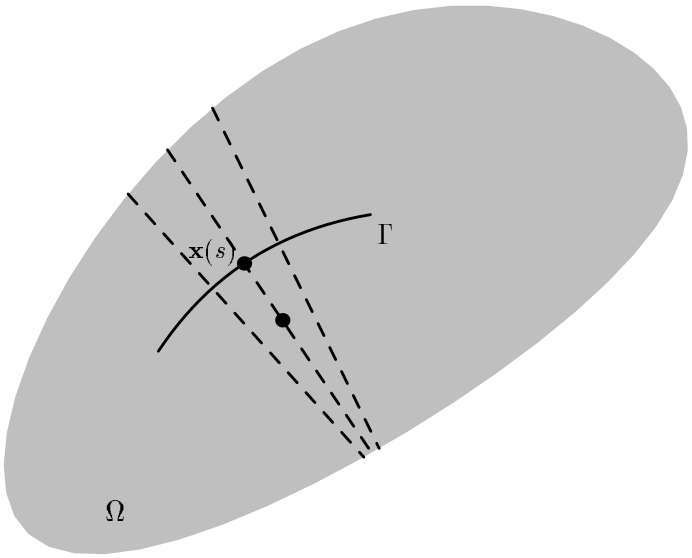

Figure 2: The center of curvature of a principal curve Γ points in the direction of the transverse mean.

5 Self-consistency as a Differential Equation

Equation (3) can be thought of as a differential equation satisfied by principal curves. Unfortunately, it cannot be used to find principal curves because the integration boundaries in the moment integrals themselves depend on the curve. In order to use a differential equation for finding principal curves, we proceed differently: We define a differential equation in such a way that its solution curves, *when they are regular*, are actually principal curves. This differential equation expresses curvature in terms of *transverse moments* of p.

Definition 4 . *The* k-th transverse moment *of the density* p *at* $(\mathbf{x}, \theta)$ *is the function* $\mu_k : \Omega \times S^1 \to \mathbf{R}$ *defined by the formula*

$$\mu_k(\mathbf{x}, \theta)) = \int_{v_-(\mathbf{x},\theta)}^{v_+(\mathbf{x},\theta)} v^k p(\mathbf{x} + v\mathbf{N}(\theta))\, dv\,.$$

The integration boundaries $v_+ = v_+(\mathbf{x}, \theta)$ *and* $v_- = v_-(\mathbf{x}, \theta)$ *are determined by the condition that* $\mathbf{x} + v_+\,\mathbf{N}$ *and* $\mathbf{x} + v_-\,\mathbf{N}$ *be the boundaries of the line segment around* $\mathbf{x}$ *obtained by intersecting the line* $\{\mathbf{x} + v\mathbf{N} : v \in \mathbf{R}\}$ *with the support* Ω *of* p. *This line segment is called the* transverse line segment *at* $(\mathbf{x}, \theta)$ *and is denoted by* $\ell(\mathbf{x}, \theta)$. *Similarly* $\mathcal{V}(\mathbf{x}, \theta) = \{v : v_-(\mathbf{x}, \theta) \le v \le v_+(\mathbf{x}, \theta)\}$ *is called the* transverse interval *of* Ω *at* $(\mathbf{x}, \theta)$.

The mean and variance of the *transverse density*

$$p^{\perp}_{(\mathbf{x},\theta)}(v) = \frac{p(\mathbf{x} + v\,\mathbf{N}(\theta))}{\mu_0(\mathbf{x}, \theta)}, \quad v_-(\mathbf{x}, \theta) \le v \le v_+(\mathbf{x}, \theta)$$

can be written in terms of transverse moments:

$$\begin{aligned} \mu_\perp(\mathbf{x}, \theta) &= \frac{\mu_1(\mathbf{x}, \theta)}{\mu_0(\mathbf{x}, \theta)} \\ \sigma^2_\perp(\mathbf{x}, \theta) &= \frac{\mu_2(\mathbf{x}, \theta)}{\mu_0(\mathbf{x}, \theta)} - \mu_\perp(\mathbf{x}, \theta)^2\,. \end{aligned}$$

Consider now the following system of first order differential equations, which we call the *principal curve equations*:

$$\frac{d\mathbf{x}}{ds} = \cos(\theta)\imath + \sin(\theta)\jmath\,; \quad \frac{d\theta}{ds} = \frac{\mu_1(\mathbf{x}, \theta)}{\mu_2(\mathbf{x}, \theta)} = \frac{\mu_\perp(\mathbf{x}, \theta)}{\mu_\perp(\mathbf{x}, \theta)^2 + \sigma^2_\perp(\mathbf{x}, \theta)}\,. \qquad (5)$$

Clearly, any solution of this system satisfies the self-consistency condition (4). This simple observation yields the following theorem.

