Principal Curves With Bounded Turn

Sathyakama Sandilya and Sanjeev R. Kulkarni, *Senior Member, IEEE*

***Abstract*—Principal curves, like principal components, are a tool used in multivariate analysis for ends like feature extraction. Defined in their original form, principal curves need not exist for general distributions. The existence of principal curves with bounded length for any distribution that satisfies some minimal regularity conditions has been shown. We define principal curves with bounded turn, show that they exist, and present a learning algorithm for them. Principal components are a special case of such curves when the turn is zero.**

***Index Terms*—Bounded turn, curve fitting, feature extraction, learning, multivariate analysis, principal components, principal curves.**

I. Introduction

Principal component analysis is a widely used tool in multivariate data analysis for purposes such as dimension reduction and feature extraction. A generalization of the idea of principal components to principal curves was introduced by Hastie and Stuetzle in [1]. There has been significant interest in principal curves after the work by Hastie and Stuetzle [2]–[8]. Principal curves have since been used for applications such as modeling outlines of ice floes [9], modeling the short-time spectrum of speech signals [10], feature extraction and pattern classification [11], [12], boundary detection [13], and analysis of ecological gradients [14]. An alternative definition and learning algorithm for principal curves have been provided in [15]. A generalization of the self-organizing map in [16] is proposed and its connections with the principal curve algorithm of [1] are explored in [17]. Principal curves by their definition in [1], however, are not guaranteed to exist for any distribution. Kégl *et al.* [18] provided a new definition for principal curves with bounded length, and showed that such curves exist for any distribution with bounded second moment. Due to the length constraint, the treatment in [18] does not encompass the case of classical principal component analysis.

In this correspondence, we penalize the turn of a curve instead of its length, and look for principal curves within the class of curves of bounded total turn. The appeal of this approach consists partly in the fact that principal components are a special case of such principal curves wherein the total turn is 0. Another motivation for the preceding definition is that it was found convenient to penalize a quantity similar to the local turn of the curve (in addition to the length) in the practical implementation of the algorithm in [18]. We define principal curves with bounded turn and show that they exist and also analyze an algorithm for learning such curves.

II. Preliminaries and Notation

In this section, we introduce some definitions and present some preliminary results that we will use in the rest of the correspondence.

Definition 1: A curve in R^d is defined as a continuous function $f\colon I \mapsto R^d$ where I is a closed connected subset of R.

We denote by I_f the domain of f and by G_f the set of all points on curve f (its range). The topology we consider on the family of curves is induced by the metric

$$\rho(f, g) = \max\{\rho_1(f, g), \rho_1(g, f)\}$$
$$\rho_1(f, g) = \inf_{\phi\colon I_f \mapsto I_g} \sup_t \|f(t) - g(\phi(t))\|. \tag{1}$$

One could define the distance between curves to be given by a metric (in our case, the L_∞ metric) on the space of functions that represent these curves. However, this approach results in a distance measure that depends on the specific parametrization of the curve being considered. In order to remove this dependence, we allow the curves to be reparametrized by any function ϕ, and take the smallest of the distances between these reparametrized functions.

Consider a curve f and a point $x \in R^d$. We define the projection of x onto f as

$$t_f(x) = \arg\min_t \|x - f(t)\|.$$

Even though the minimization is over the entire (possibly infinite-length) curve, we can show that it may be reduced to a minimization over a compact set, and hence, the minimum exists. We denote the distortion of x due to its projection onto f as

$$\Delta(x, f) = \|x - f(t_f(x))\|^2.$$

When X is a random variable, we denote the expected distortion due to its projection onto f as $\Delta(f)$. That is,

$$\Delta(f) = E[\Delta(X, f)] = E[\|X - f(t_f(X))\|^2].$$

Definition 2: Given a random variable X, we say that f is a principal curve for X in a class of curves $\mathcal{C}$ if $f \in \mathcal{C}$ and

$$\Delta(f) = \inf_{g\in\mathcal{C}} \Delta(g) \triangleq \Delta^*_{\mathcal{C}}.$$

The following lemma implies that as a function of f, the squared error distortion $\Delta(x, f)$ is continuous in the metric ρ within compact subsets of R^d. The typical compact subsets of R^d that we consider are closed balls of radius R. In the sequel, unless otherwise specified, a ball is centered at the origin. We denote the ball centered at the origin of radius R as B_R.

