International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Multi-granularity principal curves extraction based on improved spectral clustering of complex distribution data ☆

Hongyun Zhang [a,b,*], Ting Zhang [a,b], Peipei Wang [a,b], Zhihua Wei [a,b,*]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, PR China
[b] Key Laboratory of Embedded System & Service Computing, Ministry of Education of China, Shanghai 201804, PR China

ARTICLE INFO

Article history:
Received 15 March 2018
Received in revised form 3 December 2018
Accepted 6 December 2018
Available online xxxx

Keywords:
Multi-granularity principal curve
Complex data
Granulation
PL principal curves
Improved spectral clustering

ABSTRACT

In order to address the problem of instability of existing principal curve algorithms and their difficulties to deal with complex distribution data, in this study, invoking the ideas of information granulation, we propose a local-to-global multi-granularity principal curves extraction approach based on improved spectral clustering. Firstly, we propose an improved spectral clustering algorithm based on inflection point estimation to granulate complex distribution data into several granular data, and develop techniques of an automatic selection of a parameter, which determines the number of clusters (granular data). Secondly, PL (polygonal line) principal curves algorithm proposed by Kégl is utilized to extract the local principal curves of each granular data. Finally, with the use of the shortest Hamiltonian path algorithm and noise variance, the local principal curves are gradually connected together to form a global curve. A number of numeric studies completed for synthetic and publicly available data sets provide a useful quantifiable insight into the effectiveness of the proposed algorithm.

1. Introduction

As a nonlinear generalization of principal component analysis [1], principal curve is a well-known technique encountered in multivariate analysis. Principal curve was defined by Hastie [2] as a one-dimensional (1D) smooth curve, which passes through the "middle" of a set of p-dimensional data points, which can truly reflect the shape of the data. That is, the curve is the "skeleton" of the dataset, while the dataset is the "cloud" of the curve. It means that the principal curve can preserve useful information about the data. Hastie put forward the HS's principal curves algorithm (HSPC) to extract the principal curves by iterating between projecting data onto the curve and estimating conditional expectations on projectors by the scatter smoother or the spline smoother [2]. Referring to the HS's algorithm, many researchers have offered improvements to the theory as well as the ensuing algorithmic developments. To address the problem of model bias of the HSPC algorithm, Tibshirani introduced a semi-parametric principal curve model (hereafter TPC), in which an EM algorithm was used to estimate principal curves [3]. In 2000, Kégl et al. defined a polygonal curve with k segments and length L as principal curves to solve the problem of convergence of the HSPC algorithm (KPC) [4]. Delicado in 2001 defined principal curves as principal curves of oriented points to correct estimation bias (DPC) [5]. Verbeek et al. in 2001 defined soft K-segment principal curves to extract principal curves from circular data distribution [6]. More recently, Zhang et al. proposed Riemannian

☆ This paper is part of the Virtual special issue on Uncertainty in Granular Computing, Edited by Duoqian Miao and Yiyu Yao.

* Corresponding authors at: Department of Computer Science and Technology, Tongji University, Shanghai 201804, PR China.
E-mail addresses: zhanghongyun@tongji.edu.cn (H. Zhang), zhihua_wei@tongji.edu.cn (Z. Wei).

https://doi.org/10.1016/j.ijar.2018.12.006
0888-613X/

principal curves to address the problem of non-constant data distributions in 2010 (hereafter RPC) [7]. Zhang et al. extended principal curves to granular principal curves, and completed some preliminary studies to extract principal curves from large-scale data in 2014 [8]. Principal curves have been found to be an important method to summarize information residing in experimental data. Considerable work has been reported on applications of principal curves to various problems, such as shape detection [9,10], speech recognition [11], intelligent transportation analysis [12], high-dimensional data partitioning [13], image skeletonization [14], and interest point detection [15]. However, with the rapid development of the Internet and information systems, we often need to handle complex data, and here the efficiency of the existing principal curves algorithms becomes lower when dealing with complex data owing to the lack of prior knowledge.

Granular Computing (GrC) has emerged as a new way to model ways of complex problem solving by concentrating on forming information granules and realizing processing at various levels of resolution (scales) or granularities. Information granulation is a powerful tool supporting processing and interpreting complex data [16]. In this paper, referring to the idea of GrC [17–22], we granulate complex distribution data into local data, and propose a local-to-global multi-granularity principal curves extraction approach. In the design process, according to the characteristics of data distribution, we first propose an improved spectral clustering algorithm to granulate complex distribution data into several granular data, and develop a technique to automatically determine the number of such granular data. At this step, we compute low dimensional embedding of initial data with complex distribution using the spectral clustering algorithm [23,24] based on manifold distance kernel [25], and transform the initial data into low-dimensional representative data with spherical shapes. Afterwards, based on the concept of local density and minimum distance of CFSFDP (clustering by fast search and finding of density peaks) algorithm [26], we propose the method of inflection point estimation to automatically determine the number of granules and obtain the results of granulation. Then, the PL principal curve algorithm proposed by Kégl is employed to extract the local principal curves of each granular data, and the local principal curves are optimized by deleting the overfitting edges. Finally, by means of the Hamiltonian path algorithm and noise variance, the local principal curves are gradually connected to form the final smooth global curve.

The paper is organized as follows. In Section 2, the improved spectral clustering based on the inflection point estimate is introduced. Section 3 elaborates on the proposed multi-granularity principal curve algorithm. Experimental studies are covered to evaluate and analyze the performance of the proposed algorithm in Section 4. Finally, Section 5 delivers some conclusions.

2. The improved spectral clustering based granulation for complex distribution data

Manifold distance is considered to address the problem of learning the underlying global geometry of a data. As it is capable of reflecting the complex space distribution of data, we choose spectral clustering algorithm based on the manifold distance kernel [23] to granulate the complex distribution data. In the process of granulation, the selected algorithm can deal with more complex datasets, but it still exhibits two drawbacks. Firstly the number of cluster (granule) and cluster centers need be manually determined. Secondly, granulation results are unstable owing to the improper cluster centers. To alleviate these drawbacks, we improve the algorithm and propose an improved spectral clustering algorithm based on inflection point estimation to granulate complex distribution data.

2.1. The idea and drawbacks of spectral clustering algorithms based on manifold distance kernel

Spectral clustering algorithm based on manifold distance kernel introduces the manifold distance kernel into the NJW-Spectral (Ng–Jordan–Weiss-Spectral) clustering algorithm [27]. Firstly, all the data points are considered as the vertices of a graph. Secondly, the manifold distance between the vertexes is computed and the similarity matrix and the Laplacian matrix are constructed based on the manifold distance. Then, low-dimensional representative data embedded into the initial data are obtained by calculating the eigenvectors of the Laplacian matrix. Finally, the representative data are divided into k clusters by running the K-means clustering algorithm.