Theorem 1 . *A regular curve* Γ *is a principal curve of the density* p *if and only if it is a solution of the system of equations (5).*

Remark 2 . It is worth noting that the system (5) may be singular along the boundary of Ω. Suppose that Γ is a principal arc. By virtue of the transversality conditions, the boundary of the support at the endpoints of Γ is convex. If it is strictly convex, all transverse moments μ_k vanish, and the self-consistency condition degenerates at the boundary.

6 Principal Curves for Uniform Densities

Our first example is the uniform density on the rectangular strip

$$\Omega_{a,b} = \{(x, y) \ : \ 0 \leq x \leq b \, , \quad -a/2 \leq y \leq a/2\}$$

of length b and width a.

For symmetry reasons, the horizontal line $y = 0$ and the vertical line $x = b/2$ are principal curves. If the region is square, then the same is true for the two diagonals. We will shortly see, however, that for long strips many other regular principal curves exist.

Without loss of generality, any regular principal curve for the strip can be assumed to be of the form $y = f(x)$. Curves of a more general type are excluded because they necessarily have ambiguity points in $\Omega_{a,b}$, and curves of the form $x = f(y)$ are dealt with by interchanging the x and y axes. Curves which enter and/or leave the strip are also excluded by the regularity requirement.

A solution to the system of differential equations (5) is uniquely determined by the boundary values $f(0)$ and $f'(0)$. To satisfy the transversality conditions, any principle curve has to intersect the left boundary of the rectangle orthogonally, meaning $f'(0) = 0$. The requirement that the support of the density must not contain any centers of curvature places a restriction on $f(0) = y_0$:

$$-\frac{a}{6} < y_0 < \frac{a}{6}$$

Ignore for the moment the fact that the rectangle has a right boundary and consider the infinite strip. We can obtain closed form expressions for the transverse moments and find explicit solutions of the system (5) in terms of elliptic functions (see Duchamp and Stuetzle (1993) for details). Solutions

are periodic, and the period is monotonically increasing in $|y_0|$. A given starting value y_0 leads to a principal curve for the rectangle if the length b is an integer multiple of the half-period. Numerical calculations show that in order for a non-linear principal curve to exist, the aspect ratio b/a must be in an interval $I_n \approx (0.9069\ n, 0.978\ n)$, for some n. The number of such intervals containing a given value b/a increases approximately linearly with b/a. Consequently, rectangles with large aspect ratio have a large number of principal curves.

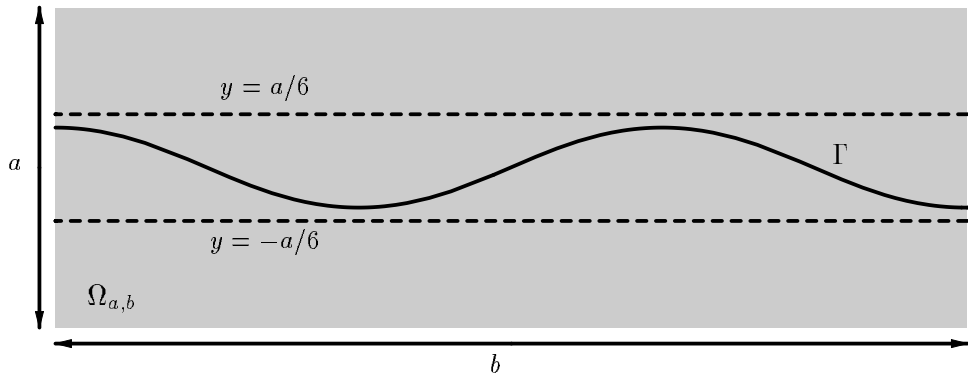

Figure 3: An oscillating principal curve for the uniform distribution on a rectangle