Lemma 1: If $G_f \cup G_g \subset B_R$, then for every $x \in B_R$

$$\Delta(x, g) - \Delta(x, f) \le 8R\rho(f, g).$$

Proof: Fix $x \in B_R$. Let $t_1 = t_f(x)$ and choose t_2 so that

$$\|f(t_1) - g(t_2)\| \le 2\rho(f, g).$$

Such a choice of t_2 is possible because from the definition of the metric ρ, we know that there exists a parametrization ϕ such that $\sup_t \|f(t) - g(\phi(t))\| \le 2\rho(f, g)$

$$\begin{aligned}\Delta(x, g) - \Delta(x, f) &\le \|x - g(t_2)\|^2 - \|x - f(t_1)\|^2\\ &= (\|x - g(t_2)\| + \|x - f(t_1)\|)\\ &\quad\times (\|x - g(t_2)\| - \|x - f(t_1)\|)\\ &\le 4R\|g(t_2) - f(t_1)\|\\ &\le 8R\rho(f, g).\end{aligned}$$

□

Manuscript received March 6, 2001; revised February 21, 2002. This work was supported in part by the National Science Foundation under NYI Grant IRI-9457645 and Grant ECS-9873451, and the MURI through the Army Research Office under Grant DAAD19-00-1-0466.

S. Sandilya was with Princeton University, Princeton, NJ 08544 USA. He is now with Siemens Corporate Research, Princeton, NJ 08540 USA (e-mail: sandilya@alumni.princeton.edu).

S. R. Kulkarni is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: kulkarni@ee.princeton.edu).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier 10.1109/TIT.2002.802614.

Definition 3: The length of a curve f over an interval $[\alpha, \beta] \subset I_f$ is given by

$$l(f, \alpha, \beta) = \sup \sum_{i=1}^{N} \|f(t_i) - f(t_{i-1})\|$$

where the supremum is over all possible finite partitions of $[\alpha, \beta]$ into intervals $[t_i, t_{i+1}]$ i.e.,

$$\alpha = t_0 < t_1 \cdots < t_{N-1} < t_N = \beta.$$

We now formally define the notion of the turn of a curve as in [19], [20].

Definition 4: Consider a piecewise-linear curve f with vertices $v_0 \cdots v_n$. Let $a_i = v_i - v_{i-1}$ and let ϕ_i be the angle between a_i and a_{i+1}. The total turn of this piecewise-linear curve is defined by

$$\kappa(f) = \sum_{i=1}^{n-1} \phi_i.$$

For a general curve f, the turn accumulated over an interval $[\alpha, \beta]$ of its domain is defined as the supremum over all piecewise-linear inscriptions in $[\alpha, \beta]$, i.e.,

$$\kappa(f, \alpha, \beta) = \sup_n \sup_g \kappa(g)$$

where g is a piecewise-linear curve with vertices $f(t_0) \cdots f(t_n)$ such that $t_i \in I_f$ and $\alpha = t_0 < t_1 < \cdots < t_{n-1} < t_n = \beta$. The turn of the entire curve f is defined as

$$\kappa(f) \overset{\Delta}{=} \sup_{\alpha, \beta \in I_f} \kappa(f, \alpha, \beta).$$

It can be shown that for smooth curves, the turn of a curve measures the integral of the curvature of the curve with respect to its arc length. The above definition of turn extends this in a natural way to nonsmooth curves. It is shown in [19] that total turn is lower semicontinuous, i.e., if $f_n \to f$, then $\kappa(f) \le \liminf \kappa(f_n)$. As an immediate consequence, we observe that $\{f\colon \kappa(f) \le K\}$ is closed in this topology.

Lemma 2 [18]: Let A be a compact subset of R^d and $L \ge 0$. The set $\{f\colon G_f \subset A, l(f) \le L\}$ is compact.

The above lemma is shown in [18] with respect to the topology that they consider, but it holds even with respect to the topology induced by the metric in (1) (our topology is coarser than that in [18]). This compactness result implies that any sequence of such curves has a convergent subsequence—a property that is crucial in our proof of existence of principal curves. The following lemma from [19] states that a curve of bounded turn that lies within a compact set has bounded length. As any closed subset of a compact set is compact, Corollary 1 follows from these two lemmas and the lower semicontinuity of turn.

Lemma 3 [19]: If $\kappa(f) \le K$ and the diameter of G_f is not more than p, then $l(f) \le p\zeta(K)$ where ζ is as follows:

$$\zeta(x) = \begin{cases} 1/\cos x/2, & 0 \le x \le \pi/2 \\ 2\sin(x/2), & \pi/2 \le x \le 2\pi/3 \\ x/2 - \pi/3 + \sqrt{3}, & x \ge 2\pi/3. \end{cases}$$

Corollary 1: If $A \subset R^d$ is compact, then $\{f\colon G_f \subset A, \kappa(f) \le K\}$ is compact.

