Journal of Computational and Applied Mathematics 298 (2016) 222–235

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Constraint local principal curve: Concept, algorithms and applications

Dewang Chen[a,*], Jiateng Yin[b], Shiying Yang[b], Lingxi Li[c], Peter Pudney[d]

[a] College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China
[b] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
[c] Department of Electrical and Computer Engineering, Purdue School of Engineering and Technology, IUPUI, IN, USA
[d] School of Mathematics and Statistics, University of South Australia, Mawson Lakes Campus, Australia

HIGHLIGHTS

- We present the concept of Constraint Local Principal Curve (CLPC) to reduce the computational complexity.
- We propose three CLPC algorithms that combine local optimization and adaptive radius to expand the range of applications and increase the solution quality.
- We present some numerical experiments with three simulation data sets and two measured GPS data sets in highway and railway.
- The CLPC algorithms can improve the accuracy and computational speed compared with the existing KPC algorithms.
- The features of the each CLPC algorithm are analyzed according to the comprehensive experiments.

ARTICLE INFO

Article history:
Received 3 April 2015
Received in revised form 12 November 2015

Keywords:
Constraint local principal curve (CLPC)
GPS
Local optimization
Adaptive radius
Principal of nearest neighbor

ABSTRACT

Existing principal curve algorithms have some drawbacks such as time consuming and narrow application scope in practice, since these algorithms are mainly based on global optimization. In this paper, we present the concept of Constraint Local Principal Curve (CLPC), which uses local optimization methods and restricts the principal curve with two fixed endpoints to reduce the computational complexity. In addition, we propose three CLPC algorithms by Local Optimization and Adaptive Radius to expand the range of applications and increase the solution quality. The first algorithm, i.e., CLPCg is based on greedy thinking. The second algorithm, i.e., CLPCs uses one dimensional search and the last algorithm CLPCc combines the greedy thinking and one dimensional search. Then, we define six performance indices to evaluate the performance of the CLPC algorithms. Finally, we present some numerical experiments with three simulation data sets and two GPS measured data sets in both highway and railway. The results indicate that all of the three CLPC algorithms can obtain high-accuracy data from multiple low-accuracy data efficiently. The CLPC algorithms can improve the accuracy and computational speed compared with the existing K-segment principal curve (KPC) algorithm. In addition, CLPCc outperforms CLPCg and CLPCs according to the comprehensive experiments while CLPCg runs much faster than other ones.

[*] Corresponding author.
E-mail addresses: dwchen@fzu.edu.cn (D. Chen), jiatyin@bjtu.edu.cn (J. Yin).

http://dx.doi.org/10.1016/j.cam.2015.11.041
0377-0427/© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Principal curves are defined to be self-consistent smooth curves that pass through the middle of a data set, which can provide a nonlinear summary of the data [1]. There are many important applications of principal curves, which include: correcting magnet locations in the Stanford Linear Collider [2]; image processing: modeling ice contours [3] and analyzing handwriting [4–6]; power systems [7–9]; speech recognition [10,11]; making complex data sets audible [12]; ecology: finding low dimensional descriptions of multiracial redundant data [13]; process control [14,15]; digital mapping: where geodetic maps and GPS play an important role in train control systems [16,17].

Spearman [18] first proposed the linear principal component analysis method for finding a vector onto which data points could be projected with minimum mean-squared deviation of projection distances. This work was followed by Hastie [1], who considered principal curves to be a generalization of Spearman's linear component analysis, and principal component analysis based on line to be a special case of a principal curve. In 1989, Hastie and Stuetzle proposed the concept of principal curves (HS principal curves) as a smooth curve that passes through the middle of the data set, where principal curve is skeleton of the data set and the data set is the cloud of the curve [2]. They required a principal curve to be self-consistent, that is, each point on the curve must be the conditional mean of all data points that are projected onto it. HS principal curve algorithms are based on the assumption that principal curves are continuously differentiable and smooth, which may lead to a series of practical problems. In particular, if there is a large curvature or a closed HS principal curve, the estimation bias will have a significant impact on the principal curve.