Consider next the uniform distribution on the annulus

$$\Omega_{R_1,R_2} = \{(r,\phi)\ :\ R_1 \le r \le R_2\}\,,$$

where (r,ϕ) are polar coordinates. For symmetry reasons there has to be a circular principal curve. Using Equation (4) relating curvature to conditional moments of the density, its radius can be shown to be

$$r_{circ} = \frac{2\left(R_1^2 + R_1R_2 + R_2^2\right)}{3\left(R_1 + R_2\right)}. \tag{6}$$

It is easy to see that *all* regular curves for the annulus are of the form $r = f(\phi)$. After a possible rotation, we may assume that f attains a local minimum at $\phi = 0$. The regularity condition requires that no center of curvature of Γ lie in the annulus. A simple calculation at a local minimum

and a local maximum of f shows that $f(\phi)$ must lie between $R_{min} = (2R_1 + R_2/3$ and $R_{max} = (R_1 + 2R_2)/3$.

Because we did not succeed in finding an analytic solution of the principal curve equations (5) for the case of the annulus, we used a variable-step 4th-order Runge-Kutta method (Runge-Kutta-Fehlberg).

Our experimental results indicate that solutions are periodic, with period $T(R_1, R_2, r_0)$ depending on the initial value r_0. In order for a solution to be a principal loop, it has to be closed, meaning that $T(R_1, R_2, r_0) = 2\pi/n$, $n = 1, 2, 3, \ldots$. This will be the case for a discrete set of values r_0. Consider, for example, the annulus $\Omega_{0.45,1}$. In addition to the circular principal loop $r = r_{circ} \approx 0.760$, there is one other principal loop, given by the initial conditions $r(0) \approx 0.646$, $r'(0) = 0$. This principle loop has period $T = \pi/4$ and oscillates between the values $r = 0.646$ and $r = 0.874$. Initial conditions of the form $r'(0) = 0$ and $r(0) < 0.646$ give curves whose periods are slightly greater than $\pi/2$ and so cannot be closed; for $r(0) > 0.646$, the period is slightly smaller than $\pi/2$ and again the curve cannot close (see Figure 4).

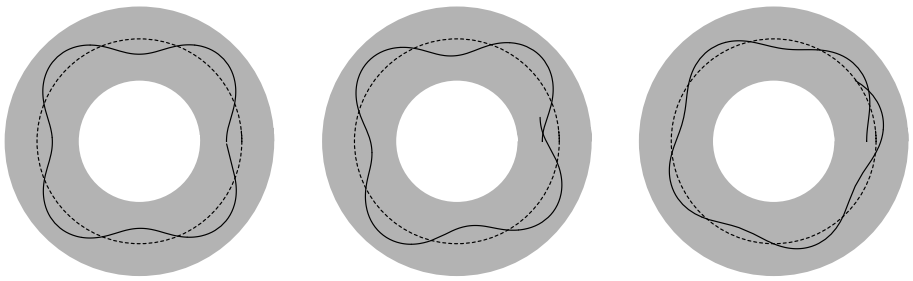

Figure 4: Three solutions of the principal curve equations on the annulus $\Omega_{0.45,1}$. From left to right the initial conditions are: $r(0) = 0.646$, $r(0) = 0.634$ and $r(0) = 0.690$. Only the first is a closed curve, and thus a principal loop

Computing principal curves for a number of annuli with different aspect ratios suggests that the number of principal loops increases with increasing aspect ratio, while the period $T(R_1, R_2, r_0)$ decreases with increasing aspect ratio (see Figure 5).

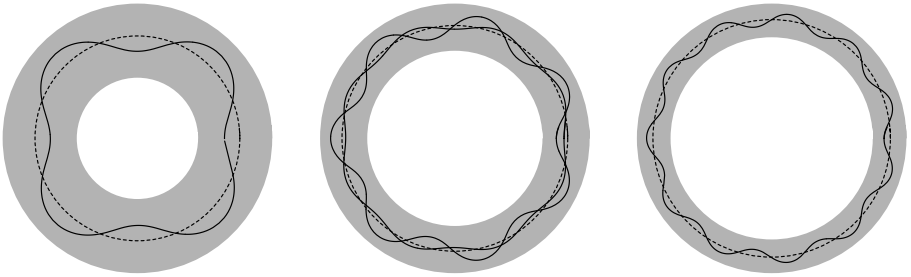

Figure 5: As the aspect ratio increases, the period of principal curves decreases. From left to right are the three annuli $\Omega_{0.45,1}$, $\Omega_{0.65,1}$ and $\Omega_{0.75,1}$ with some principal curves on each. Two principal curves are shown for the middle annulus. The right annulus also supports several principal curves (not shown).