III. Existence of Principal Curves of Bounded Turn

We shall now proceed to the proof of existence of principal curves of bounded total turn. The main idea in the construction is to use the compactness property of the set of curves of bounded turn within a compact subset of R^d. We know that for any class of curves $\mathcal{C}$, there exists a sequence of curves in $\mathcal{C}$ whose distortions converge to $\Delta^*_{\mathcal{C}}$. From this sequence, we construct a subsequence of curves such that this subsequence converges on any compact subset of R^d. We then obtain a "limiting" curve from this subsequence and show that it achieves the minimum distortion in the class and, therefore, is a principal curve.

We need to make more precise the meaning of curves converging on a compact set. To that end, we define the restriction of a curve to a set as follows.

Definition 5: Consider a closed set $A \subset R^d$ and a curve f such that $f(0) \in A$. We define $f|_A$ (the restriction of f to A) as the curve g with $I_g = [a, b]$ where $0 \in [a, b], \forall\, t \in [a, b],\ f(t) = g(t) \in A$, and $[a, b]$ is the largest such interval.

Such a definition is necessary as the intersection of a curve with a set A can, in general, be a union of curves. We would like to pick one of these connected components as the restriction of the curve to A. The above definition gives a consistent way to make the choice of this component. We say that f_n converges to f on A if

$$f_n|_A \to f|_A.$$

The existence result we would like to state is that for any finite K, there exists a principal curve with total turn at most K. However, this is not true as demonstrated by the following example. Let $X = (X_1, X_2) \in R^2$ be distributed according to the one-dimensional Gaussian distribution along two parallel lines with X being equally likely to occur on either of the two lines, i.e.,

$$P(X_2 = 0) = P(X_2 = 1) = 1/2$$

and

$$f_{X_1}(x_1) = \exp(-x_1^2/2)/\sqrt{2\pi}.$$

In this case, if we consider all curves with total turn less than or equal to π, then the infimum of the distortions is 0. This infimum is, however, not achieved by any curve as the distribution of X is concentrated on a set that is not connected.

Hence, we need to impose more stringent regularity conditions on the class of curves we consider to ensure that minimizers of our objective function exist. The problem with the above example is that the curves are permitted to accumulate their turn arbitrarily far from the origin resulting in the "limit" of these curves not being a curve, but a union of curves. In order to circumvent the above problem, we impose a uniform bound on the rate at which the turn accumulated within B_R (see Fig. 1) converges to the total turn of the curve, i.e., fix $\tau(R)$ continuous and decreasing (to zero) in R and consider the class of curves

$$\mathcal{C}_{K,\tau} = \{f\colon \kappa(f) \le K, \kappa(f) - \kappa(f|_{B_R}) \le \tau(R)\}. \tag{2}$$

An alternate approach would be to consider a family of curves that includes unions of finitely many curves and introduce a composite cost function that penalizes both the cumulative turn and the number of connected components in the union.

In the following lemma, we show that if a curve has bounded turn, then it does not cross any annulus infinitely often. In particular, for any R, there exists an R' such that once the curve exits the ball of radius R', it does not re-enter the ball of radius R. The lemma is true for any curve of bounded turn, but we prove it here under the assumption in (2). The assumption guarantees the existence of an R' such that this is true uniformly across the entire class $\mathcal{C}_{K,\tau}$.

Lemma 4: Let f be a curve in $\mathcal{C}_{K,\tau}$ as in (2). For any $R > 0$, there exists R' such that $G_f \cap B_R$ is contained in $G_{f|_{B_{R'}}}$.

Proof: Choose R' so that $\tau(R') < \arccos(R/(R' - R))$ and $f(0) \in B_{R'}$. We now show the contrapositive. Assume that there exists a point on f and within B_R that is not on $f|_{B_{R'}}$. As f is continuous, and $f|_{B_{R'}}$ is not the entire curve, we know that there exists x_1 that is

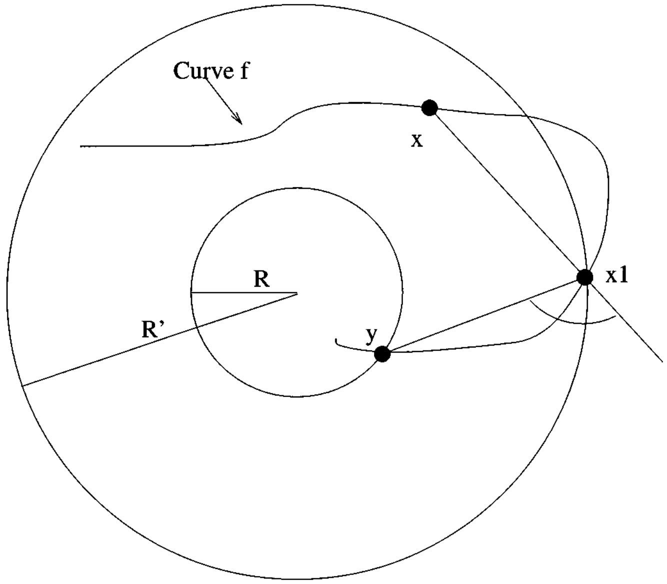

Fig. 1. Turn accumulated outside $B_{R'}$.

