Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 3930957, 12 pages
https://doi.org/10.1155/2017/3930957

Research Article

Robust L-Isomap with a Novel Landmark Selection Method

Hao Shi,[1] Baoqun Yin,[1] Yu Kang,[1] Chao Shao,[2] and Jie Gui[3]

[1]*Department of Automation, University of Science and Technology of China, Hefei, China*
[2]*College of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou, China*
[3]*Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, China*

Correspondence should be addressed to Hao Shi; haoshi@mail.ustc.edu.cn

Received 17 October 2016; Accepted 16 March 2017; Published 24 May 2017

Academic Editor: Guangming Xie

Isomap is a widely used nonlinear method for dimensionality reduction. Landmark-Isomap (L-Isomap) has been proposed to improve the scalability of Isomap. In this paper, we focus on two important issues that were not taken into account in L-Isomap, landmark point selection and topological stability. At first, we present a novel landmark point selection method. It first uses a greedy strategy to select some points as landmark candidates and then removes the candidate points that are neighbours of other candidates. The remaining candidate points are the landmark points. The selection method can promote the computation efficiency without sacrificing accuracy. For the topological stability, we define edge density for each edge in the neighbourhood graph. According to the geometrical characteristic of the short-circuit edges, we provide a method to eliminate the short-circuit edge without breaking the data integrity. The approach that integrates L-Isomap with these two improvements is referred to as Robust L-Isomap (RL-Isomap). The effective performance of RL-Isomap is confirmed through several numerical experiments.

1. Introduction

Real-world data such as voice signals, gene microarray, or hyperspectral imagery data usually has a high dimensionality, which makes them difficult to analyze. This is known as the "curse of dimensionality" [1]. In order to analyze and process the high-dimensional real-world data adequately, data dimensionality reduction has been attracting significant interest. Generally, real-world data is found to lie on a low-dimensional manifold embedded in the high-dimensional observation space. Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. It can remove redundant information from the original data and alleviate the "curse of dimensionality" problem.

In the last few decades, many dimensionality reduction algorithms have been proposed. Principal component analysis (PCA) [2] and multidimensional scaling (MDS) [3] are traditionally linear methods. However, these linear methods fail to deal with complex nonlinear data. Currently, a number of nonlinear dimensionality reduction methods are available, for example, isometric feature mapping (Isomap) [4], local linear embedding (LLE) [5], and local tangent space alignment (LTSA) [6]. Among these methods, Isomap is representative of global manifold learning methods, which attempts to preserve global geometrical features of data set in the embedding space. It substitutes Euclidean distance in MDS for geodesic distance. Hindered by MDS and geodesic distance computation, Isomap would become very time-consuming as the amount of the input data increases. To improve the scalability of Isomap, L-Isomap was proposed by Silva and Tenebaum [7] in which the time-consuming computation was performed on a subset of points referred to as landmark points. However, two main problems still exist with L-Isomap.

The first problem is how to select landmarks for L-Isomap. So far, several landmark selection methods have been proposed. In [7], landmarks are selected randomly from the given data. An interesting approach called Fast-Isomap based on integer optimization was proposed in [8]. In [9], landmarks are chosen from an approximate central part of the given data. In this paper, we propose a novel method for landmark point selection. First, the method selects some points as candidates using a greedy strategy. After that, it obsoletes the candidate points that are neighbours of other candidate points. The remaining candidate points are determined as

the landmark points. Our method is proven to be more efficient without sacrificing accuracy by experiments.

Another critical problem of L-Isomap algorithm is its topological instability [10], which is caused by short-circuit edges [11]. In order to compute the geodesic distance, a neighbourhood graph is constructed by connecting every data point with its k-nearest neighbours. The short-circuiting problem occurs when an oversized neighbourhood is chosen so that the neighbourhood distance is larger than the distance between the folds in the manifolds. Short-circuit edges severely damage the manifold structure. Thus even a single short-circuit edge may produce many errors when computing geodesic distance which in turn severely impairs the performance of L-Isomap. A simple way to avoid short-circuit edges is to decrease the neighbourhood size. But a very small neighbourhood size will fragment the manifold into a large number of disconnected regions [10]. Thus, choosing an appropriate neighbourhood size requires a priori information about the global geometry of the high-dimensional manifold. However, it is difficult to know the information beforehand [4, 10]. H. Choi and S. Choi [12] solve the short-circuiting problem by removing data points with extremely large total flows when computing the shortest path. In this paper, we define edge density for every edge in the neighbourhood graph using multivariate kernel density estimation (KDE) and propose a method to make L-Isomap robust to short-circuit edge. Different from the former method which tries to delete abnormal points, our method aims at short-circuit edge itself. According to the geometric feature of the short-circuit edge that the areas which short-circuit edges go through have very few data points, the edge which has extraordinarily low edge density is claimed as the short-circuit edge and our method removes such edges from the neighbourhood graph. Numerical experiments on several data sets demonstrate the effectiveness of our method.

This paper is organized as follows. Section 2 briefly reviews the Isomap algorithm and L-Isomap algorithm. A novel landmark selection method is given in Section 3. Section 4 presents a method of eliminating short-circuit edges. Experiment results on several data sets are given in Section 5, in order to validate the useful performance of RL-Isomap. Finally, conclusions and future extensions are discussed in Section 6.

2. Isomap and L-Isomap

For the given original input data $X \in R^{D\times N}$ with N samples and D dimensions, Isomap attempts to embed those samples into a lower dimensional space $X \in R^{d\times N}$, while preserving the geodesic distance between all the input points as faithfully as possible. Isomap first constructs a weighted undirected neighbourhood graph $G = (V, E)$, where each node $v_i \in V$, corresponding to the point $x_i \in X$, is connected with its k-nearest neighbours and each edge $e_{ij} \in E$ is assigned weight D_{ij} that represents the Euclidean distance between points x_i and x_j. Second, Isomap computes the shortest path between every two points in the graph to approximate the geodesic distance D_{ij}^G using Dijkstra's or Floyd's shortest-path algorithm [13, 14]. The geodesic distance between all the data points forms the geodesic distance matrix D_G. Finally, Isomap applies MDS on matrix D_G to find the low-dimensional embedding.