7 Principal Curves Cross

In analogy to linear principal components, densities in general will have multiple principal curves. These curves satisfy conditions that are vaguely analogous to the orthogonality of linear principal components.

Proposition 2 *Suppose that* Γ_1 *and* Γ_2 *are principal curves for a density* p*. Then they are not linearly separable.*

Proof. Let $\mathbf{X}$ be a random vector with density p, and let Γ be a principal curve. Then

$$\mathrm{E}(\mathbf{X}) = \mathrm{E}_\Gamma \, \mathrm{E}(\mathbf{X} \mid \pi_\Gamma(\mathbf{X}) = \gamma) = \mathrm{E}(\pi_\Gamma(\mathbf{X}))$$

As $\mathrm{E}(\pi_\Gamma(\mathbf{X}))$ is in the convex hull of Γ, the convex hulls of any two principal curves have to intersect. ∎

We next show that, under mild convexity conditions on Ω, any two regular principal curves of a density must cross.

Theorem 2 . *Let Γ_1 and Γ_2 be two regular principal curves of the density p. If Γ_1 has endpoints, suppose that $\partial\Omega$ is strongly convex at those endpoints. Then Γ_1 and Γ_2 intersect.*

Proof. The proof is by contradiction and proceeds in two steps. Suppose that Γ_1 and Γ_2 do not intersect. Then there must be points $\mathbf{x}_1 \in \Gamma_1$ and $\mathbf{x}_2 \in \Gamma_2$ for which $\mathrm{dist}(\Gamma_1, \Gamma_2) = \mathrm{dist}(\mathbf{x}_1, \mathbf{x}_2) > 0$. According to Lemma 2 below these points must lie in the interior of the respective curves. This implies that the line L joining $\mathbf{x}_1$ and $\mathbf{x}_2$ intersects both curves orthogonally, and that the transverse distributions coincide. We will use the self-consistency condition, which is satisfied by both curves, to locate their centers of curvature on L.

Let $\mathcal{V} \subset L$ denote the connected component of the intersection of L with Ω containing $\mathbf{x}_1$ and $\mathbf{x}_2$. Let $\mathbf{x}_0 \in \ell$ be the mean of the transverse density induced on $\mathcal{V}$, and let $\sigma_\perp^2$ be its variance. Parameterize $\mathcal{V}$ by $\mathcal{V} = \{\mathbf{x}_0 + u\mathbf{N}_1 \mid a < u < b\}$, and set $\mathbf{x}_i = \mathbf{x}_0 + u_i\mathbf{N}_1$.

Writing the self-consistency condition (4) in the form $\kappa_i = -u_i/(u_i^2 + \sigma_\perp^2)$ shows that the u-coordinate of the center of curvature of Γ_i is given by the formula

$$c_i = u_i - \frac{u_i^2 + \sigma_\perp^2}{u_i} = -\frac{\sigma_\perp^2}{u_i}. \tag{7}$$

First consider the case where u_1 and u_2 have opposite signs, as shown in Figure 6(a). In this case, $\mathrm{d}(\mathbf{x}_1, \mathbf{x}_2)$ is not a minimum. Next consider the case where u_1 and u_2 have the same sign. Without loss of generality assume that $u_1 < u_2 \leq 0$, as shown in Figure 6(b). Equation (7) then implies that $c_1 < c_2$, so again $\mathrm{d}(\mathbf{x}_1, \mathbf{x}_2)$ is not a minimum[1]. ▌

Lemma 2 *Let Γ_1, Γ_2 be as in Theorem 2, and assume that they do not intersect. Then the distance between Γ_1, Γ_2 is realized at interior points $\mathbf{x}_1 \in \Gamma_1, \mathbf{x}_2 \in \Gamma_2$.*

Proof. Note first that the projection of $\mathbf{x}_2$ onto Γ_1 is $\mathbf{x}_1$. This shows that $\mathbf{x}_1$ cannot be an endpoint of Γ_1, for, by virtue of the strong convexity assumption, the only points in Ω projecting onto the endpoints of Γ_1 are the endpoints themselves.

[1]We wish to thank Andreas Buja for this observation.