on f such that $\|x_1\| = R'$ (see Fig. 1). From the definition of turn in terms of polygonal line inscriptions, we have

$$\begin{aligned}
&(\kappa(f) - \kappa(f|_{B_{R'}})) \\
&\quad \geq \inf_{x \in B_{R'} \cap G_f} \inf_{y \in B_R \cap G_f} \arccos\left(\frac{\langle x_1 - x,\, y - x_1\rangle}{\|x_1 - x\|\|y - x_1\|}\right) \\
&\quad \geq \inf_{x \in B_{R'}} \inf_{y \in B_R} \arccos\left(\frac{\langle x_1 - x,\, y - x_1\rangle}{\|x_1 - x\|\|y - x_1\|}\right) \\
&\quad \geq \arccos\left(\sup_{x \in B_{R'}} \sup_{y \in B_R} \frac{\langle x_1 - x,\, y - x_1\rangle}{\|x_1 - x\|\|y - x_1\|}\right) \\
&\quad \geq \arccos\left(\sup_{x \in B_{R'}} \sup_{y \in B_R} \frac{\langle x_1 - x,\, y\rangle}{\|x_1 - x\|(R' - R)}\right) \\
&\quad \geq \arccos\left(\frac{R}{R' - R}\right)
\end{aligned}$$

which is a contradiction. The preceding inequalities are justified by the fact that $\arccos(\cdot)$ is monotonically decreasing. □

Proposition 1: Consider the class of curves $\mathcal{C}_{K,\tau}$ as detailed in (2). If $E[\|X\|^2] < \infty$, then there exists a principal curve in $\mathcal{C}_{K,\tau}$.

Proof: Let $\Delta^* = \inf_{f \in \mathcal{C}_{K,\tau}} \Delta(f)$. We know that there exists a sequence of curves $f_n \in \mathcal{C}_{K,\tau}$ such that $\Delta(f_n) \to \Delta^*$.

First we show that there exists a compact set (a ball B_R) in R^d that has nonempty intersection with infinitely many f_n. Suppose that such a ball does not exist, i.e., for every m, for large enough n, $G_{f_n} \cap B_m = \emptyset$. Choose $R > 0$ such that

$$(R^2/4)P[X \in B_{R/2}] > E[\|X\|^2].$$

If $G_{f_n} \cap B_R = \emptyset$, then

$$\begin{aligned}
\Delta(f_n) &= E\left[\|X - f(t_f(X))\|^2 1_{\{\|X\| \leq R/2\}}\right] \\
&\quad + E\left[\|X - f(t_f(X))\|^2 1_{\{\|X\| > R/2\}}\right] \\
&> (R^2/4)P[X \in B_{R/2}] \\
&\quad + E\left[\|X - f(t_f(X))\| 1_{\{\|X\| > R/2\}}\right] \\
&> E[\|X\|^2].
\end{aligned}$$

As the distortion of the curve that contains only the origin is the second moment of X, and this trivial curve is in the family of curves we consider, the least distortion among this family of curves cannot strictly exceed $E[\|X\|^2]$. The distortion of the sequence f_n converges to this minimum. Thus, this is a contradiction. Hence, there is a ball B_R with which infinitely many f_n have nonempty intersections. Consider this B_R and let f_{n_k} be a subsequence such that $G_{f_{n_k}} \cap B_R \neq \emptyset$. Reparametrize the sequence of curves f_{n_k} by a translation of the domain so that $f_{n_k}(0) \in B_R$. This ensures that the restriction of f_{n_k} to any ball containing B_R is a nonempty curve. From the lemmas and corollary in the last section, we have that for each $r > R$, $f_{n_k}|_{B_r}$ has a convergent subsequence. Construct a set of sequences n_k^j (for $j \geq R$) such that

- $\{n_k^{j+1}\}$ is a subsequence of n_k^j and
- $f_{n_k^j}|_{B_j}$ is convergent.

