An Effective Principal Curves Extraction Algorithm for Complex Distribution Dataset

Hongyun Zhang*, Duoqian Miao, Lijun Sun, and Ying Ye

The Key Laboratory of Embedded System and Service Computing, Ministry of Education, China, Tongji University, Shanghai 201804, China
School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China
Zhangongyun583@sina.com

Abstract. This paper proposes a new method for finding principal curves from complex distribution dataset. Motivated by solving the problem, which is that existing methods did not perform well on finding principal curve in complex distribution dataset with high curvature, high dispersion and self-intersecting, such as spiral-shaped curves, Firstly, rudimentary principal graph of data set is created based on the thinning algorithm, and then the contiguous vertices are merged. Finally the fitting-and-smoothing step introduced by Kégl is improved to optimize the principal graph, and Kégl's restructuring step is used to rectify imperfections of principal graph. Experimental results indicate the effectiveness of the proposed method on finding principal curves in complex distribution dataset.

Keywords: Principal curves, Complex distribution dataset, Thinning algorithm, Fitting-smoothing step, Image skeletonization.

1 Introduction

Hastie and Stuetzle introduced the notion of principal curves to solve the problems in traditional machine learning and multivariate data analysis in 1989 [1]. Principal curves are self-consistent smooth curves that pass through the middle of a data set. More specifically, we hope to find curves passing through the middle of the datasets, which can truly reflect the shape of the data. Principal curves are non-linear generalizations of principal components, the basic idea of which is to seek low-dimensional manifolds embedded in the high-dimensional space [2]. Due to all the satisfying properties and advantages, principal curves have gained its rapid development since 1990's with various definitions of principal curves having been proposed. Banfield and Raftery gave their principal curve definition called BR principal curve in 1992 [3]. Kégl proposed PL principal curve definition in 2000 [4]. Verbeek defined K-segment principal curve in 2002 [5], while Delicado introduced D principal curve in 2003 [8]. Currently, a

* Corresponding author, Mobile: 13917907676; Email: zhanghongyun583@sina.com

J. Yu et al. (Eds.): RSKT 2010, LNAI 6401, pp. 328–335, 2010.

great number of achievements have been reported concerning the applications of principal curves, such as shape detection [3], intelligent transportation analysis [2], speech recognition [7], image skeletonization, feature extraction [6] and Data Compression and Regression analysis [9].

Based on these definitions, ways of finding principal curves from data sets have been proposed one after the other. Hastie and Stuetzle offered an alternation between projecting data onto the curve and estimating conditional expectations on projectors by the scatter smoother or the spline smoother [1]. Banfield and Raftery modified the Hastie-Stuetzle method, using the projection residual of the data, instead of the data themselves, to estimate conditional expectations, for reducing both bias and variance [3]. Verbeek et al. proposed a k-segments algorithm which incrementally combines local line segments into the polygonal line to achieve an objective [5]. Kégl et al. presented the polygonal line algorithm which starts with an initial polygonal line, adds a new vertex to the polygonal line at each iteration, and updates the positions of all vertices so that the value of a penalized distance function is minimized [4]. Delicado found the principal oriented points one by one and orderly linked them to estimate principal curves [8].

However, for complex distribution dataset with high curvature, high dispersion and self-intersecting, such as spiral-shaped curves and spring-shaped curves, existing methods did not work well. Verbeek et al. attempted to solve this problem by combining line segments, which were optimized to minimize the total squared distance of all points to their closest segments into a polygonal line, but did not fit the curve well. All of these algorithms almost use the first principal component of all data as the initial estimation of the principal curve when lacking the prior knowledge. Unfortunately it is bad initialization for complex distribution dataset with high curvature, high dispersion and self-intersecting. So we need to consider the global structure feature of data set from the beginning. In this paper, an effective strategy is introduced to extract principal curves for complex distribution dataset. Instead of starting with a simple topology such as the first principal component and then increasing its complexity iteratively, we directly span the sufficient complex topology and then refine it iteratively. In our algorithm, instead of principal component analysis, thinning-based method is used to initialize data and create rudimentary principal graph of data set. Then, considering the large scale of dispersion and amount, we merging the contiguous vertices and improve the fitting-smoothing step of Kégl's principal curves algorithm [4]to optimize the principal graph. Finally, Kégl's restructuring step [4] is used to refine the graph.