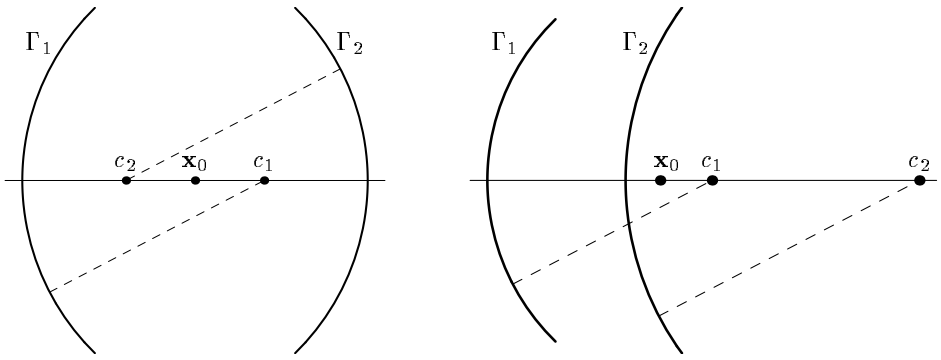

Figure 6: (a) Curves Γ_1, Γ_2 intersect the normal L on opposite sides of the mean; (b) Curves intersect the normal on the same side of the mean.

We now show that $\mathbf{x}_2$ cannot be an endpoint of Γ_2. Let $\mathbf{x}_e$ be an endpoint of Γ_2. By the transversality condition (2) of Proposition 1, Γ_2 intersects the boundary of Ω orthogonally at $\mathbf{x}_e$. By virtue of Remark 1, the vector from from $\mathbf{x}_e$ to $\pi_{\Gamma_1}(\mathbf{x}_e)$ intersects the boundary transversely, and therefore its inner product with the (inward pointing) tangent vector of Γ_2 at $\mathbf{x}_e$ is positive. Consequently, x_e cannot be the point on Γ_2 closest to $\mathbf{x}_1$. ∎

8 Principal Curves on the Disk and the Ellipse

As shown in Section 6, any spherically symmetric density has a circle as a principal loop. We now consider principal arcs for spherically symmetric densities:

Proposition 3 *Let p be a spherically symmetric density with compact support Ω a disk centered at the origin. Then the only regular principal arcs are straight lines through the origin.*

Proof. Suppose Γ is a regular principal arc. Without loss of generality we may assume that Γ enters Ω at the point $(-r, 0)$, where r is the radius of Ω. The transversality conditions require that the tangent of Γ at $(-r, 0)$ be horizontal. The straight line $y = 0$ obviously is a principal arc that satisfies

these conditions. We will use the fact that Γ is a solution to the principal curve equations (5) and thus is uniquely determined by a point and a tangent vector to conclude that Γ and the straight line $y = 0$ coincide. However, using a point and tangent vector on the boundary is too simplistic, because the principal curve equations are singular on the boundary. We thus have to argue as follows:

Suppose Γ is not a straight line. Then there has to be a point $\mathbf{x}_1$ with curvature $\kappa(\mathbf{x}_1) \neq 0$. Assume without loss of generality that $\kappa(\mathbf{x}_1) > 0$. This implies that that there must also be a point $\mathbf{x}_2$ with $\kappa(\mathbf{x}_2) < 0$ (otherwise, Γ would not cross the principal curve $y = 0$, contradicting Theorem 2). Therefore, there is a point $\mathbf{x}_0 \in \Gamma$ with $\kappa(\mathbf{x}_0) = 0$. As a consequence of the self-consistency condition (4) and spherical symmetry, Γ intersects the normal line $\{\mathbf{x}_0 + v\mathbf{N}(\mathbf{x}_0)\}$ in the center of its intersection with Ω. Thus the tangent line $\{\mathbf{x}_0 + v\mathbf{T}(\mathbf{x}_0)\}$ passes through the origin. Because Γ is a solution of the principal curve equations (5), it is uniquelly determined by $\mathbf{x}_0$ and $\mathbf{T}(\mathbf{x}_0)$. This means that it has to be a straight line through the origin, and we have reached a contradiction. ▌

Our understanding of principal curves for elliptical, non-spherical densities is not as complete. Assume that p is a uniform density on elliptical region Ω. Obviously, the major and minor axes of the ellipse are principal arcs. We do not know if there are any others. Given that spherically symmetric densities have circles as principal loops, one might suspect that elliptical densities have ellipses as principal loops. However, this is not true in general:

Proposition 4 *Let p be the uniform density on an ellipse that is not a circle. Then there is no elliptical principal loop.*

Proof.