We now show that the sequence of curves $\{f_{n_j^j}\}$ converges on every compact subset of R^d. Consider a ball of radius a. We know that the sequence $\{f_{n_j^j},\, j > \lceil a \rceil\}$ is a subsequence of $f_{n^j_{\lceil a \rceil}}$, and hence, its restriction to $B_{\lceil a \rceil}$ is convergent. Therefore, $f_{n_j^j}$ is a sequence of curves whose restriction to any ball is convergent. For notational simplicity, we now denote this newly constructed sequence of curves as f_n.

We now define the "limiting curve" f^* as the curve that satisfies for every $r > R$

$$f_n|_{B_r} \to f^*|_{B_r}.$$

In order to show that this is indeed a principal curve, we need now to show that it is in $\mathcal{C}_{K,\tau}$ and that it minimizes distortion within $\mathcal{C}_{K,\tau}$, i.e., $\Delta(f^*) = \Delta^*$.

First we show that $f^* \in \mathcal{C}_{K,\tau}$. As the total turn of a curve is the limit of the turns of finite polygonal line inscriptions and the total turn is lower semicontinuous, we have

$$\begin{aligned}
\kappa(f^*) &= \lim_{R \to \infty} \kappa(f|_{B_R}) \\
&\leq \lim_{R \to \infty} \liminf_n \kappa(f_n|_{B_R}) \\
&\leq \lim_{R \to \infty} K = K.
\end{aligned}$$

We have effectively shown that if a sequence of curves converges on all compacts, then the turn of the "limiting" curve is less than the liminf of the turns of the sequence of curves. Next, we need to show that the decay rate of the turn outside B_r is smaller than $\tau(r)$.

Fix r and δ such that $r - \delta > \|f^*(0)\|$. Let

$$\begin{aligned}
s_0 &= \sup\{t < 0, \text{ such that } \|f^*(t)\| = r - \delta\} \\
t_0 &= \sup\{t > 0, \text{ such that } \|f^*(t)\| = r - \delta\} \\
s_n &= \sup\{t < 0, \text{ such that } \|f_n(t)\| = r - \delta\} \\
t_n &= \sup\{t < 0, \text{ such that } \|f_n(t)\| = r - \delta\}.
\end{aligned}$$

We now construct the following curves:

$$\begin{aligned}
g(t) &= f(t) \text{ for } t \in I_f \text{ and } t \leq s_0 \\
h(t) &= f(t) \text{ for } t \in I_f \text{ and } t \geq t_0.
\end{aligned}$$

$$\begin{aligned}
g_n(t) &= f_n(t) \text{ for } t \in I_{f_n} \text{ and } t \leq s_n \\
h_n(t) &= f_n(t) \text{ for } t \in I_{f_n} \text{ and } t \geq t_n.
\end{aligned}$$

Then

$$\begin{aligned}\kappa(f^*) &\le \kappa(g) + \kappa(f^*|_{B_r}) + \kappa(h)\\ &\le \kappa(f^*|_{B_r}) + \liminf_n \kappa(g_n) + \kappa(h_n)\\ &\le \kappa(f^*|_{B_r}) + \liminf_n \kappa(f_n) - \kappa(f_n|_{B_{r-\delta}})\\ &\le \kappa(f^*|_{B_r}) + \tau(r-\delta).\end{aligned}$$

The second inequality in the preceding expression holds because g_n and h_n converge on all compacts to g and h, respectively. As that inequality is true for every sufficiently small $\delta > 0$, and $\tau(\cdot)$ is decreasing and continuous, we have that $\kappa(f^*) \le \kappa(f^*|_{B_R}) + \tau(r)$.

On the other hand, if $r < \|f(0)\|$, then $f|_{B_r}$ is empty, and hence, for large enough n, $f_n|_{B_r}$ is empty, and hence,

$$\begin{aligned}\kappa(f^*) - \kappa(f^*|_{B_r}) &= \kappa(f^*)\\ &\le \liminf_n \kappa(f_n)\\ &= \liminf_n \kappa(f_n) - \kappa(f_n|_{B_r})\\ &\le \tau(r)\end{aligned}$$

as each $f_n \in \mathcal{C}_{K,\tau}$. Moreover, $\kappa(f^*) \le \tau(r)$ for all $r < \|f(0)\|$ implies $\kappa(f^*) \le \tau(\|f(0)\|)$.