2 Principal Curve

Loosely speaking, principal curves are smooth one-dimensional (1D) curves that pass through the "middle" of a set of p-dimensional data points[1]. The goal is to provide smooth and low dimensional summaries of the data. Here, a 1D curve in a p-dimensional space is a vector f of p functions indexed by one single variable s. The parameter$s$ is the arc length along the curve. For any density$h$ inR^p

with finite second moments, the curve$f$ is a principal curve of $h$if the following self-consistent criterion is satisfied for almost every$s$:

$$f(s) = E[X|s_f(X) = s] \tag{1}$$

And

$$s_f(X) = sup\{s : \|X - f(s)\| = inf\|Y - f(\tau)\|\} \tag{2}$$

In the above X is a random vector from h, $s_f$is the projection index function which maps any value of $X = x$ to the value of s for which $f(x)$ is closest to x. The definition of a principal curve indicates that any point of a principal curve is the condition expectation of those points that project to this point, and a principal curve satisfies the property of self-consistent. Fig. 1 shows a first principal component line and a principal curve, Compared with corresponding first principal component, two obvious advantages of a principal curve can be observed: first, a principal curve can keep more information of data; and second, it can describe the outline of primitive information better.

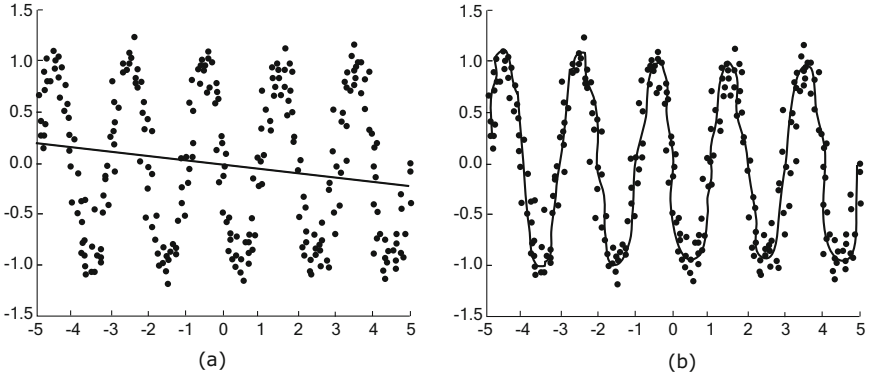

Fig. 1. Comparison between principal curve and first principal component

3 Principal Curves Extraction Algorithm

To overcome the ineffectiveness of the first principal component which is used to initialize data with high curvature, high dispersion and self-intersecting, such as spiral-shaped curves, thinning-based method is used to initialize data and create rudimentary principal graph of data set. Then, considering the large scale of dispersion and amount, we merging the contiguous vertices and improve the fitting-smoothing step of Kégl's principal curves algorithm [4] to optimize the principal graph. Finally, Kégl's restructuring step[4] is used to refine the graph. It contains the following steps:

STEP 1. Global Structure Extraction Step. Zhang-Suen thinning algorithm is adopted to generate sufficiently complex topology, that is, directly obtain the approximate initial skeleton of the original data template. However, it is not smooth

and usually contains a number of spurious branches and inadequate structural elements. The skeleton is denoted by G_{vs}, which consists of V and S, where $V = \{v_1, \ldots, v_m\}$ is a set of *vertices*, and $S = \{(v_{i1}, v_{j1}), \ldots, (v_{ik}, v_{jk})\}, 1 \leq i1, j1, \ldots, ik, jk \leq m$ is a set of *edges*, while S_{ij} is a line segment that connects v_i and v_j.

STEP 2.Vertices-Merge Step. Considering two remarkable characteristics of complex distribution dataset, high level of dispersion and large quantity of data points, in this process, we merge the adjacent vertices in terms of distance and curvature. The distance criterion makes sure that at least a certain number of vertices are retained in a certain area coverage while the curvature criterion is set to reduce the number of vertices merged in areas with big curvature. Vertices-Merges step has two advantages: (a) complex distribution dataset includes thousands of vertices, and the Vertices-Merge step can effectively reduce the number of vertices which need to be adjusted. As a result, the efficiency of the algorithm is improved; (b) Vertices -Merge step increases the ratio of data points to skeleton vertices. In that case, more data points on average are involved in adjusting a single skeleton vertex so that the deviation of skeleton is controlled under a certain degree.

STEP 3. The Improved Fitting-Smoothing Step. Iteratively fit and smooth the skeleton by repeatedly projecting data point and optimizing vertex, until convergence is achieved while keeping the skeleton approximately equidistant from the contours of the dataset. The vertices smoothing and fitting process includes the following two steps.

STEP 3.1. The Improved Projection Step: The original Kégl algorithm points out: given a data set $X_n = \{x_1, \ldots, x_n\}$, scan the whole skeleton for every data point x_i , the data point x_i is partitioned into "the nearest neighbor regions" according to which segment or vertex it projects. This step is time-consuming, since thousands of scans are required. Considering the characteristic of complex distribution dataset, the Projection Step of original algorithm is improved. The step only scans certain areas of the vertices around the data point x_i instead of the whole skeleton. Through a large scale of samples-training, we set the width of 30 to 60 pixels which can lead to relatively good results both on effectiveness and time-costing.