Assume that the ellipse $\Gamma = \{(x, y) \mid x^2 + y^2/b^2 = 1\}$ is a principal curve for the uniform distribution on a convex region Ω with smooth boundary $\partial\Omega$, which is symmetric with respect to both the x- and y-axes. Using the self-consistency condition, we determine a parameterization for the curve $\partial\Omega$ and show that it is not an ellipse.

Parametrize Γ as $\mathbf{f}(\theta) = (\cos(\theta), b\sin(\theta))$. Then $\partial\Omega$ can be written as $\mathbf{z}(\theta) = \mathbf{f}(\theta) + v_-(\theta)\mathbf{N}(\theta)$, where $v_-(\theta)$ is the negative distance along the normal between $\mathbf{f}(\theta)$ and $\partial\Omega$ (see Figure 7). Let $\mathbf{z}_+(\theta) = \mathbf{f}(\theta) + v_+(\theta)\mathbf{N}(\theta)$ be the point of intersection of the normal line with the x-axis.

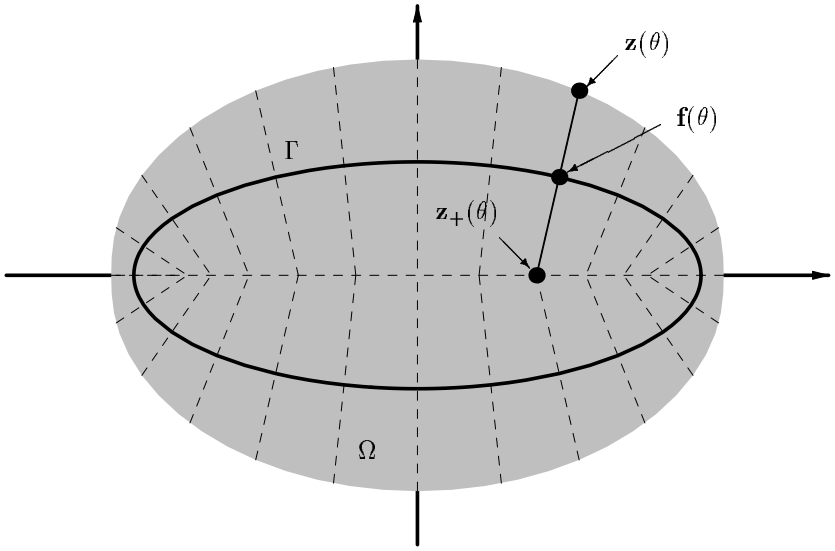

Figure 7:

It is not difficult to show that the set of ambiguity points of Γ is the interval of the x-axis joining the foci of the ellipse Γ. It follows that the inverse image $\mathcal{V}(\theta)$ of $\mathbf{f}(\theta)$ under the projection map $\pi_\Gamma : \Omega \to \Gamma$ is the interval joining $\mathbf{z}(\theta)$ and $\mathbf{z}_+(\theta)$. We now use Equation (3) and the fact that p is uniform to express the self-consistency condition as

$$(v_+(\theta)^2 - v_-(\theta)^2)/2 + (v_+(\theta)^3 - v_-(\theta)^3)/3\kappa(\theta) = 0. \tag{8}$$

Because Γ is assumed to be an ellipse, both $\kappa(\theta)$ and $v_+(\theta)$ are known, and Equation (8) can then be solved for $v_-(\theta)$.

It remains to show that $\mathbf{z}(\theta) = (x(\theta), y(\theta))$ is not an ellipse, unless $b = 1$. If the boundary were an ellipse, then its semi-major and semi-minor axes A, B would be determined by the formulas $(A, 0) = \mathbf{z}_-(0)$ and $(0, B) = \mathbf{z}_-(\pi/2)$; moreover, the error $e(\theta) = (x(\theta)/A)^2 + (y(\theta)/B)^2 - 1$ would be identically zero.