The manifold distance kernel based NJW algorithm can obtain acceptable clustering results on some datasets, but it still exhibits two drawbacks.

(1). In the K-means step, the number of clusters needs to be predetermined; the cluster number exhibits a great influence on the clustering results. For example, a spiral dataset consists of three spirals. When we specify the number of clusters to be 3, the results are correct. When we set the number of clusters to be 2, 4 or 5, the results exhibit a large deviation from the original structure. As shown in Fig. 1, the number of clusters has a visible impact on the final clustering results, however a “correct” number of clusters is not known in advance.
(2). The clustering results obtained by the algorithm proposed in [23] are not stable as it uses the K-means clustering algorithm to cluster the representative data. The performance of the K-means clustering algorithm depends on the random selection of initial cluster centers; when these centers are not properly selected, the algorithm may be trapped in a local minimum. As presented in Fig. 2, the results show the drawback mentioned above.

Please cite this article in press as: H. Zhang et al., Multi-granularity principal curves extraction based on improved spectral clustering of complex distribution data, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.12.006

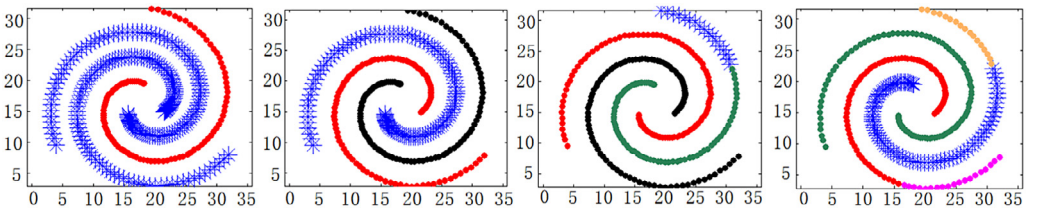

Fig. 1. Cluster results obtained for different number of clusters.

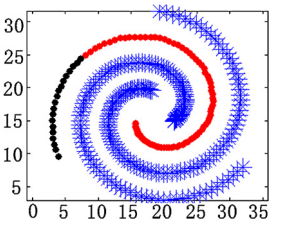

Fig. 2. Example of unstable clustering result.

With the use of the manifold distance, the NJW algorithm can deal with more complex datasets. However the K-means step of the algorithm cause the algorithm to become incorrect and unstable if the cluster number and cluster centers are not correctly determined. Hence, an improved spectral clustering algorithm is proposed here to deal with those drawbacks.

2.2. An improved spectral clustering based on inflection point estimate

CFSFDP [26] algorithm is robust with respect to data distribution. Hence, compared to K-means clustering algorithm, it is more suitable to cluster the data exhibiting complex distribution, such as spherical-like data and spiral-like data. Besides, it is able to automatically find the cluster centers. Considering that the distribution of representative data is spherical, CFSFDP is a viable option.

The CFSFDP algorithm relies on the assumptions that cluster centers are surrounded by neighbors with lower local density and that they are positioned at a relatively large distance from any points with a higher local density. For each data point i, we compute two quantities: its local density ρ_i and distance δ_i. Both these quantities depend only on the distances d_{ij} between data points, which are assumed to satisfy the triangular inequality. The local density ρ_i of data point i is defined as

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise, and d_c is a cutoff distance. Basically, ρ_i is equal to the number of points that are closer than d_c to point i. The algorithm is only sensitive to the relative magnitude of ρ_i in different points, implying that, for large datasets, the results of the analysis are robust with respect to the choice of d_c.

δ_i is measured by computing the minimum distance between the point i and any other point with higher density.

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{2}$$

For the point with highest density, we let $\delta_i = \max_j(d_{ij})$. Note that δ_i is much larger than the typical nearest neighbor distance only for the points that are local or global maxima of the density. Thus, cluster centers are recognized as points for which the value of δ_i is anomalously large.

The cluster centers are automatically determined by the CFSFDP but the cluster numbers are required to be set up manually. Based on the concept of the local density and minimum distance of CFSFDP algorithm, we develop a method of inflection point estimation to automatically determine the number of clusters and produce clustering results.

Specifically, a new quantity γ_i called decision attribute is introduced:

$$\gamma_i = \rho_i * \delta_i \tag{3}$$

where ρ_i is the local density and δ_i is the minimum distance proposed by the CFSFDP algorithm. Hence, we can obtain γ_i in the dimension reduction step of the spectral clustering algorithm. As shown in Fig. 3, the values of γ_i of the Spiral dataset are sorted from the largest to the smallest. The overall distribution of the values is similar to the log function, and most of the values are close to 0.

Please cite this article in press as: H. Zhang et al., Multi-granularity principal curves extraction based on improved spectral clustering of complex distribution data, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.12.006

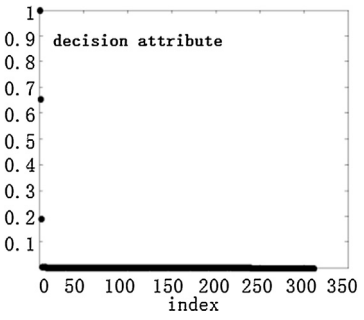

Fig. 3. Values of γ_i of Spiral dataset.

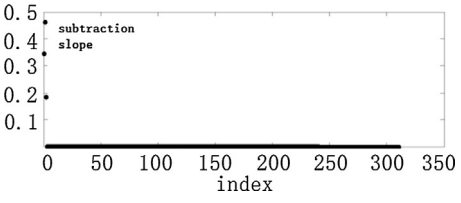

Fig. 4. Subtraction slope.

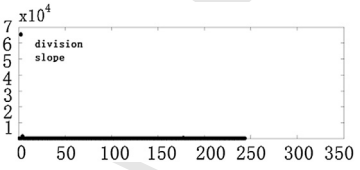

Fig. 5. Division slope.

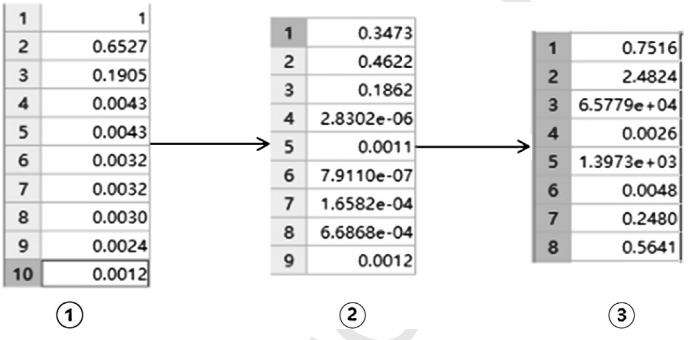

Fig. 6. Inflection point estimate.