Now, we show that f^* achieves Δ^*. Let $r > R$. For all $x \in B_r$, we have $\Delta(x, f_n) < 4r^2$ and hence, $f_n(t_f(x)) \in B_{3r}$. Moreover, according to Lemma 4, we have that there exists r' such that

$$G_{f_n|B_{r'}} \supset G_{f_n} \cap B_{3r}.$$

Hence,

$$\begin{aligned}\Delta(f^*) &= E[\Delta(X, f^*)]\\ &= E[\Delta(X, f^*) - \Delta(X, f_n) + \Delta(X, f_n)]\\ &= E[\Delta(X, f_n)] + E[(\Delta(X, f^*) - \Delta(X, f_n))1_{B_r}(X)]\\ &\quad + E[(\Delta(X, f^*) - \Delta(X, f_n))1_{B_r^C}(X)]\\ &\le E[\Delta(X, f_n)] + E[(\Delta(X, f^*) - \Delta(X, f_n))1_{B_r}(X)]\\ &\quad + E[4\|X\|^2 1_{B_r^C}(X)]\\ &\le E[\Delta(X, f_n)] + 8r'\rho(f^*|_{B_{r'}}, f_n|_{B_{r'}}) + E[4\|X\|^2 1_{B_r^C}(X)]\end{aligned}$$

where the inequality for the second term stems from the continuity of Δ with respect to ρ (Lemma 1). The rest of the first inequality follows from the following explanation. As $\Delta(x, f_n)$ is a positive function

$$E[(\Delta(X, f^*) - \Delta(X, f_n))1_{B_r^C}(X)] \le E[\Delta(X, f^*)1_{B_r^C}(X)]$$

and as $G_f \cap B_r \ne \emptyset$ we have for $x \in B_r^C$, $\Delta(x, f) \le \|x\| + r < 2\|x\|$, which implies

$$E[\Delta(X, f^*)1_{B_r^C}(X)] \le E[4\|X\|^2 1_{B_r^C}(X)].$$

Choose r so that $E[4\|X\|^2 1_{B_r^C}(X)] < \epsilon/3$, and n large enough so that $\Delta(f_n) < \Delta^* + \epsilon/3$ and $8r'\rho(f^*|_{B_{r'}}, f_n|_{B_{r'}}) < \epsilon/3$. These conditions together imply that for any $\epsilon > 0$, $\Delta(f^*) < \Delta^* + \epsilon$, and hence $\Delta(f^*) = \Delta^*$. □

IV. A Learning Result

We now consider the problem of learning a principal curve when the distribution of X is unknown but we are given observations of a random process $X_1, X_2, \ldots$ drawn independent and identically distributed (i.i.d.) according to F_X. The goal is to arrive at a curve in $\mathcal{C}_{K,\tau}$ such that its expected distortion is small and gets closer to that of a principal curve as we get more data to learn from. As the distribution of X is unknown, we estimate the distortion of a curve based on the empirical measure that assigns a weight $1/n$ to each observed data point. We denote the empirical distortion of curve f by

$$\Delta_n(f) = \frac{1}{n}\sum_{i=1}^{n} \Delta(X_i, f).$$

In order to arrive at the principal curve, we resort to structural risk minimization (see [21], [22]). When we have a finite amount of data, we cannot optimize over the entire class $\mathcal{C}_{K,\tau}$ as this may lead to overfitting to the data. Hence, we choose a sequence of classes of increasing complexity within which the optimization is conducted. Just as in [18], we consider classes of polygonal lines with increasing number of segments. A distinction that we make is that we also expand the set in which these polygonal lines lie as the random variable X is not assumed to be bounded. The classes S_k are more precisely defined below

$$S_k = \{f \in \mathcal{C}_{K,\tau},\ G_f \subset B_{R_k} \text{ and } f \text{ is a polygonal line with } k \text{ segments}\}$$

where R_k is an increasing unbounded sequence. After n observations, we choose a curve in S_{k_n} that minimizes the empirical distortion as the candidate f_n for the principal curve. That is,

$$f_n = \arg\min_{f \in S_{k_n}} \Delta_n(f). \tag{3}$$

For judiciously chosen k_n, the distortion of the curves f_n converges to that of a principal curve.

In order to obtain convergence rates for the learning process, we need to impose further regularity on the distribution of X. Heretofore, we only assumed that $E\|X\|^2 < \infty$. Now we assume that the decay rate of the contribution to the second moment from outside balls of radius R is bounded by some known function of R. Namely

$$E[\|X\|^2 1_{B_R^C}(X)] \le \beta(R). \tag{4}$$

Proposition 2: Suppose that $\beta(R) = R^{-\alpha}$, and that $k_n = n^{1/3}$. Then the expected distortion of f_n as in (3) converges to the distortion of a principal curve at a rate

$$\Delta(f_n) - \Delta(f^*) = O\left(n^{-\frac{\alpha}{6+3\alpha}}\right).$$

In order to show this proposition, we need a lemma shown in [18] which we state below in a slightly modified form (as a result of applying Lemma 3).