So far, Isomap has been successfully applied in many different fields. Unfortunately, when the amount of input data, N, is too large, Isomap may become too time-consuming in terms of the shortest-path construction ($O(kN^2 \log N)$) and the MDS eigenvalue decomposition ($O(N^3)$). In order to speed up these two computations, L-Isomap is proposed. L-Isomap randomly selects n points from X, denoted as landmark points. Instead of computing the shortest path between all data points, L-Isomap only computes the shortest path from each data point to the landmark points. Then classical MDS is applied to the resulting $n \times N$ geodesic distance matrix to find the low-dimensional embedding of the landmark points. The embeddings of the remaining points are obtained by a fixed linear transformation of their geodesic distance to the landmark points. This way the time complexities of the shortest path and the MDS computation are, respectively, reduced to $O(knN \log N)$ and $O(n^2 N)$.

3. Landmark Selection

3.1. Landmark Candidate. A very important procedure for L-Isomap is to build the k-nearest neighbourhood graph. However, many neighbourhoods are similar because they share common points. Thus, some neighbourhoods can be deleted to get a simpler neighbourhood graph. Based on this idea, we select the candidate landmark points by simplifying neighbourhood graph. Formally, let $\Omega = \{S_1, S_2, \ldots, S_N\}$ be the neighbourhood set, one for each point $x_i \in X$, where $\bigcup_{i=1}^{N} S_i = X$, and each set $S_i \in \Omega$ is assigned a nonnegative cost $c(S_i)$. The goal is to find a set cover $\Omega^* = \{S_{i_1}, S_{i_2}, \ldots, S_{i_v}\} \subset \Omega$, satisfying $\bigcup_{r=1}^{v} S_{i_r} = X$ and minimizing $c(\Omega^*) = \sum_{r=1}^{v} c(S_{i_r})$. Once the cover Ω^* is obtained, the corresponding landmark candidates are determined. The problem can be resolved by a greedy strategy. Let u_{S_i} be the number of uncovered points in S_i and the ratio of S_i is $r_{S_i} = u_{S_i}/c(S_i)$ which counts the number of points covered by S_i per unit cost. The probability of including S_i in Ω^* increases with the ratio r_{S_i}. The landmark candidate selection problem is an unweighted case, where $c(S_i) = 1$. A sketch of the landmark candidate selection method can be summarized as Algorithm 1.

Algorithm 1 can run in time $O(N \log N)$ and it achieves an approximation ratio of $H(p)$, where

$$H(p) = \sum_{k=1}^{p} \frac{1}{k}, \tag{1}$$

and p is the size of the largest set in Ω [15, 16]. We apply Algorithm 1 to get the neighbourhood subset $\Omega^* = \{S_{i_1}, S_{i_2}, \ldots, S_{i_v}\}$ and the corresponding landmark candidate set $X_c = \{x_{i_1}, x_{i_2}, \ldots, x_{i_v}\}$, where x_{i_r} corresponds to each neighbourhood S_{i_r} in Ω^* and $r \in \{1, 2, \ldots, v\}$.

3.2. Landmark Selection Algorithm. The landmark candidates may be neighbours of each other and some of them are unnecessary [17]. Thus we can further optimize the set of

(1) Let $\Omega^* = \Phi$.
(2) If $u_{S_i} = 0$ for all i then stop: Ω^* is the cover, where $i = 1, 2, \ldots, N$. Otherwise find a subscript $i_r \in \{1, 2, \ldots, N\}$, maximizing the ratio $u_{S_{i_r}}/c(S_{i_r})$ and proceed to Step (3).
(3) Add S_{i_r} to Ω^*, replace each S_i by $S_i - S_{i_r}$ and return to Step (1).

Algorithm 1: Selection of landmark candidates.

(1) $L = X_c$
(2) for $j = i_1 : i_v$
if $\beta_j \neq \Phi$, then for all r, let $\beta_{i_r} = \beta_{i_r} - \beta_j$, delete β_j from B, and $L = L - x_j$,
end
(3) L is the landmark point set.

Algorithm 2: Landmark selection.

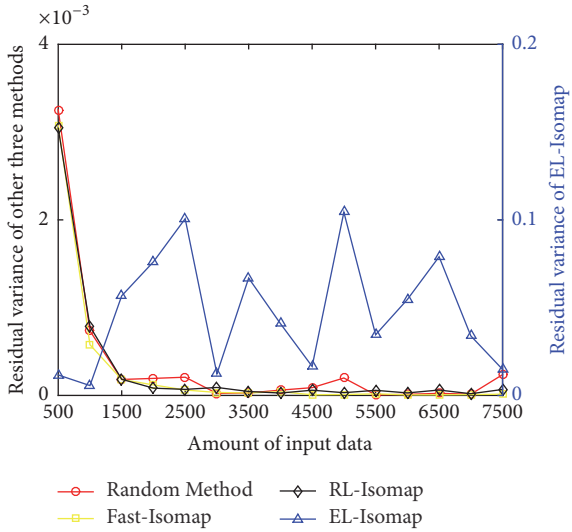

Figure 1: Residual variance comparison of four methods.

landmark candidate points to obtain the landmark points that are nonneighbouring to each other. Formally, let $B = \{\beta_{i_1}, \beta_{i_2}, \ldots, \beta_{i_v}\}$, where β_{i_r} represents the set of points whose neighbourhood include x_{i_r} and $r \in \{1, 2, \ldots, v\}$. The landmark point selection method can be summarized as Algorithm 2.