The number of cluster centers is determined by finding the inflection point of the decision attribute distribution. The change of slope is maximal at the location of inflection point. Due to the x axis is 1, the slope can be obtained by subtracting the adjacent decision attributes, which is denoted as subtraction slope. As shown in Fig. 4, the subtraction slopes of the first few data are larger, and the subtraction slopes of the majority of the data are smaller. By observing the subtraction slopes of decision attribute, we find values of the subtraction slopes have a sharp difference between the non-cluster centers and the cluster centers. The sharp difference can be obtained by dividing the adjacent subtraction slopes, which is denoted as division slope. As shown in Fig. 5, the first convexity point (division slope of the point is larger than its adjacent points) is considered as an inflection point. That is, the number of the cluster centers is the index of the inflection point.

Fig. 6 shows a number of Spiral datasets with larger decision attributes. There are four steps to estimate the inflection point. Firstly, compute the decision attribute γ_i, as step 1 in Fig. 6. Secondly, obtain subtraction slope by subtracting the right adjacent decision attribute, as step 2 in Fig. 6. Thirdly, we obtain a division slope by dividing the right adjacent subtraction slope, shown in the step 3 of Fig. 6. Finally, take the first convex point as the inflection point and select the sorted index of the inflection point as the number of clusters. For the Spiral data in Fig. 6, the cluster number is 3.

To sum up, the complete algorithm is outlined as follows:

Algorithm 1. An improved spectral clustering based on inflection point estimate

Input: Gauss kernel parameter σ, sample points $X_N = \{x_1, x_2, \ldots, x_N\} \in R^d$

Output: Clustering result *cluster_labels*

Step 1: Compute the manifold distance $d_G(x_i, x_j)$ [25]. Compute the similarity between point and point produced by the Gaussian kernel. As shown below, we obtain the similarity matrix $A \in R^{n \times n}$.

$$A_{ij} = e^{\frac{-d_G(x_i, x_j)^2}{2\sigma^2}} \tag{4}$$

Please cite this article in press as: H. Zhang et al., Multi-granularity principal curves extraction based on improved spectral clustering of complex distribution data, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.12.006

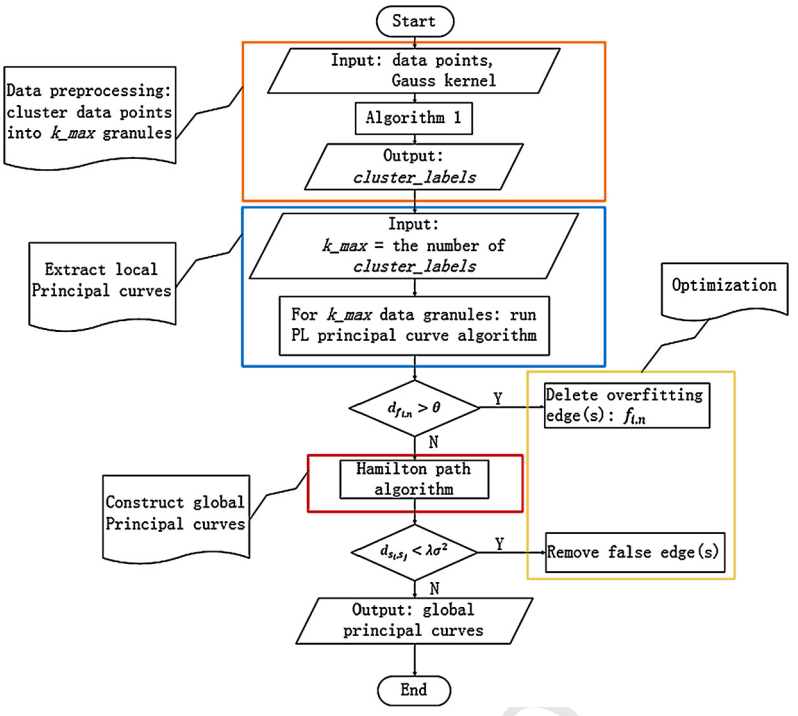

Fig. 7. The flowchart of Algorithm 2.

Step 2: Compute the Laplacian matrix L following the manifold distance kernel-based spectral algorithm [23].

Step 3: Suppose $\lambda_1, \lambda_2, \ldots, \lambda_m$ are m largest eigenvalues of the Laplacian matrix; $v^1, v^2, \ldots, v^m$ are eigenvectors associated with the corresponding eigenvalues. We construct the matrix $V = [v^1, v^2, \ldots, v^m] \in R^{n\times m}$, Only when the row i of V is assigned to class j, the sample X_i is assigned to class j.

Step 4: We take each row of V as the representative data of every initial data and use CFSFDP to cluster them. According to the two parameters ρ_i and δ_i, the decision variables γ_i are calculated and sorted from the largest to the lowest value. We use the inflection point estimation to calculate the number of cluster centers (n) and select n points which have the largest values of γ_i to be the cluster centers. Label other points as the same to the closest points whose density is larger.

Step 5: Get the clustering result *cluster_labels*.

The algorithm can solve the problem of manually inputting clustering number and instability. Meanwhile, the algorithm needs fewer parameters and has strong robustness in a certain range.

3. Multi-granularity principal curve algorithm based on improved spectral clustering

Assume that a set of numeric data $X_N = \{x_1, x_2, \ldots, x_N\} \in R^d$ is given. In the multi-granularity principal curve algorithm, we follow the strategy covered by following algorithm

Algorithm 2. Multi-granularity principal curve algorithm based on improved spectral clustering
Input: *cluster_labels*
Output: principal curves

Step 1: Form *k_max* granular data by running the improved spectral clustering algorithm;
Step 2: Extract the local principal curves of each granular data with the use of the PL principal curves algorithm [4];
Step 3: Delete overfitting edges with a designed objective function;
Step 4: Construct multi-granularity principal curves based on the Hamilton path algorithm;
Step 5: Optimize the global principal curves by removing false edges.

To clarify the essence of the algorithm, we illustrate it in Fig. 7.
In the following sections, we elaborate on these steps in more detail.

Please cite this article in press as: H. Zhang et al., Multi-granularity principal curves extraction based on improved spectral clustering of complex distribution data, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.12.006

3.1. Formation of the local principal curves

Considering information granulation as a powerful tool of processing complex data and addressing the instability of the existing principal curve algorithms, the improved spectral clustering proposed in Section 2 is used to construct granular data. Hence, Algorithm 1 is employed to form *k_max* granular data, where *k_max* is the number of clusters.