In 1992, Banfield and Raftery developed principal curves (BR principal curves) which solved these problems [3]. But, the BR principal curves algorithm may obtain smooth but false principal curves in practice. Tibshirani introduced a semi-parametric method and redefined the principal curves based on a hybrid model [4]. But Tibshirani principal curves (T principal curves) cannot be proved to exist uniquely in some situations. In 1996, Duchamp and Stuetzle studied the global differential geometry of the principal curve [5,6]. They also proved that the principal curves are not unique. In 1999, Kegl and Krzyzak proposed principal curves with length constraint and proved the existence and uniqueness of the Kegl principal curves (K principal curves), as long as the finite two moments of the data set exist [19–21]. They also proposed a polygon algorithm to find K principal curves. One disadvantage of the K principal curves method, however, is that the length of the curve needs to be fixed in advance. Combining the generative topographical mapping (GTM), Chang and Ghosh defined the probabilistic principal curves (PPCs) [22,23]. Different from all above works, PPCs aim to find the low dimensional manifold embedded in high-dimensional data space, which retain the self-consistency of HS principal curves and can extend principal curves to higher dimensional manifold. To deal with large data sets, Zhang et al. applied granular computation on principal curves, which were shown to be more effective and robust [24]. Verbeek et al. proposed the concept of K-segment principal curve algorithm (KPCv) [25], which uses local principal component method to form K vectors, and connects them to generate a principal curve. But it may be not applicable for high dimensional data sets.

Many engineering applications require the approximating curve that has two fixed end points. An important example is a curve that defines a fixed path, e.g., a road or a rail line that connects two cities or places. In many transportation application projects, we aim to construct a curve describing a path from noisy GPS measurements taken while traveling along the path (a road or a rail line). An important feature of this problem is that the data are ordered by approximated distance along the path, and the path is traversed only once. This implies that we can traverse the data from start to finish, and use efficient local optimization methods to construct the approximated curves. Hence, we need to modify the concept, model, and algorithm of principal curves to obtain broader application ranges.

In 2008, we proposed a constraint principal curves algorithm to handle problems with endpoint constraints [26]. But the method is based on global optimization, which is too time-consuming to be practical for large scale data sets. To realize the fusion of multiple GPS track data, we proposed a heuristic algorithm based on continual split and non-linear optimization in 2008 [27]. This algorithm also has some practical defects, such as time consuming which limits the range of its applications. In 2011, Jia et al. proposed a Max Point Method optimization algorithm for constraint K-segment principal curve (CKPC), named as CKPCm [28]. However, this algorithm is difficult to deal with complex curves with good fitness and robustness. Zhang et al. proposed two principal curve algorithms for partitioning high-dimensional data spaces [29]. Although they advance one step for reducing computational complexity, it is still too complex to be used for engineers. In 2013, Chen et al., proposed three practical constraint principal curve algorithms based on global optimization [30]. Although they can generate better initial solutions, the computational complexity is still high for the large data sets, and the algorithms cannot work well for the spiral data sets.

To address these mentioned problems, we focus on an innovative constraint principal curve with less computational complexity, broader application range, and ease to understand. Hence, we put forward the idea of CLPC (Constraint Local Principal Curve). For this CLPC approach, we use some constraints on principal curves with two fixed endpoints and local optimization methods to reduce the computational complexity, and we also apply an adaptive radius to divide a big data set into some small subsets efficiently. Moreover, we describe three efficient algorithms in this paper for constructing constraint principal curves that can be used to estimate fixed paths from GPS measurements. On the basis of a greedy algorithm, (i.e. CLPCg [31]), we propose two improved CLPC algorithms i.e., CLPCs and CLPCc, by using adaptive radius and local optimization methods. To test the effectiveness and efficiency of the proposed algorithms, we compare their performance (with the KPCv method proposed by Verbeek [25] and the CKPCm proposed by Jia [28]) using three simulated data sets and two GPS measured data sets.