In order to test the effectiveness of our landmark selection method, we, respectively, run RL-Isomap and other three methods on Swiss roll data, where the amount of the input data varies from 500 to 7500. Tenebaum et al. [4] used residual variance to characterize how well the embedding result preserves the geometry features of the high-dimensional manifold. The smaller the residual variance value, the better the embedding result. As shown in Figure 1, EL-Isomap has the worst performance of the four methods in the experiment because it always tends to select landmarks which are situated around the circumcenter [9]. The random scheme performs well but it is not stable enough because of its random strategy and unpredictability. Fast-Isomap has similar performance to RL-Isomap; however, the number of landmark points selected by RL-Isomap is much smaller than that of Fast-Isomap as shown in Figure 2(a). In this experiment, the landmark points selected by Fast-Isomap is averagely 48.55% more than RL-Isomap. As a result, RL-Isomap will be faster than Fast-Isomap when computing the low-dimensional embedding. As shown in Figure 2(b), RL-Isomap is averagely 33.56% faster than Fast-Isomap in the experiment. So RL-Isomap performs best in both speed and accuracy in this experiment.

4. Robust L-Isomap

4.1. Short-Circuit Edge Elimination. As pointed out in Section 1, L-Isomap faces topological instability problem because of the short-circuit edge. An oversized neighbourhood might result in short-circuit edges that destruct the manifold structure. As shown in Figure 3, the short-circuit edge, denoted by black solid line, directly links up two points which are supposed to be very far on the manifold. A simple way to avoid these short-circuit edges is to decrease the neighbourhood size; however too small neighbourhood will break the connectivity of the neighbourhood graph, as shown in Figure 4. Thus it is not an easy job to choose an appropriate neighbourhood size. The previous method tries to delete some outliers to mitigate the short-circuiting problem. But the method breaks the integrity of the data set. In this section, we present a novel method which directly deletes the short-circuit edges in the neighbourhood graph.

As shown in Figure 3, the areas which short-circuit edges go through usually have very sparse data points. Based on this fact, we claim that the edge that goes through the area which has extremely low data density is a short-circuit edge. In order to quantify the data density of the area which the edges go through, we introduce edge density for each edge in the neighbourhood graph using KDE method.

KDE is a nonparametric tool for estimating the distribution of data [18]. The multivariate KDE is a direct extension of the univariate estimator. For the D-variate random sample

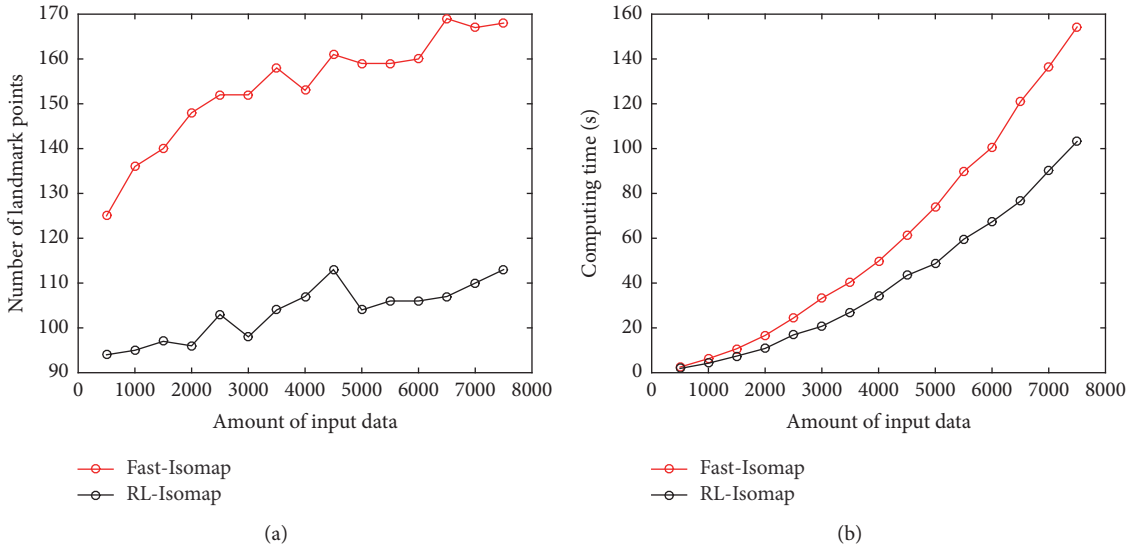

Figure 2: Comparison between RL-Isomap and Fast-Isomap.

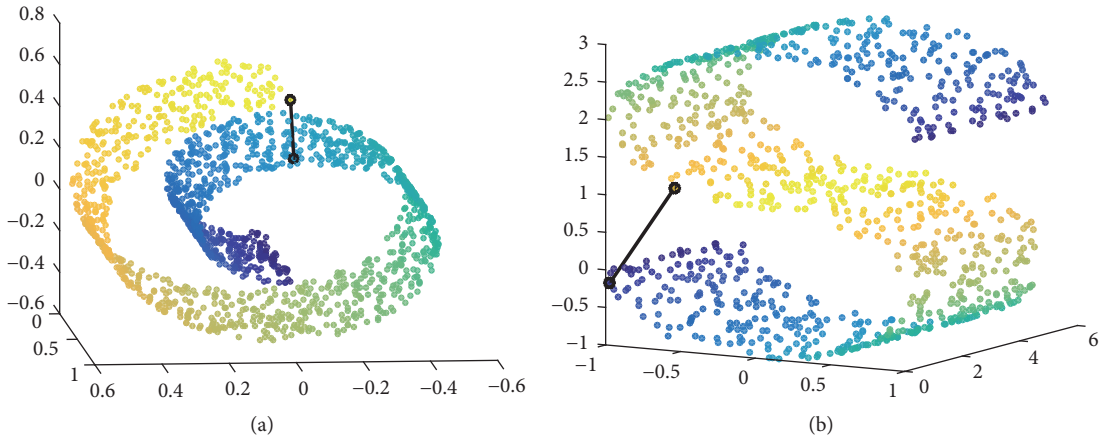

Figure 3: Short-circuit edges connecting two surfaces on different manifolds.