STEP 3.2. The Improved Vertex Optimization Step: In original algorithm, every vertex v_i in the skeleton is optimized by using a gradient method to adjust the positions of vertices and segments for finding a local minimum of $E(G)$. The penalized distance function $E(G)$ is as follows:

$$E(G) = \frac{1}{n} \sum_{i=1}^{n} \Delta(x_i, G) + \lambda P(G) \tag{3}$$

$$P(G) = \frac{1}{m} \sum_{i=1}^{m} P_v(v_i) \tag{4}$$

$\frac{1}{n}\sum_{i=1}^{n}\Delta(x_i, G)$ is the average squared distance of all points from the G_{vs}. The smaller the value of $\frac{1}{n}\sum_{i=1}^{n}\Delta(x_i, G)$ is, the better G_{vs} fits the data. λ is a penalty coefficient that determines the trade-off between the accuracy of the approximation and smoothness of the curves. $P(G)$ is a penalty on the total curvature of the skeleton. The smaller the value of $P(G)$, the smoother G_{vs} is . n is the number of the data points. $\Delta(x_i, G)$ is the Euclidean squared distance between a point x_i and the nearest point of the skeleton to x_i. m is the number of vertices. $P_v(v_i)$ is the curvature penalty at vertex v_i.

According to the distribution of complex pattern data, since lots of triangle functions and branch structures are used in the original penalty function $P(G)$, the calculation process is really time-costing. We have figured out a way to solve this problem by replacing $P(G)$ with a new penalty function $D(G)$.

$$D(G) = \frac{1}{m}\sum_{i=1}^{m}\sum_{x\epsilon V_i \bigcup S_i}\Delta(x, v_i) \tag{5}$$

From the experimental results, we sum up that the function $D(G)$ has three advantages: (a) it helps the skeleton convergence much more; (b) it reduces the skeleton deviation; (c) since $D(G)$ only involves simple calculation of addition and average, it is more efficient than $P(G)$ which uses triangle functions.

Obviously, we redefine the penalized distance function $E(G)$:

$$E(G) = \Delta(G) + \lambda D(G) \tag{6}$$

STEP 3.3. The Judge Step: Judge if the adjusted skeleton meets the convergent condition, if true, goto STEP 4, else goto STEP 3.1.

STEP 4. The Restructuring Step: Rectify the structural imperfections of the skeleton graph by deleting short paths and small loops to get more accurate skeleton.

STEP 5. END.

4 Experimental Results and Analysis

We have carried on two typical experiments on complex distribution dataset. The first experiment tested the capability of the our algorithm for extracting principal curves from data distribution of simulated data sets, and the second one's aim was to test the effects of the proposed algorithm in the applications of image skeletonization with images of logo and fingerprint.

4.1 Simulated Data Sets

We generated data sets along some curves by the commonly used additive noise model.

$$X^{'} = X + e \tag{7}$$

Where X is a data set uniformly distributed on a smooth curve which is called generating curve. $e = (e_1, e_2)$ is a bivariate additive noise which is independent of X , and $X^{'}$ is the generated data set.

We constructed various shaped curves, such as line, circle, half-circle, S-shaped, etc. Then we tested the proposed algorithm on these data sets. It turned out to have exciting results. In order to emphasize the algorithm's effectiveness for finding principal curves from complex distribution dataset with high curvature, high dispersion and self-intersecting, we select experimental result on the spiral-shaped curve and spring-shaped curve. Fig. 2 proves the effectiveness of our proposed algorithm.