To extract the local principal curves of each granular data, we employ the PL principal curves algorithm [4]. The algorithm defines principal curves as continuous curves of a given length L, which minimizes the expected squared distance between the curve and points of the space randomly chosen according to a given distribution. The basic idea is to start with a straight line segment $f_{0,n}$, the shortest segment of the first principal component line which contains all of the projected data points, and in each iteration of the algorithm it increases the number of segments by one by adding a new vertex to the polygonal line $f_{k,n}$ produced in the previous iteration. After adding a new vertex, the positions of all vertices are updated so that the value of a penalized distance function becomes minimized.

The algorithm mainly consists of a projection step and an optimization step. In the projection step the data points are partitioned into "nearest neighbor regions" according to which segment or vertex they project on. The "nearest neighbor regions" is denoted as *Voronoi* region [4]. In the optimization step, the new position of a vertex is determined by minimizing an average squared distance criterion penalized by a measure of the local curvature, while all other vertices are kept fixed. These two steps are iterated so that the optimization step is applied to each vertex in a cyclic fashion (so that after the last vertex, the procedure starts again with the first vertex), until convergence has been achieved and $f_{k,n}$ produced. Then a new vertex is added.

The algorithm stops when k exceeds a certain threshold $c(n, \Delta)$. This stopping criterion is based on a heuristic complexity measure, determined by the number of segments k, the number of data points n, and the average squared distance $\Delta_n(f_{k,n})$.

During running the PL principal curves algorithm, in order to avoid overfitting, we construct a new cost function to delete false polygonal lines

$$d_{f_{i,n}} = \frac{num\_Voronoi'}{S(f_{i,n})} \tag{5}$$

where *Voronoi'* means the *Voronoi* region of the polygonal line for $f_{i,n}$, and *num_Voronoi'* means the number of the data points in the *Voronoi'* region. $S(f_{i,n})$ is the length the line $f_{i,n}$. If $d_{f_{i,n}} > \theta$, we delete $f_{i,n}$. If not, $f_{i,n}$ is preserved, where θ is an empirical parameter. The cost function implies that we delete those polygonal lines that can represent too few data points.

For all the *k_max* granular data, we extract the local principal curve $\{f^i_{0,n}, f^i_{1,n}, \ldots, f^i_{k,n}\}$ by PL principal curve algorithm, where $i = 1, 2, \ldots, k\_max$. Thus we can denote all of the local principal curves as $\{s_1, s_2, \ldots, s_{k\_max}\}$, where $s_i = \{f^i_{0,n}, f^i_{1,n}, \ldots, f^i_{k,n}\}$.

3.2. Construction of multi-granularity principal curves from local to global

After obtaining *k_max* local principal curves, multi-granularity global principal curves are constructed with the use of the Hamilton path algorithm. Firstly, we define a fully connected graph $G = (V, E)$, where the set of vertices V consists of the $2k\_max$ end-points of the *k_max* local principal curves. Also, we define a set of edges $A \subset E$ which contains all edges that correspond to the local principal curves. A sequence of the edges $\{(v_0, v_1), (v_1, v_2), \ldots, (v_{k\_max-2}, v_{k\_max-1}), (v_{k\_max-1}, v_{k\_max})\}$ in which all edges are distinct is called a 'path'. A path is 'open' if $v_0 \neq v_m$. An open path that passes through every vertex in the graph exactly once is called a 'Hamiltonian path' (HP). Note that we can consider a HP as a set $P \subset E$. We wish to find the HP P minimizing the total cost of the path, under the constraint: $A \subset P \subset E$. The cost function of a path P is defined as $l(P) + \lambda a(P)$, with $0 \leq \lambda \in \mathbb{R}$ being a parameter to be set manually. The term $l(P)$ denotes the length of the path, defined as the sum of the length of the edges in P. The length of an edge $e = (v_i, v_j)$ is taken as the Euclidean distance between its vertices, namely $l(e) = \|v_i - v_j\|$. The second term, $a(P)$, is a penalty term equal to the sum of the angles between adjacent edges. The parameter λ controls the trade-off between preferring short paths and paths that do not contain sharp turns. The $\lambda a(P)$ term is introduced to implement a preference for 'smooth' (not having sharp turns) curves. The smaller λ, is the smaller the preference for the smooth curves becomes.

Greedy strategy outlined below is employed to construct the global multi-granularity principal curve. A HP on a subset of V is called as a sub-HP. We start with the *k_max* local principal curves as *k_max* sub-HPs. At each step we connect two sub-HPs with an edge e. Note that the total cost of a (sub-)HP consisting of two sub-HPs P_i and P_j linked together by an edge e is the sum of the costs of each sub-HP plus $l(e)$ plus an angle penalty $a(e)$. Fig. 8 illustrates the angle penalty $a(e) = \alpha + \beta$ incurred by an edge $e = (v_i, v_j)$ that connects two sub-HPs P_i and P_j. We assign to each edge $e \in (E - A)$ cost $c(e) = l(e) + \lambda a(e)$. The edge that minimizes $c(e)$ over all edges that connects two sub-paths is inserted at each step. Summarizing, the procedure reads as follows:

1. Start with *k_max* sub-HPs defined by A.
2. While there are at least two sub-HPs.
3. Join those two sub-HPs P_i and P_j $i \neq j$ by edge $e \in (E - A)$ such that e minimizes $c(e)$ over all edges connecting two distinct sub-HPs.

Please cite this article in press as: H. Zhang et al., Multi-granularity principal curves extraction based on improved spectral clustering of complex distribution data, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.12.006

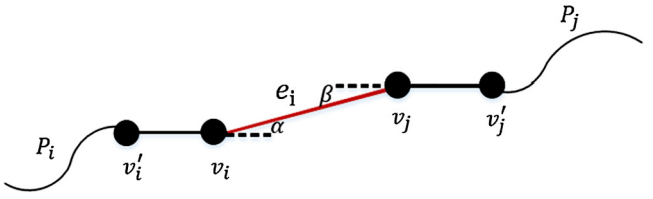

Fig. 8. Connecting two sub-paths: the angle penalty of (v_i, v_j) is the sum of the angles between an edge and the adjacent edges, i.e. $\alpha + \beta$.