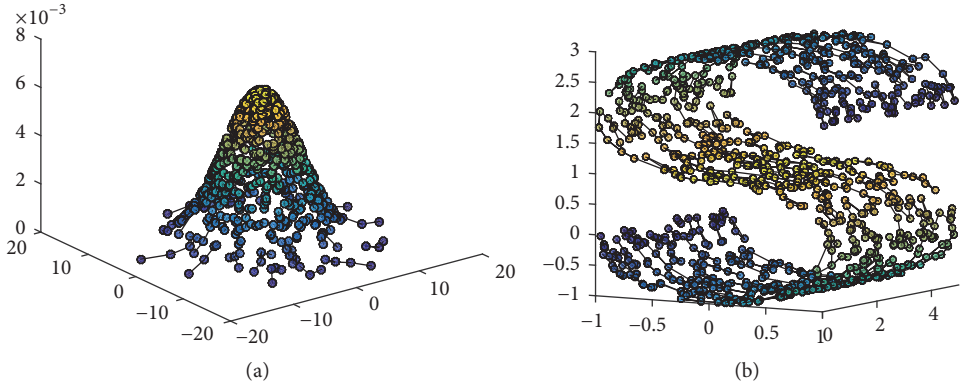

Figure 4: Disconnectivity caused by too small neighbourhood size.

$z_1, z_2, \ldots, z_m$ having density f, the D-dimensional KDE is defined as follows:

$$\widehat{f}(z;H) = m^{-1}\sum_{i=1}^{m} K_H(z - z_i), \tag{2}$$

where $z \in \mathbb{R}^D$, H is a symmetric positive $D \times D$ matrix called the bandwidth matrix, and

$$K_H(z) = |H|^{-1/2} K\left(H^{-1/2}z\right), \tag{3}$$

where $|H|$ is the determinant of H and K is D-dimensional kernel function that satisfies

$$\int K(z)\, dz = 1, \tag{4}$$

and $\int$ is shorthand for $\int \cdots \int_{\mathbb{R}^D}$.

However, KDE is linear method, such that it cannot be applied directly on complex nonlinear data. In mathematics, the manifold is a topological space that locally resembles Euclidean space near each point. More precisely, each point of a D-dimensional manifold has a neighbourhood that is homeomorphic to the D-dimensional Euclidean space [5]. Therefore, the points within the same neighbourhood can be approximately seen to have linear structure. For the input data $X \in R^{D\times N}$, L-Isomap uses k-NN rule to build the neighbourhood graph G with edge set $E = \{e_{ij} \mid i > j,\ x_i \in S_j \text{ or } x_j \in S_i\}$, where S_i is the neighbourhood of point x_i and e_{ij} denotes the edge connecting points x_i and x_j. Let $\mathcal{F}_i = \{x_i\} \cup S_i$ and, according to the local linearity properties of the manifold, data points in $\mathcal{F}_i$ have nearly linear structure. This way, we define the point density of the manifold at point x_i as

$$\widehat{g}_i(x;H) = \left|\mathcal{F}_i\right|^{-1} \sum_{x_u \in \mathcal{F}_i} K_H(x - x_u), \tag{5}$$

where $|\mathcal{F}_i|$ denotes the number of points included in $\mathcal{F}_i$.

Given an edge e_{ij}, the quartiles of e_{ij} can be calculated as follows:

$$e_{ij}^m = \frac{(4-m)\cdot x_i + m \cdot x_j}{4}, \quad m = 1, 2, 3. \tag{6}$$

After that, we have

$$\widehat{g}_{ij}^m(x;H) = \left|\mathcal{F}_i \cup \mathcal{F}_j\right|^{-1} \sum_{x_u \in \mathcal{F}_i \cup \mathcal{F}_j} K_H(x - x_u). \tag{7}$$

Then let

$$\widehat{q}_{ij} = \frac{1}{3}\sum_{m=1}^{3} \widehat{g}_{ij}^m\left(e_{ij}^m; H\right). \tag{8}$$

We define edge density of each edge e_{ij} in E as follows:

$$d_{ij} = \frac{\widehat{q}_{ij}}{\max\left(\widehat{g}_i(x_i;H), \widehat{g}_j(x_j;H)\right)}. \tag{9}$$

(1) Define a neighbourhood graph by using k-NN rule.
(2) Eliminate the edges whose d_{ij} is below η.
(3) Select landmark candidates by Algorithm 1.
(4) Determine landmark by Algorithm 2.
(5) Apply LMDS to find a low-dimensional embedding.

Algorithm 3: Robust L-Isomap.

In this paper, we take the standard D-variate normal density, $K(x) = (2\pi)^{-D/2}\exp(-(1/2)x^Tx)$ as the kernel function. The choice of bandwidth matrix H is intractable in KDE [19]. In the univariate case, too big values of h will make the estimate too smooth and may not uncover the structural features, whereas small values of bandwidth h yield "wiggly" estimate and show spurious features [20]. In the multivariate case, the choice of the bandwidth matrix H faces the same dilemma. Fortunately, our experiments show that the choice of H has little effect on the results because the edge density values of short-circuit edges are much lower than that of the normal edges. For simplicity, we take the unit matrix I as bandwidth matrix.