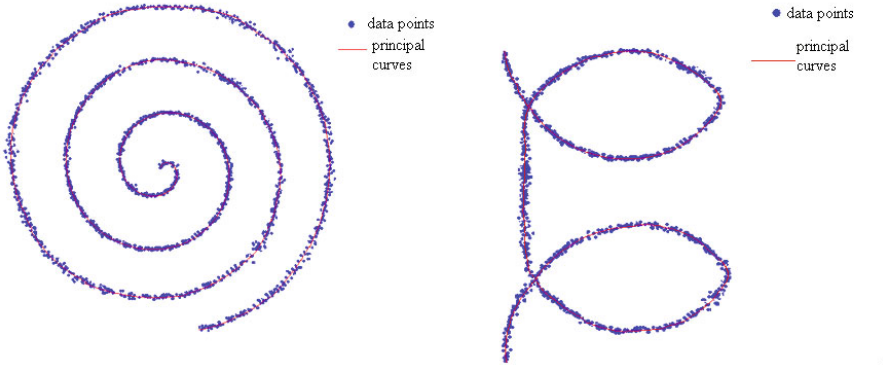

Fig. 2. Principal curves on simulated data set

Kégl et al. [4] pointed out the gradual degradation of their polygonal line algorithm on spiral-shaped generating curves of increasing length. Their algorithm performed well on spirals with one and a half, and two loops, but failed when the number of loops reached three. The Hastie-Stuetzle algorithm performed worse in the same experiments. Delicado and Hueta [8] also compared the performances of their algorithm with the PL algorithm and the Hastie-Stuetzle algorithm on the spiral with two loops and concluded that the Kégl's algorithm and theirs behave quite similarly which are both better than the Hastie-Stuetzle algorithm. Verbeek et al. [5] made further efforts, summed up the failures of different algorithms and successfully validated the effectiveness of their k-segment algorithm on the spiral with two and a half loops. The resultant polygonal line using twelve line segments by the k-segment algorithm recovered the basic shape of the spiral, but did not fit the curve well. In our experiment, the loops of spiral were more than that of all the above experiments. The number of loops reached four and the experimental result was almost indistinguishable from the generating spiral-shaped curve. In addition, our algorithm fitted the curve well.

4.2 Image Skeletonization

The self-consistency property of principal curves is quite similar to the equidistance property of medial axis of shapes. If foreground pixels of a shape are represented by a two dimensional data set, then the principal curves of this data set is the approximation to its skeleton of shape. Singh et al. and Kégl et al. have already applied the principal curves in image skeletonization. Now we also used the proposed algorithm to look for skeletons of images.

The performance of the algorithm was tested on three suites of images. The first one is the logo of Tongji university. The second one is the fingerprint pictures from FVC2002 fingerprint database [10]. Experimental results are listed as Fig. 3 Results confirmed the effectiveness of our algorithm on skeletonization of images whose dataset distribution meets complex distribution.

Fig. 3. Result of logo and fingerprint image skeletonization

5 Conclusions

Since the notion of principal curves was put forward, researchers have already introduced several methods of how to find principal curves from data sets. However, as for the complex distribution dataset such as high degree of dispersion, bending or self-intersecting, those existing methods are not ideal. To solve this problem, the paper proposes a new method. Firstly, it initializes a set of vertices to create the preliminary skeleton by using thinning algorithm after which adjacent vertices are merged. Then the fitting-smoothing step of Kgl's principal curves algorithm is improved to smooth vertices positions and construct the principal graph from these vertices through iteration. Finally, Kgl's restructuring step is used to refine the graph. Algorithm has been tested on simulated data sets and applied to image skeletonization. Experimental results confirm the effectiveness of the proposed method on finding principal curves from complex distribution dataset.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (granted No. 60775036, 60970061), National 973 Program (granted No. 2003CB316902).

References

1. Hastie, T., Stuetzle, W.: Principal curves. Journal of the American Statistical Association 84(406), 502–516 (1989)
2. Zhang, J.P., Chen, D.W., Kruger, U.: Adaptive Constraint K-segment Principal Curves For Intelligent Transportation Systems. IEEE Transactions on Intelligent Transportation Systems 9(4), 666–677 (2008)
3. Banfield, J.D., Raftery, A.E.: Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. Journal of the American Statistical Association 87(417), 7–16 (1992)
4. Kégl, B., Krzyzak, A., Linder, T.: Learning and Design of Principal Curves. IEEE Trans. on Distribution Analysis and Machine Intelligence 22(3), 281–297 (2000)
5. Verbeek, J.J., Vlassis, N., Krose, B.: A k-segments algorithm for finding principal curves. Distribution Recognition Letter 23(10), 1009–1017 (2002)
6. Zhang, H.Y., Miao, D.Q., Zhang, D.X.: Analysis and Extraction of Structural Features of Off -Line Handwritten Digits Based on Principal Curves. Journal of Computer Research and Development 42(8), 1344–1349 (2005)
7. Reinhard, K., Niranjan, M.: Parametric subspace modeling of speech transitions. Speech Communication 27(1), 19–42 (1999)
8. Delicado, P., Huetra, M.: Principal curves of oriented points.theoretical and computational improvements. Comput. Stat. 18(2), 293–315 (2003)
9. Jochen, E., Gerhard, T., Ludger, E.: Data Compression and Regression Based on Local Principal Curves. In: Advances in Data Analysis, Data Handling and Business Intelligence. Springer, Heidelberg (2009)
10. FVC2002 web site, `http://bias.csr.unibo.it/fvc2000/download.asp`