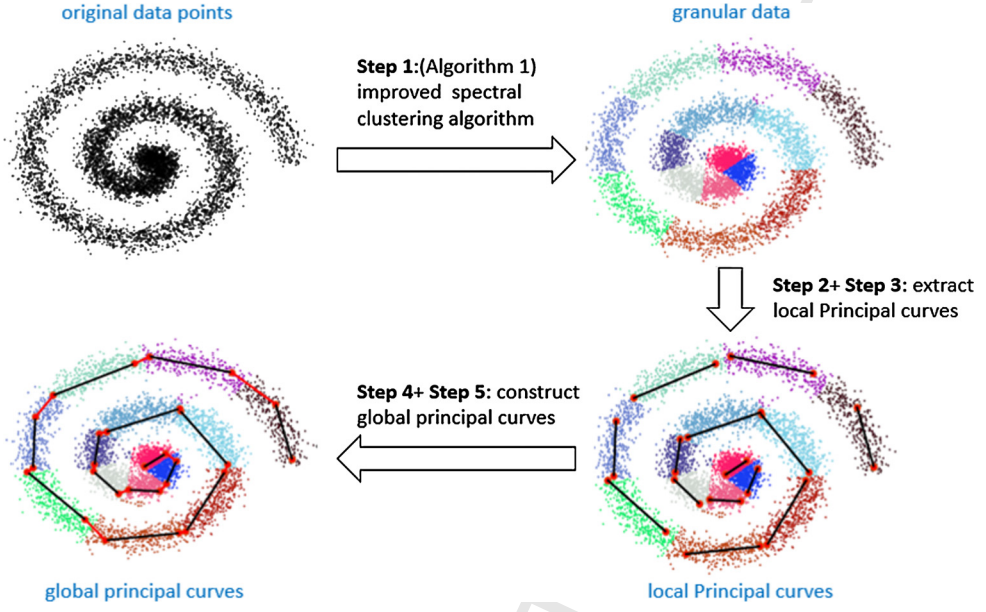

Fig. 9. Multi-granularity principal curves extraction.

To remove false edges in the process of connecting local principal curves to global principal curves, unnecessary connections between local principal curves should be removed. Firstly, we compute the distance between pairwise principal curves in $s_1, s_2, \ldots, s_{k\_max}$. Take s_i as an example: we compute the projection distance between the endpoints of other principal curves and s_i, denoted as d_{s_i,s_j}, where $i = 1, 2, \ldots, k\_max$. The adaptive threshold is selected as:

$$OF = \lambda\sigma_2^2 \tag{6}$$

If $d_{s_i,s_j} < \lambda\sigma^2$, delete the shorter line between s_i and s_j. If $d_{s_i,s_j} \geq \lambda\sigma^2$, preserve the two line segments. Here λ is the adaptive parameter, σ^2 is the variance of noise.

Fig. 9 illustrates an underlying idea of the proposed methodology. First, a large collection of similar objects is arranged together by running the improved spectral clustering algorithm to form a few granular data. Next, following the PL principal curve algorithm and the procedure of deleting overfitting edges, the local principal curves are obtained. Finally, based on the Hamilton path algorithm and false edge removal method, the global principal curve is gradually obtained and optimized.

With the optimization method outlined before, the advantage of the proposed algorithm is obvious. Firstly, instead of manual inputting of the granular number, the granulation of the initial data is more robust by means of the proposed improved spectral clustering algorithm (Algorithm 1). Secondly, with the elaborately designed cost function in step 3, the overfitting edges are deleted. The third advantage is the proposed adaptive threshold for removing the false edges in step 5.

3.3. Space and time complexity of the approach

As before, let N be the number of the data points, d be the dimensionality of the data, our algorithm extracts the local principal curves of one data granule at one time and we have k_max data granules in total. Hence, the space complexity of the proposed algorithm is $O(Nd/k\_max)$, while the space complexity of PL algorithm is $O(Nd)$.

As to time complexity, it is composed of two parts: the complexity of the improved spectral clustering algorithm and the complexity of the principal curves extraction process. The time complexity of the first part is $O(N^3)$, which is the same to the time complexity of NJW-Spectral clustering algorithm because the complexity of the inflection estimate can be ignored compared to NJW-Spectral clustering. Besides, the complexity PL algorithm is $O(kN^2)$, where k is the number of iterations

Please cite this article in press as: H. Zhang et al., Multi-granularity principal curves extraction based on improved spectral clustering of complex distribution data, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.12.006

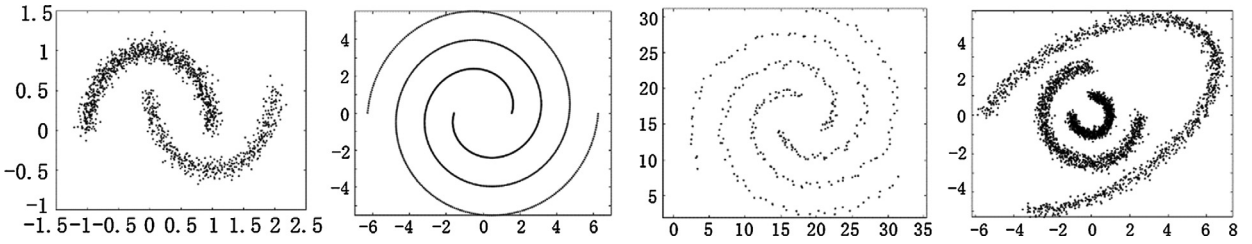

Fig. 10. Synthetic data sets.

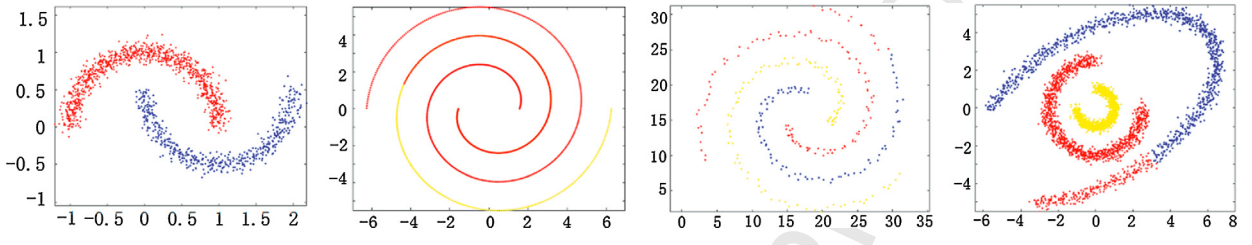

Fig. 11. Clustering result reported in [23].

