Principal Curves: Learning and Convergence[1]

Balázs Kégl
Dept. of Computer Science
Concordia University
1450 de Maisonneuve Blvd.
West, Montreal PQ, Canada
H3G 1M8
email: kegl@cs.concordia.ca

Adam Krzyżak
Dept. of Computer Science
Concordia University
1450 de Maisonneuve Blvd.
West, Montreal PQ, Canada
H3G 1M8
email: krzyzak@cs.concordia.ca

Tamás Linder
Dept. of Mathematics and Statistics
Queen's University
Kingston, Ontario, Canada
K7L 3N6
email: linder@mast.queensu.ca

Kenneth Zeger
Dept. of Electrical and Computer Engineering
University of California
San Diego, La Jolla CA
92093-0407 USA
email: zeger@ucsd.edu

Abstract — **Principal curves have been defined as "self consistent" smooth curves which pass through the "middle" of a d-dimensional probability distribution or data cloud. We take a new approach by defining principal curves as continuous curves of a given length which minimize the expected squared distance between the curve and points of the space randomly chosen according to a given distribution. The new definition makes it possible to carry out a theoretical analysis of learning principal curves from training data and it also leads to a new practical construction.**

I. Introduction

Hastie and Stuetzle [1] (hereafter HS) generalized the self consistency property of principal components and introduced the notion of *principal curves.* Consider a d-dimensional random vector $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$ with finite second moments, and let $\mathbf{f}(t) = (f_1(t), \ldots, f_d(t))$ be a smooth curve in $\mathbb{R}^d$ parametrized by $t \in \mathbb{R}$. For any $\mathbf{x} \in \mathbb{R}^d$ let $t_f(\mathbf{x})$ denote the parameter value t for which the distance between $\mathbf{x}$ and $\mathbf{f}(t)$ is minimized. By the HS definition, $\mathbf{f}(t)$ is a principal curve if it does not intersect itself and

$$\mathbf{f}(t) = \mathbf{E}(\mathbf{X} | t_f(\mathbf{X}) = t),$$

that is, $\mathbf{f}(t)$ is the conditional expectation of $\mathbf{X}$ given that $\mathbf{X}$ is closer to $\mathbf{f}(t)$ than to any other point of $\mathbf{f}$.

There remains an unsatisfactory aspect of the definition of principal curves in the original HS paper as well as in subsequent works. Although principal curves have been defined to be *nonparametric*, their existence for a given distribution or probability density is an open question, except for very special cases such as elliptical distributions. This also makes it very difficult to theoretically analyze any estimation scheme for principal curves. Below we give a new definition of principal curves which resolves this problem and which leads to a new effective algorithm.

II. Principal Curves with a Length Constraint

A *curve* in d-dimensional Euclidean space is a continuous function $\mathbf{f} : I \to \mathbb{R}^d$, where I is a closed interval of the real line. Let the expected squared distance of $\mathbf{X}$ from $\mathbf{f}$ be defined by

$$\Delta(\mathbf{f}) = E\big[\inf_t \|\mathbf{X} - \mathbf{f}(t)\|^2\big] = E\|\mathbf{X} - \mathbf{f}(t_\mathbf{f}(\mathbf{X}))\|^2.$$

We give the following new definition of principal curves.

Definition 1 *A curve* $\mathbf{f}^*$ *is called a* principal curve *of length* L *for* $\mathbf{X}$ *if* $\mathbf{f}^*$ *minimizes* $\Delta(\mathbf{f})$ *over all curves of length less than or equal to* L*.*

A useful advantage of the new definition is that principal curves of length L always exist if $\mathbf{X}$ has finite second moments, as the next result shows.

Lemma 1 *Assume that* $E\|\mathbf{X}\|^2 < \infty$*. Then for any* $L > 0$ *there exists a curve* $\mathbf{f}^*$ *with* $l(\mathbf{f}^*) \le L$ *such that*

$$\Delta(\mathbf{f}^*) = \inf\{\Delta(\mathbf{f}) : l(\mathbf{f}) \le L\}.$$

III. Learning Principal Curves

Suppose that the distribution of $\mathbf{X}$ is concentrated on a closed and bounded convex set $K \subset \mathbb{R}^d$, and we are given n training points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ drawn independently from the distribution of $\mathbf{X}$. Let $\mathcal{S}$ denote the family of curves taking values in K and having length not greater than L. For $k \ge 1$ let $\mathcal{S}_k$ be the set of polygonal curves (broken lines) in K which have k segments and whose lengths do not exceed L.

Let $\Delta(\mathbf{x}, \mathbf{f}) = \min_t \|\mathbf{x} - \mathbf{f}(t)\|^2$ denote the squared distance between $\mathbf{x}$ and $\mathbf{f}$. For any $\mathbf{f} \in \mathcal{S}$ the empirical squared error of $\mathbf{f}$ on the training data is the sample average $\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^n \Delta(\mathbf{X}_i, \mathbf{f})$. Let our theoretical algorithm choose an $\mathbf{f}_{k,n} \in \mathcal{S}_k$ which minimizes the empirical error, i.e, let

$$\mathbf{f}_{k,n} = \arg\min_{\mathbf{f} \in \mathcal{S}_k} \Delta_n(\mathbf{f}).$$

The efficiency of the estimator is measured by the difference $J(\mathbf{f}_{k,n})$ between the expected squared loss of $\mathbf{f}_{k,n}$ and the optimal expected squared loss achieved by $\mathbf{f}^*$, i.e.,

$$J(\mathbf{f}_{k,n}) = \Delta(\mathbf{f}_{k,n}) - \Delta(\mathbf{f}^*).$$

The next theorem upper bounds the loss of the estimator in terms of the training data size n.

Theorem 1 *Assume that* $\mathbf{P}\{\mathbf{X} \in K\} = 1$ *for a compact and convex set* K*, let* n *be the number of training points, and let* k *be chosen to be proportional to* $n^{1/3}$*. Then the expected squared loss of the empirically optimal broken line with* k *segments and length at most* L *converges, as* $n \to \infty$*, to the squared loss of the principal curve of length* L *at a rate*

$$J(\mathbf{f}_{k,n}) = O(n^{-1/3}).$$

Based on the theoretical learning scheme above, a practical algorithm [2] for constructing principal curves has been developed. The new algorithm compares favorably with existing methods both in terms of complexity and performance.

References

[1] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[2] B. Kégl, A. Krzyżak, T. Linder, and K. Zeger, "Learning and design of principal curves" *preprint*, 1997. (Java implementation: `http://www.cs.concordia.ca/~grad/kegl/pcurvedemo.html`)

[1] This research was supported in part by the NSF and NSERC.