We compute the edge density for every edge in E by (9) and have $D = \{d_{ij} \mid i > j,\ x_i \in S_j \text{ or } x_j \in S_i\}$. As mentioned above, short-circuit edges have extremely low edge density. In such case, if d_{ij} is less than a certain threshold, we believe that the corresponding edge e_{ij} is a short-circuit edge and remove it from the neighbourhood graph. For example, edge density values for Swiss roll data are illustrated for two different sizes of neighbourhoods, where $k = 10$ and $k = 15$ in Figure 5. In Figure 5(c), there are a few short-circuit edges in the neighbourhood graph where $k = 15$. In this case, a few edges have extremely low edge density values shown in Figure 5(d). While, for the case of $k = 10$, the neighbourhood graph in Figure 5(a) is very healthy and the corresponding edge density values are well-distributed as shown in Figure 5(b). From Figure 5, it is clear that the edge density values of short-circuit edges are much lower than that of normal edges. In order to determine the threshold, we sort all the elements in D in ascending order and use d_{ij} which has the maximum increment in the first half of the sequence as our threshold. The threshold η can be obtained adaptively as follows.

(1) Arrange all elements in D in ascending order and the first half of the sequence is $d_1 \le d_2 \le \cdots \le d_{[|D|/2]}$, where $[|D|/2] = \max\{n \in \mathbb{Z} \mid n \le |D|/2\}$ with $\mathbb{Z}$ denoting the set of all integers.

(2) $\eta = d_t$, where $t = \min\{\arg\max_k(d_k - d_{k-1})\}$, and $k = 1, 2, \ldots, [|D|/2]$.

The algorithm that integrates L-Isomap with the landmark selection method and the short-circuit edge elimination method is called RL-Isomap. The main procedure of RL-Isomap is summarized as Algorithm 3.

The previous method proposed by H. Choi and S. Choi [12] must recompute the neighbourhood graph because a few points which have extremely high total flows have been eliminated from the original input data set. On the contrary, RL-Isomap directly eliminates the short-circuit edges from

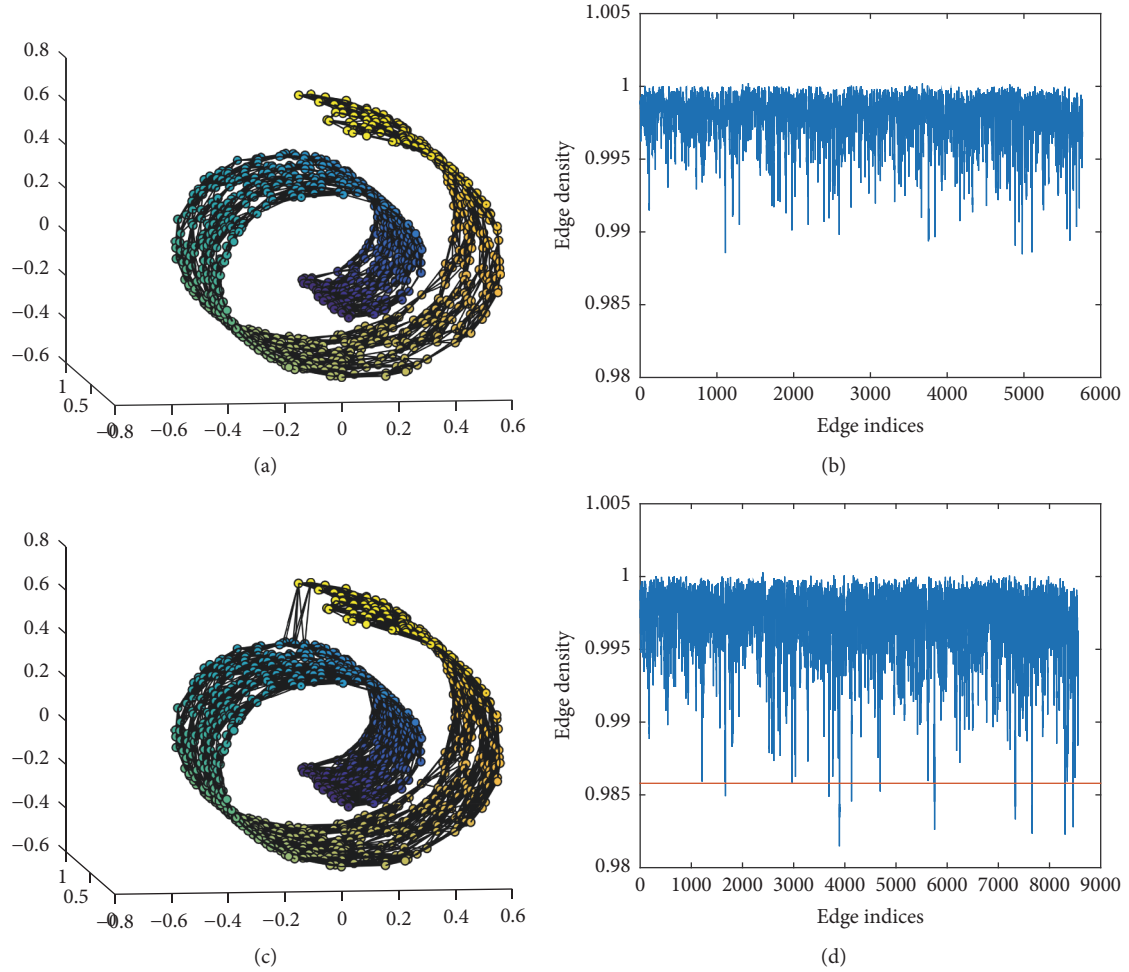

Figure 5: Edge density of each edge in the neighbourhood graph for the case of Swiss roll data. The number of the data points is 1000. (a) The k-nearest neighbourhood graph with $k = 10$. (b) The edge density values are well-distributed for the case of $k = 10$. (c) The k-nearest neighbourhood graph with $k = 15$. There are some short-circuit edges in this case. (d) The orange line denotes the threshold $\eta = 0.9858$. Some edges have extremely low edge density values that are considered as short-circuit edges.

the neighbourhood graph and will not destroy the data's integrity.