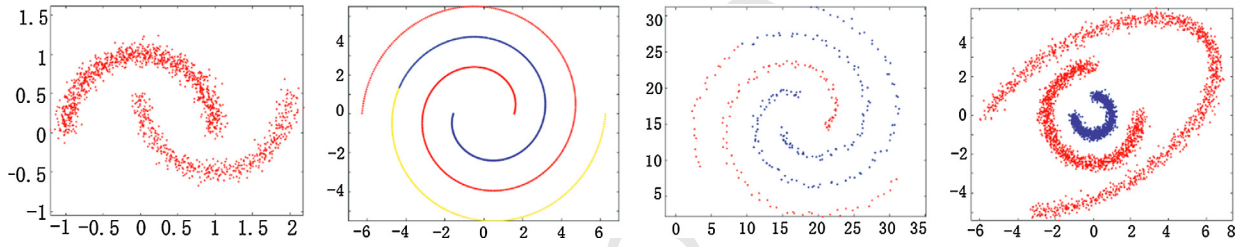

Fig. 12. Clustering result of CFSFDP [26].

of the PL algorithm. After the initial data have been granulated, the number of data which need to be dealt with by PL algorithm is reduced to $\frac{N}{k\_max}$. Thus, the complexity of the second part is $O(k\frac{N^2}{k\_max})$ In conclusion, the time complexity of the proposed algorithm is $O(N^3 + k\frac{N^2}{k\_max})$.

4. Experimental results and analysis

In this section, we first present a validation experiment to explain the effectiveness of the improved spectral clustering algorithm. Then, in order to demonstrate the effectiveness of the multi-granularity principal curve algorithm, we design two experiments to compare the PL principal curve algorithm and the proposed method. Finally, the impact of the objective function OF on the performance of the algorithm is investigated. All the experiments are completed for three public datasets, say Twomoons, Jain and Spiral and a single synthetic dataset (DisorderThreeCircles). The four datasets are displayed in Fig. 10.

4.1. Granulation results and analysis

Compared with [23] and the CFSFDP algorithm [26], we use four datasets to realize a comparative study. The parameters of the algorithm are the same and the number of clusters is selected manually as presented in [23].

As shown in Fig. 11, when the number of cluster is unknown, the algorithm proposed in [23] cannot produce good clustering results. The algorithm is not stable. Fig. 12 contains the clustering results produced by the CFSFDP algorithm. As visualized in Fig. 13, the proposed algorithm not only can determine the number of cluster centers, but also it yields very good results. The reasons are as follows.

Firstly, we use manifold distance to alleviate the drawback of the spectral clustering. After we get the representative of the original data, we invoke an effective method to determine the cluster number based on CFSFDP. In order to draw the

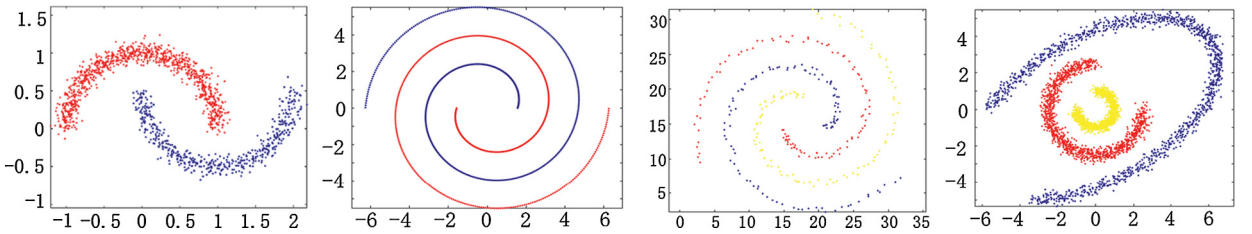

Fig. 13. Clustering results of improved spectral clustering algorithm based on inflection point estimate.

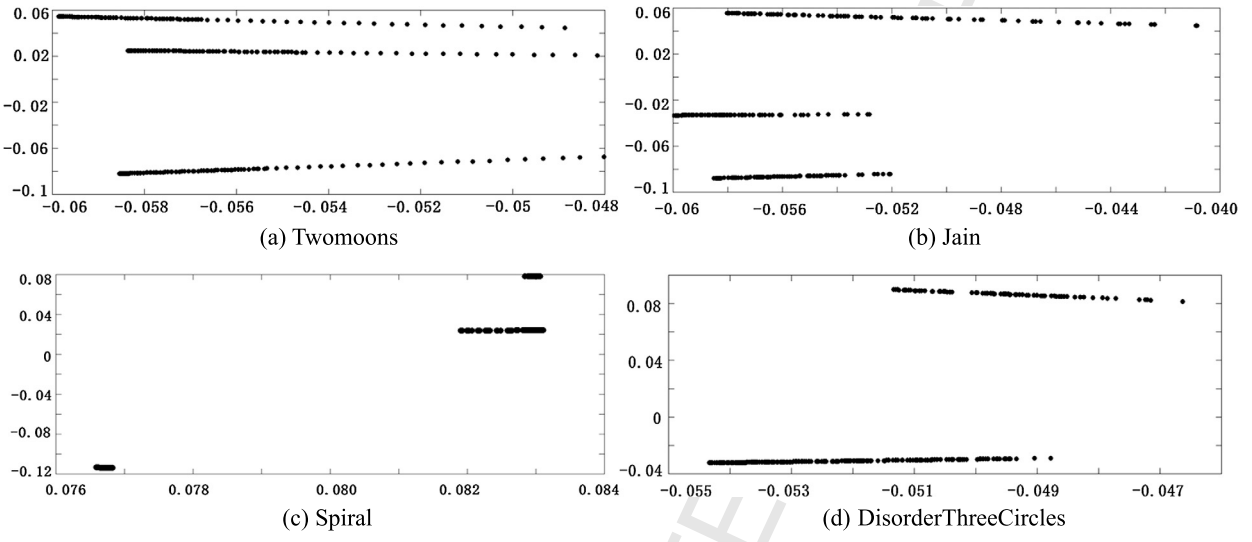

Fig. 14. Representative points of original data.

representative points of the original data in the two-dimensional plane, we take eigenvectors corresponding the first two largest eigenvalues, then the representative points are also two-dimensional.

As shown in Fig. 14, the representative points form spherical data and each cluster has some density extremum, which can be processed by the CFSFDP. Therefore, the use of inflection point estimate to automatically determine the cluster center makes the algorithm more accurate and stable.

In the calculation of similarity, we need to use the Gaussian kernel parameter σ and the neighbor number n. In the CFSFDP algorithm, we need neighborhood parameter k. The method of selecting neighbor number n is the one described in [23].

In order to observe the effect of kernel parameter σ and neighborhood parameter k on clustering results, we run the improved spectral clustering algorithm on the four datasets. By the means of counting the number of the data points which are assigned to incorrect clusters with different kernel parameter σ and neighborhood parameter k, we evaluate the performance of the improved spectral clustering algorithm. On the one hand, we keep $k = 20$ unchanged, and we change the kernel parameter σ by ranging it from 0.1 to 20. Experiment results show that there are no mistakenly clustered data. We can conclude that the proposed algorithm is robust to the Gaussian kernel parameter σ. On the other hand, Gaussian kernel parameter σ is set to 5 and six different neighborhood parameter k ranging from 10 to 100 is considered. When k is 10, we observed that some points are assigned to the wrong cluster on Jain and Spiral. The reason is that when the neighborhood range is too low, there is not a big difference in the density of the data, and data points tend to be assigned to adjacent clusters, especially when the data distribution is scattered (see Jain and Spiral). Thus, in the process of determining the cluster centers, errors occur. But when k assumes values above 20, the clustering results are accurate, which turns out that the algorithm is not sensitive to the parameter k.