To show that $e(\theta)$ is not identically zero, it suffices to compute the second derivative of $e(\theta)$ at $\theta = 0$. One obtains the formula

$$e''(0) = 6b^2 \left(\frac{12b^2}{\left(6b^2 - 3 + \sqrt{3(1+2b^2)(3-2b^2)}\right)^2} - \frac{1}{2+b^2} \right)$$

It is not difficult, but messy, to give an analytic proof that $e''(0)$ is positive for $0 < b < 1$. The proof proceeds by simplifying the term in parentheses and isolating the numerator. Setting it equal to zero leads to the problem of finding the zeros of a fourth degree polynomial in b^2, which can be checked to have no zeros between 0 and 1. ∎

9 Conclusion

Principal curves were introduced to formalize the notion of "a curve passing through the middle of a dataset". Vaguely speaking, a curve is said to pass through the middle of a dataset if every point on the curve is the average of the observations projecting onto it. This idea can be made precise by defining principal curves for probability densities. Principal curves can be regarded as a generalization of linear principal components — if a principal curve happens to be a straight line, then it is a principal component. In this paper we study principal curves in the plane. We show that principal curves are solutions of a differential equation. By solving this differential equation, we find principal curves for uniform densities on rectangles and annuli. There are oscillating solutions besides the obvious straight and circular ones, indicating that principal curves in general will not be unique. If a density has several principal curves, they have to cross, a property somewhat analogous to the orthogonality of principal components. Finally, we investigate principal curves for spherical and elliptical distributions.

In a companion paper (Duchamp and Stuetzle (1995)) we analyze the extremal properties of principal curves. Like principal components, principal curves are critical points of the expected squared distance to the data. However, the largest principal component minimizes the distance, whereas all principal curves are saddlepoints. This explains why cross-validation does not appear to be a viable method for choosing the complexity of principal curve estimates.

References

[1] J. D. Banfield and A. E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *JASA*, 87:7–16, 1992.

[2] J.F. Brinkley. Knowledge-driven ultrasonic three-dimensional organ modeling. *IEEE Trans. Pat. Anal. Mach. Intell.*, 7(4):431–441, July 1985.

[3] T. Duchamp and W. Stuetzle. The geometry of principal curves in the plane. TR 250, Department of Statistics, University of Washington, 1993.

[4] T. Duchamp and W. Stuetzle. Extremal properties of principal curves in the plane. Submitted to *Annals of Statistics*, 1995.

[5] L. Fang and D.C. Gossard. Reconstruction of smooth parametric surfaces from unorganized data points. In J. Warren, editor, *Curves and Surfaces in Computer Vision and Graphics 3*, volume 1830, pages 226–236. SPIE, 1992.

[6] A. Goshtasby. Surface reconstruction from scattered measurements. In J. Warren, editor, *Curves and Surfaces in Computer Vision and Graphics 3*, volume 1830, pages 247–256. SPIE, 1992.

[7] T.J. Hastie. *Principal Curves and Surfaces*. PhD thesis, Stanford University, 1984.

[8] T.J. Hastie and W. Stuetzle. Principal curves. *JASA*, 84:502–516, 1989.

[9] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Surface reconstruction from unorganized points. *Computer Graphics*, 26(2):71–78, 1992.

[10] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Mesh optimization. *Computer Graphics*, 27:19–26, 1993.

[11] R.W. Martin, G. Bashein, M.L. Nessly, and F. Sheehan. Methodology for three-dimensional reconstruction of the left ventricle from transesophageal echocardiograms. *Ultrasound in Med. and Biol.*, 19(1):27–38, 1993.

[12] S. Muraki. Volumetric shape description of range data using "blobby model". *Computer Graphics*, 25(4):227–235, July 1991.

[13] R.B. Schudy and D.H. Ballard. Model detection of cardiac chambers in ultrasound images. Technical Report 12, Computer Science Department, University of Rochester, 1978.

[14] R.B. Schudy and D.H. Ballard. Towards an anatomical model of heart motion as seen in 4-d cardiac ultrasound data. In *Proceedings of the 6th Conference on Computer Applications in Radiology and Computer-Aided Analysis of Radiological Images*, 1979.

[15] F.H. Sheehan, M.P. Feneley, N.P. DeBruijn, J.S. Rankin, J.W. Davis, E.L. Bolson, P.S. Glass, and F.M. Clements. Quantitative analysis of regional wall thickening by transesophageal echocardiography. *Journal of Thoracic and Cardiovascular Surgery*, 103(2):347–354, 1992.

[16] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Trans. Pat. Anal. Mach. Intell.*, 12(2):131–147, February 1990.

[17] B.C. Vemuri, A. Mitiche, and J.K. Aggarwal. Curvature-based representation of objects from range data. *Image and Vision Computing*, 4(2):107–114, 1986.