4.2. Complexity of RL-Isomap. RL-Isomap algorithm runs in $O(N^2 + k^2N + N \log N + v^2 + knN \log N + n^2N)$, where $O(N^2)$ is the complexity of constructing the neighbourhood graph of all input data. Then, we compute the edge density for the neighbourhood graph to eliminate the short-circuit edges and its time complexity is $O(k^2N)$. Next term $O(N \log N)$ is the complexity using Algorithm 1 to get the landmark candidate points. Algorithm 2, determining the landmark points, runs in $O(v^2)$ where v is the number of the candidate points. The fifth term $O(knN \log N)$ is the complexity of computing the geodesic distance using Dijkstra's algorithm. The last term $O(n^2N)$ is the complexity of computing the embedding of the input data.

5. Numerical Experiments

In this section, we conduct several experiments on different data sets: (1) Swiss roll data; (2) Toroidal Helix data; (3) face image data, to test RL-Isomap. After that, we use RL-Isomap to analyze the nonlinear structure of the high-dimensional Internet traffic matrix.

5.1. Experiment on Swiss Roll Data. In Figure 5(c), 1000 data points were used to build the neighbourhood graph where the neighbourhood size $k = 15$. In such case, there are several short-circuit edges appearing in the neighbourhood graph shown in Figure 5(c). First, we apply Robust Kernel Isomap [12] to the neighbourhood graph. After deleting two points which are considered as outliers by Robust Kernel Isomap, a new neighbourhood graph is shown in Figure 6(a) and there are still some short-circuit edges left. In such case, the embedding result by Robust Kernel Isomap is not correct shown in Figure 6(e). The distribution of edge density values for the neighbourhood in Figure 5(c) is shown in Figure 5(d) and it is clear that some values are extremely low. We then apply RL-Isomap to compute the threshold, marked by an orange line in Figure 5(d), where the threshold $\eta = 0.9858$ and the edges whose edge density values are below the threshold are deleted to get a new neighbourhood

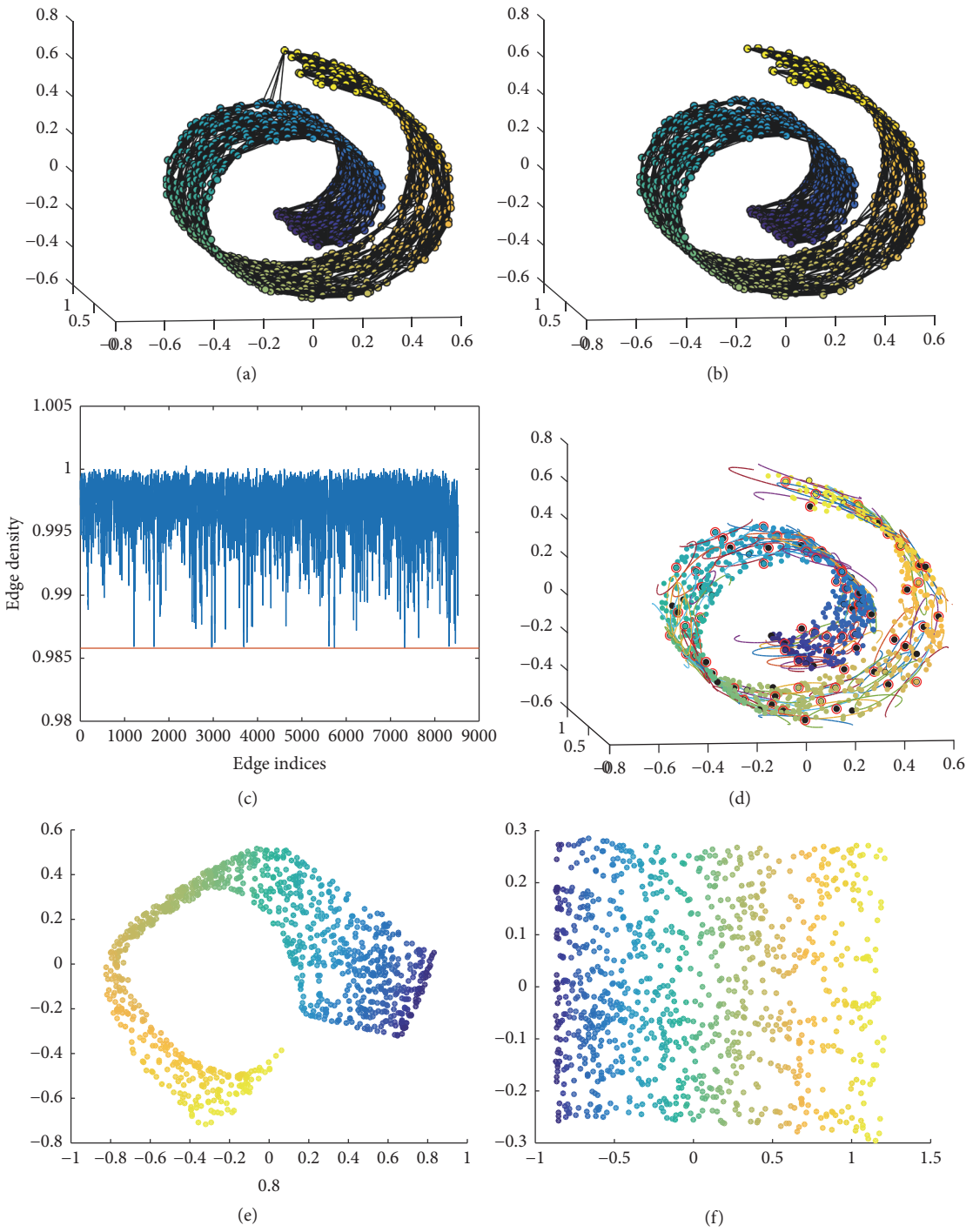

FIGURE 6: Comparison of RL-Isomap with Robust Kernel Isomap for the case of Swiss roll data. (a) Neighbourhood graph by Robust Kernel Isomap. (b) Neighbourhood graph after removing the edges whose edge density values are below the threshold by RL-Isomap. (c) The distribution of the edge density values of edges in the new neighbourhood graph. (d) The landmark candidate points and the landmark points. (e) Embedding result by Robust Kernel Isomap. (f) Embedding result by RL-Isomap.