4.2. Principal curves result and analysis

To evaluate the performance of the multi-granularity principal curves algorithm, it is tested on four complex datasets. Also, to realize a comparative study, the PL algorithm is tested as well. Fig. 15 and Fig. 16 show the results of the proposed algorithm and the PL principal curve algorithm. The black line segments shown in the figure are the principal curves extracted. We can note that the principal curves extracted by our algorithm can fit data better than the “conventional” PL principal curve algorithm. The curves are smoother and exhibit lower noise. On the contrary, the PL principal curve algorithm cannot ignore the false edges occurring between different granules.

Please cite this article in press as: H. Zhang et al., Multi-granularity principal curves extraction based on improved spectral clustering of complex distribution data, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.12.006

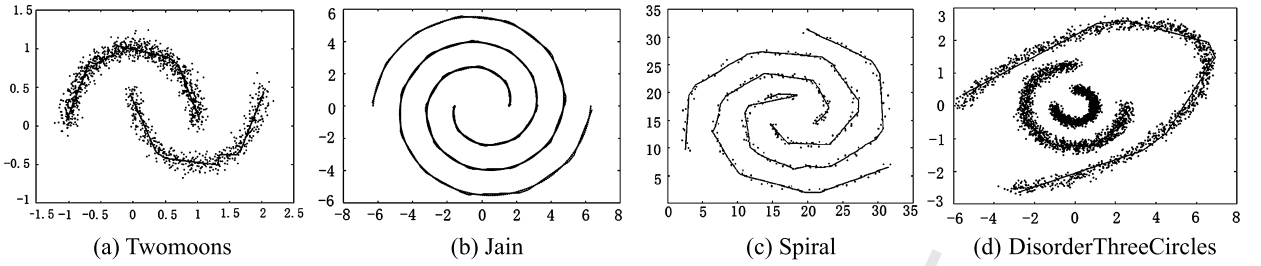

Fig. 15. Principal curves extracted by the proposed algorithm.

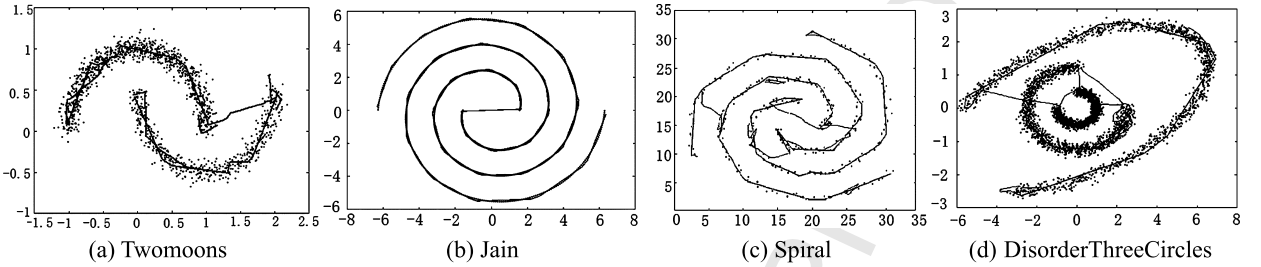

Fig. 16. Principal curves extracted by PL principal curve algorithm.

Table 1
The values of the objective function OF.

Distance	39.9815	7.2633	12.3076	8.7402	4.1456	0.1871	0.1498
Log_dis	3.1353	3.9556	3.5356	3.6982	3.7774	4.0919	4.0953
OF	39.9928	7.2775	12.3204	8.7535	4.1592	0.2018	0.1645

Based on the principle analysis, the traditional PL principal curve algorithm inserts a new segment by comparing all the points in the dataset. It calculates the sum of the distances between each point and the data points around it. Meanwhile, it calculates the sum of the distances between the surrounding data points and the nearest segment. If the two distance sum differ greatly, the region of the point and its surrounding points is taken as a data block to extract the first principal component segment. The principal curve extraction is completed for the entire dataset, which can result in overfitting and emergence of false edges. In the PL principal curve algorithm, there is no measure to rectify these problems. However, in our algorithm, we granulate the data first, which can generate data with simple distribution. Then, accurate local principal curves are obtained to construct global principal curves. The two improvements proposed here also cope well with the overfitting problem. As is shown in Fig. 15 and Fig. 16, the experimental results demonstrate that the multi-granularity principal curves exhibit some advantages when compared with the corresponding PL principal curves.

4.3. An impact of the objective function OF on the performance of our algorithm

In this section, the objective function OF is investigated. It determines the termination condition of the algorithm. Taking the Spiral dataset as an example, when the minimum value of OF is 0.1645 (see Table 1) and the maximum number of the principal curves is 8, overfitting problem appears (as shown in Fig. 17(a)). Therefore, it is unreliable to only rely on the objective function as the termination condition of the algorithm. To deal with the problem, we introduce an auxiliary function (step 3) and an adaptive parameter based on noise variance (step 5) in the Algorithm 2. Fig. 17(b) validates the effectiveness of the improvements.

4.4. Comparison of the running time

Table 2 shows the running time of PL algorithm and our algorithm applied to the four datasets. The running time of the algorithm is composed of two parts: the clustering time and the principal curve extraction time. With the use of the improved spectral clustering algorithm to granulate data, this is not surprising as our algorithm has no advantage in runtime. It is noticeable that the distribution of the dataset has an impact on the run time of both of the two algorithms. However, the principal curves extracted by our algorithm are more precise and smooth than those extracted by PL algorithm. Besides, our algorithm sacrifices a little bit of time complexity, but according to the analysis in Section 3, the space complexity of PL algorithm is *k_max* times the size of that of our algorithm.