Lemma 5: For any $\epsilon > 0$, there exists a finite collection $S_{k,\epsilon}$ of curves in B_{R_k} such that for any $f \in S_k$, there exists $f' \in S_{k,\epsilon}$ such that

$$\sup_{x \in B_{R_k}} |\Delta(x, f) - \Delta(x, f')| < \epsilon$$

and the cardinality of $S_{k,\epsilon}$ satisfies

$$\begin{aligned}|S_{k,\epsilon}| &\le 2^{D_k^2\zeta(K)/\epsilon + 2k + 1} V_d^{k+1}\\ &\quad \times \left((D_k^2/\epsilon + 1)\sqrt{d}\right)^d \left((D_k^2\zeta(K)/k\epsilon + 3)\sqrt{d}\right)^{kd}\end{aligned}$$

where V_d is the volume of the d-dimensional unit sphere and $D_k = 2R_k$.

Proof (of Proposition 2): This proof very closely follows the corresponding proof in [18]. We consider unbounded random variables X, and curves of possibly infinite length, and so we separate the contribution to error from within and without a ball of finite radius. We use the relationship between turn and length of a curve to cast our problem in the framework of [18] and then use techniques therein to bound the contribution from within these balls. The error from outside such balls is bounded by the extra regularity (4) we imposed.

Let $f_n^* = \arg\min_{f \in S_k} \Delta(f)$

$$\Delta(f_n) - \Delta(f^*) = \Delta(f_n) - \Delta(f_n^*) + \Delta(f_n^*) - \Delta(f^*)$$

$$\begin{aligned}\Delta(f_n^*) - \Delta(f^*) &= E[(\Delta(X, f_n^*) - \Delta(X, f^*))1_{B_{R_k}}(X)] \\ &\quad + E[(\Delta(X, f_n^*) - \Delta(X, f^*))1_{B_{R_k}^C}(X)] \\ &\le \frac{8R_k^2\zeta(K)}{k} + 4E\left[\|X\|^2 1_{B_{R_k}^C}(X)\right].\end{aligned}$$

The second term in the preceding inequality follows when $G_{f_n^*} \cap B_{R_k} \neq \emptyset$, which is eventually true, and the first term follows from Lemmas 1 and 3 as there exists a curve $g \in S_k$ such that $\rho(g, f^*|_{B_{R_k}}) \le l(f^*|_{B_{R_k}})/2k$

$$\begin{aligned}\Delta(f_n) - \Delta(f_n^*) = E&\left[(\Delta(X, f_n) - \Delta(X, f_n^*))1_{B_{R_k}}(X)\right] \\ &+E\left[(\Delta(X, f_n) - \Delta(X, f_n^*))1_{B_{R_k}^C}(X)\right].\end{aligned}$$

Using the same techniques as in [18], it can be shown that

$$\begin{aligned}E&\left[(\Delta(X, f_n) - \Delta(X, f_n^*))1_{B_{R_k}}(X)\right] \\ &\le 2\epsilon + O(n^{-1/2}) + \sqrt{\frac{2D_k^4\log(|S_{k,\epsilon}|+1)}{n}}.\end{aligned}$$

Hence, we have

$$\begin{aligned}\Delta(f_n) - \Delta(f^*) &\le \frac{8R_k^2\zeta(K)}{k} + 2\epsilon + O(n^{-1/2}) \\ &\quad + \sqrt{\frac{2D_k^4\log(|S_{k,\epsilon}|+1)}{n}} \\ &\quad + 8E\left[\|X\|^2 1_{B_{R_k}^C}(X)\right].\end{aligned}$$

Setting $R_k = k^{1/(2+\alpha)}$, $\epsilon = k^{-\alpha/(2+\alpha)}$ and $k = n^{1/3}$ we obtain

$$\Delta(f_n) - \Delta(f^*) \le Qn^{-\alpha/(6+3\alpha)}$$

for some constant Q. □

We note here that when X is a bounded random variable, we may let $\alpha \to \infty$ and obtain a convergence rate of $n^{-\frac{1}{3}+\epsilon}$ for any ϵ, which is close to that obtained in [18] for learning principal curves of bounded length. Moreover, the same idea may be used to extend the learning result of [18] to unbounded distributions.

V. Conclusion

We proposed a definition for principal curves with bounded total turn. Under this definition, we showed the existence of principal curves, and demonstrated a theoretical algorithm based on complexity regularization for learning such curves from i.i.d. data.