graph that is shown in Figure 6(b) where the short-circuit edges disappear. Then we recalculate the edge density value for the edges in the new neighbourhood graph and the results are shown in Figure 6(c). It is clear that the edge density values are well-distributed by this time. Figure 6(d) shows the landmark candidate points (black solid points) and the landmark points (marked by red circles) obtained, respectively, by Algorithms 1 and 2. As shown in Figure 6(f), RL-Isomap finds the correct two-dimensional embedding. In this experiment, RL-Isomap outperforms Robust Kernel Isomap

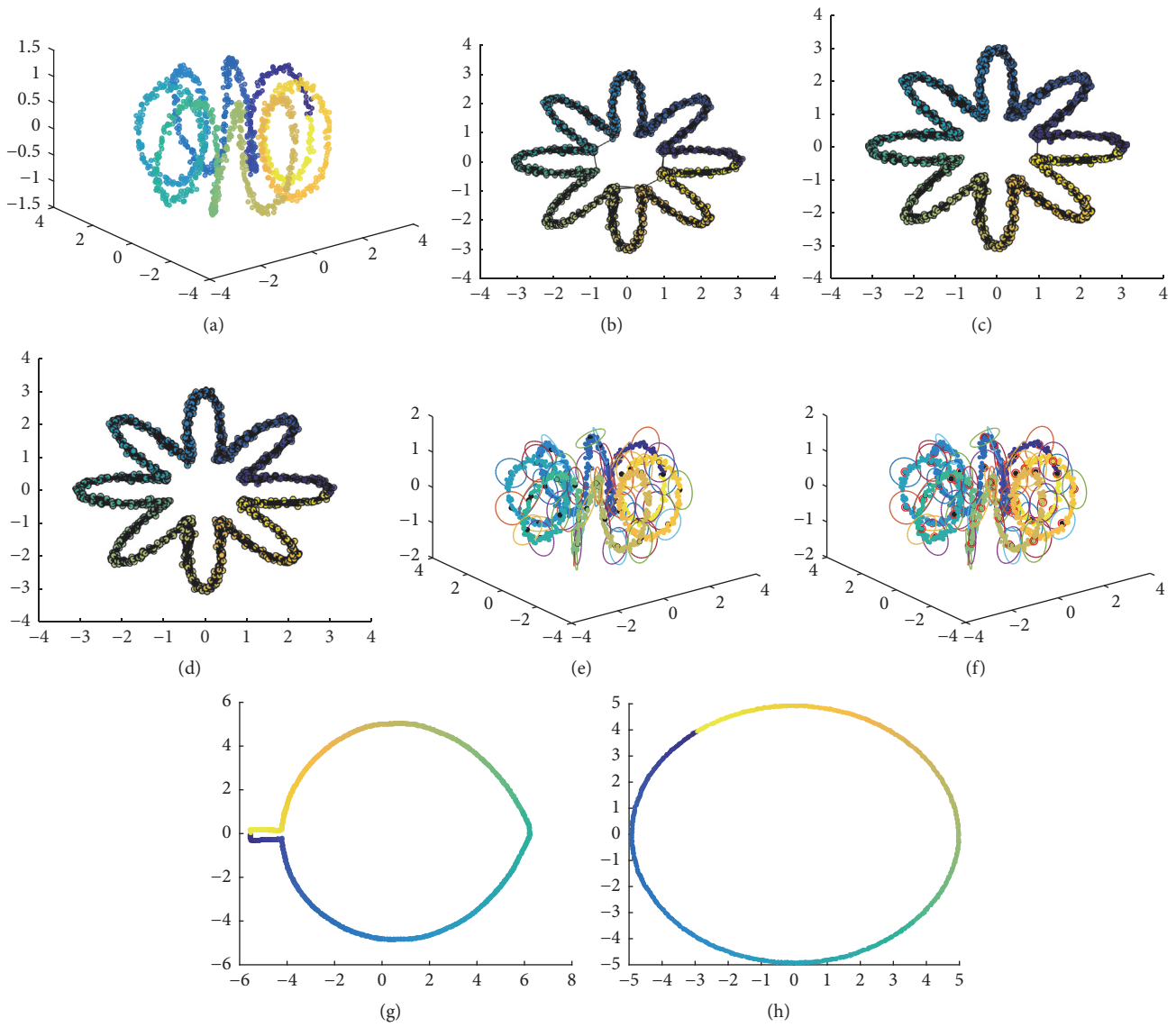

Figure 7: Comparison of RL-Isomap with Robust Kernel Isomap for the case of Toroidal Helix data. (a) The Toroidal Helix data. (b) The k-nearest neighbourhood graph with $k = 25$. (c) Neighbourhood graph constructed by Robust Kernel Isomap. (d) The new neighbourhood graph after removing the edges whose edge density values are below the threshold. (e) The landmark candidate points. (f) The landmark points. (g) Embedding result by Robust Kernel Isomap. (h) Embedding result by RL-Isomap.

and the experiment results demonstrate the effectiveness of RL-Isomap.