Please cite this article in press as: H. Zhang et al., Multi-granularity principal curves extraction based on improved spectral clustering of complex distribution data, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.12.006

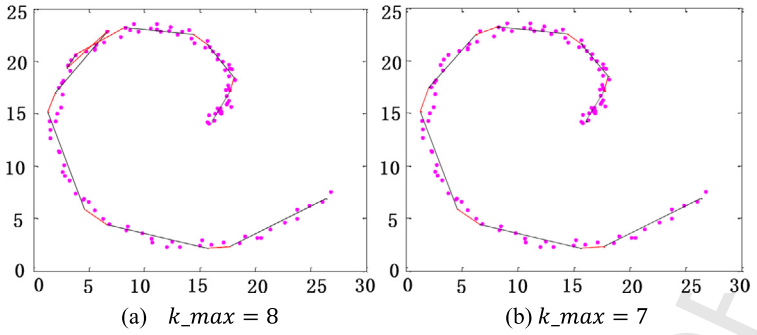

Fig. 17. Results obtained before and after optimization.

Table 2
Comparison of running time.

	Twomoons	Jain	Spiral	DisorderThreeCircles
PL algorithm	0.5071	34.3108	0.8667	4.3233
Our algorithm	1.0075	34.5324	0.9329	5.6917

5. Conclusions

For large scale complex data, although the traditional principal curve algorithm can greatly improve the processing speed, the performance of them is not stable. For instance, the existing PL principal curve algorithm still cannot well handle the complex data with self-intersecting characteristics, high curvature, and significant dispersion. Therefore, new approaches are badly needed to solve the principal curve learning problem realized in the presence of complex data. In our study, we combined the idea of GrC to improve the algorithm of the principal curve. We proposed the improved spectral clustering algorithm as the way of granulation and validation of its effectiveness. Based on the idea of GrC, this study designs and realizes the algorithm of extracting the principal curve from local to global, and compares and analyzes the performance of the algorithm in the public datasets and the artificial dataset. Experiments demonstrate that the constructed algorithm exhibits a significant level of stability and robustness. Several further pursuits can be envisioned including investigations dealing with set valued (interval) data, nominal data, and mixed data.

Acknowledgements

Authors would like to thank the anonymous reviewers for their constructive comments and valuable suggestions. This work is supported by the National Science Foundation of China (Grant Nos. 61573255, 61673301, 61673299, 61573259), National Key R&D Program of China (Grant No. 213), and Major Project of Ministry of Public Security (Grant No. 20170004).

References

[1] P. Baldi, K. Hornik, Neural networks and principal component analysis: learning from examples local minima, Neural Netw. 2 (1) (1989) 53–58.
[2] T. Hastie, W. Stuetzle, Principal curves, J. Am. Stat. Assoc. 84 (406) (1989) 502–516.
[3] R. Tibshirani, Principal curves revisited, Stat. Comput. 2 (4) (1992) 183–190.
[4] B. Kegl, A. Krzyzak, T. Linder, K. Zeger, Learning and design of principal curves, IEEE Trans. Pattern Anal. Mach. Intell. 22 (3) (2000) 281–297.
[5] P. Delicado, Another look at principal curves and surfaces, J. Multivar. Anal. 77 (1) (2001) 84–116.
[6] J. Verbeek, N. Vlassis, B.J.A. Kröse, A soft k-segments algorithm for principal curves, in: International Conference on Artificial Neural Networks, vol. 2130, 2001, pp. 450–456.
[7] J. Zhang, U. Kruger, X. Wang, D. Chen, A Riemannian distance approach for constructing principal curves, Int. J. Neural Syst. 20 (3) (2010) 209–218.
[8] H.Y. Zhang, W. Pedrycz, D.Q. Miao, Z.H. Wei, From principal curves to granular principal curves, IEEE Trans. Cybern. 44 (6) (2014) 748–760.
[9] D.C. Stanford, A.E. Raftery, Finding curvilinear features in spatial point patterns: principal curve clustering with noise, IEEE Trans. Pattern Anal. Mach. Intell. 22 (6) (2000) 601–609.
[10] H. Wang, T.C.M. Lee, Extraction of curvilinear features from noisy point patterns using principal curves, Pattern Recognit. Lett. 29 (16) (2008) 2078–2084.
[11] Z.Y. He, M.L. Gu, T.J. Wang, X.X. Shi, Principal component feature for speech recognition, J. Appl. Sci. 17 (4) (1999) 427–432.
[12] J.P. Zhang, D.W. Chen, U. Kruger, Adaptive constraint K-segment principal curves for intelligent transportation systems, IEEE Trans. Intell. Transp. Syst. 9 (4) (2008) 666–677.
[13] J.P. Zhang, X.D. Wang, U. Kruger, F.Y. Wang, Principal curve algorithms for partitioning high-dimensional data spaces, IEEE Trans. Neural Netw. 22 (3) (2011) 367–380.
[14] R.S. Bradley, P.J. Withers, Post-processing techniques for making reliable measurements from curve-skeletons, Comput. Biol. Med. 72 (2016) 120–131.
[15] Y.J. Yu, J. Wang, Enclosure transform for interest point detection from speckle imagery, IEEE Trans. Med. Imaging 36 (3) (2016) 769–780.
[16] Y.Y. Yao, Information granulation and rough set approximation, Int. J. Intell. Syst. 16 (1) (2001) 87–104.
[17] L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets Syst. 90 (90) (1997) 111–127.

[18] W. Pedrycz, P. Rai, A multifaceted perspective at data analysis: a study in collaborative intelligent agents, in: Special Issue on Cybernetics and Cognitive Informatics, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 39 (4) (2009) 834–844.
[19] Y.Y. Yao, Interpreting concept learning in cognitive informatics and granular computing, in: Special Issue on Cybernetics and Cognitive Informatics, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 39 (4) (2009) 855–866.
[20] W. Pedrycz, M.L. Song, Analytic hierarchy process (AHP) in group decision making and its optimization with an allocation of information granularity, IEEE Trans. Fuzzy Syst. 19 (3) (2011) 527–539.
[21] J.H. Yu, B. Zhang, M.H. Chen, W.H. Xu, Double-quantitative decision-theoretic approach to multigranulation approximate space, Int. J. Approx. Reason. 98 (2018) 236–258.
[22] Y.H. She, X.L. He, H.X. Shi, Y.H. Qian, A multiple-valued logic approach for multigranulation rough set model, Int. J. Approx. Reason. 82 (2017) 270–284.
[23] X.M. Tao, S.Y. Song, P.D. Cao, D.D. Fu, A spectral clustering algorithm based on manifold distance kernel, Inf. Control 41 (3) (2012) 307–313.
[24] Z. Halim, M. Waqas, A.R. Baig, A. Rashid, Efficient clustering of large uncertain graphs using neighborhood information, Int. J. Approx. Reason. 90 (2017) 274–291.
[25] J.B. Tenenbaum, V.D. Sliva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[26] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.
[27] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and algorithm, NIPS Proc. 14 (2001) 849–856.

Please cite this article in press as: H. Zhang et al., Multi-granularity principal curves extraction based on improved spectral clustering of complex distribution data, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.12.006