Principal curves reduce the representation of the given data from d dimensions to one dimension. We may, however, be interested in increasing the accuracy of the reduced representation by increasing the dimension of the representation. Principal component analysis, being linear in nature, provides a natural way of adding these extra dimensions in a recursive manner. Even though projection onto principal curves is a nonlinear operation, it may be interesting to investigate the possibility of the existence of a second principal curve that represents in some way the residual error of projection on the first. Another approach to higher dimension representations is to look for principal manifolds. The techniques we presented here only assumed that with respect to a topology on the set of curves that we were optimizing over, this set is compact on compact subsets of R^d, and the loss function is continuous. If there exist regularity conditions on manifolds that imply the above, then the same approach we outlined here may be used to show existence of principal manifolds. A more interesting question in this arena may be to arrive at learning algorithms for these manifolds. A slightly different approach to principal manifolds is presented in [23].

References

[1] T. Hastie and W. Stuetzle, "Principal curves," *J. Amer. Statist. Assoc.*, pp. 502–516, 1989.

[2] D. C. Stanford and A. E. Raftery, "Finding curvilinear features in spatial point patterns: Principal curve clustering with noise," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 601–609, June 2000.

[3] T. Duchamp and W. Stuetzle, "Extremal properties of principal curves in the plane," *Ann. Statist.*, vol. 24, no. 4, pp. 1511–1520, 1996.

[4] W. J. Krzanowski, "Recent trends and developments in computational multivariate analysis," *Statist. and Comput.*, vol. 7, no. 2, pp. 87–99, 1997.

[5] A. R. Webb, "A loss function approach to model selection in nonlinear principal components," *Neural Net.*, vol. 12, no. 2, pp. 339–345, 1999.

[6] B. Chalmond and S. C. Girard, "Nonlinear modeling of scattered multivariate data and its application to shape change," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 422–432, May 1999.

[7] K. Chang and J. Ghosh, "A unified model for probabilistic principal surfaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 22–41, Jan. 2001.

[8] P. Delicado, "Another look at principal curves and surfaces," *J. Multivariate Anal.*, vol. 77, pp. 84–116, 2001.

[9] J. D. Banfield and A. E. Raftery, "Ice floe identification in satellite images using mathematical morphology and clustering about principal curves," *J. Amer. Statist. Assoc.*, pp. 7–16, 1992.

[10] K. Reinhard and M. Niranjan, "Subspace models for speech transitions using principal curves," *Proc. Inst. Acoust.*, pp. 53–60, 1998.

[11] K. Chang and J. Ghosh, "Principal curves for nonlinear feature extraction and classification," *Proc. SPIE: Applications of Artificial Neural Networks in Image Processing III*, pp. 120–129, 1998.

[12] ——, "Principal curve classifier—A nonlinear approach to pattern classification," in *Proc. IEEE Int. Joint Conf. Neural Networks*, 1998, pp. 695–700.

[13] J. B. Chen and I. Zurbenko, "Nonparametric boundary detection," *Commun. Statist.—Theory and Methods*, vol. 26, no. 12, pp. 2999–3014, 1997.

[14] G. De'ath, "Principal curves: A new technique for indirect and direct gradient analysis," *Ecology*, vol. 80, no. 7, pp. 2237–2253, 1999.

[15] R. Tibshirani, "Principal curves revisited," *Statist. and Comput.*, pp. 183–190, 1992.

[16] T. Kohonen, "Clustering, taxonomy, and topological maps of patterns," in *Proc. 6th Int. Conf. Pattern Recognition*, 1982, pp. 114–128.

[17] F. Mulier and V. Cherkassky, "Self-organization as an iterative kernel smoothing process," *Neural Comput.*, pp. 1165–1177, 1995.

[18] B. Kégl, A. Krzyżak, T. Linder, and K. Zeger, "Learning and design of principal curves," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 281–297, Mar. 2000.

[19] A. D. Alexandrov and Yu. G. Reshetnyak, "General theory of irregular curves," in *Mathematics and Its Applications (Soviet Series)*. Norwell, MA: Kluwer, 1989.

[20] S. R. Kulkarni, S. K. Mitter, J. N. Tsitsiklis, and O. Zeitouni, "PAC learning with generalized samples and an application to stochastic geometry," *IEEE Trans. Pattern Anal. Machine Intell.*, pp. 933–942, Sept. 1993.

[21] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[22] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[23] A. J. Smola, S. Mika, B. Schölkopf, and R. C. Williamson, "Regularised principal manifolds," in *Lecture Notes on Artificial Intelligence: Computational Learning Theory*, P. Fischer and H. U. Simon, Eds. Berlin, Germany: Springer Verlag, 1999, pp. 214–229.