5.2. Experiment on Toroidal Helix Data. Toroidal Helix data is shown in Figure 7(a), with data number $N = 1200$. The neighbourhood graph is constructed with $k = 25$ and Figure 7(b) gives an overhead view of the neighbourhood graph and in such case there are a few short-circuit edges. First, we use Robust Kernel Isomap and 8 points are removed. After that, Robust Kernel Isomap reconstruct the neighbourhood graph, but there are still short-circuit edges existing, shown in Figure 7(c) and the corresponding embedding result in Figure 7(g) is not correct. Then we use (9) to compute edge density for each edge in the neighbourhood graph in Figure 7(b) and the values are illustrated in Figure 8(a) where the threshold $\eta = 0.9313$. The edges whose edge density is below the threshold are removed from the neighbourhood graph and then get a new neighbourhood graph without short-circuit edge shown in Figure 7(d). The edge density values of the new neighbourhood distribute more evenly than that of the original neighbourhood graph (see Figure 8(b)). After that, we run Algorithm 1 on the new neighbourhood graph and get the landmark candidate points denoted by black solid point shown in Figure 7(e) and then Algorithm 2 is applied to the candidate points and the resulting landmark points, marked by red circles, are illustrated in Figure 7(f). Finally, RL-Isomap finds the correct two-dimensional embedding, shown in Figure 7(h), that faithfully preserves the geometric features of the original manifold. This experiment further validates the effectiveness of RL-Isomap.

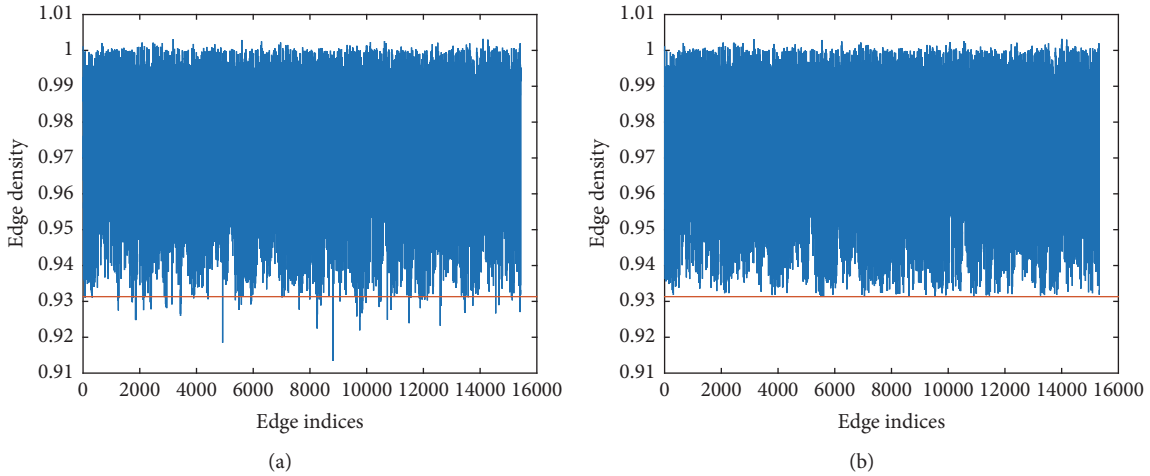

Figure 8: The comparison of distribution of edge density value between the two neighbourhood graphs.

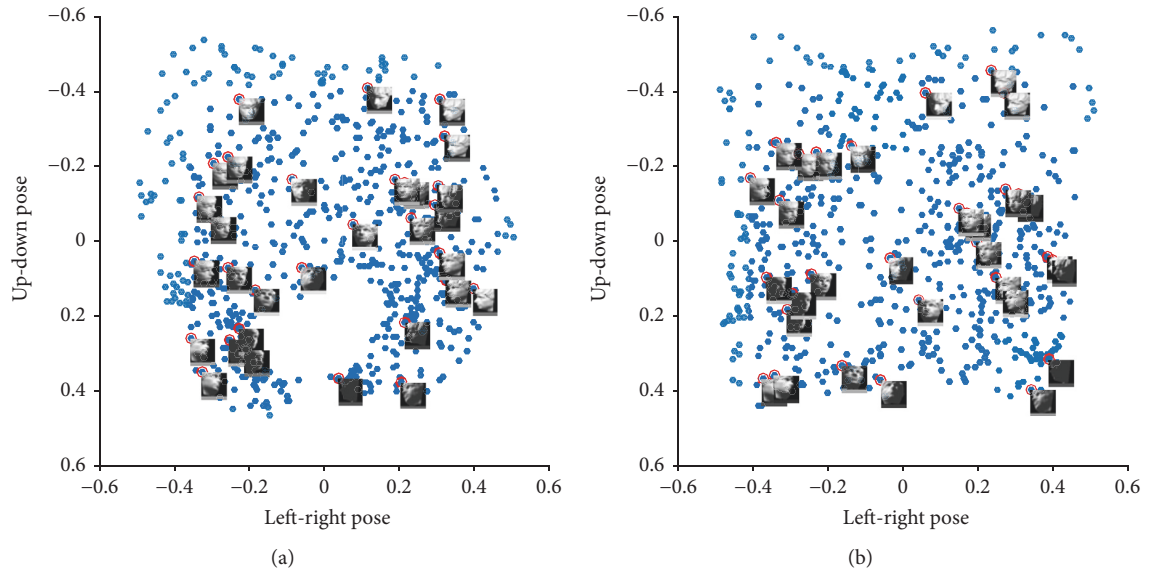

Figure 9: The two-dimensional projection of the face data.

5.3. Experiments on Face Data. In addition to the previous experiments on synthetic data set, we apply RL-Isomap on real-world data, face images. The input consists of 64-pixel-by-64-pixel images of a face with different lighting directions and poses, represented by a sequence of 4096-dimensional vectors. Thus the data set actually is a 3-dimensional manifold embedded in the 4096-dimensional observational space. We apply L-Isomap and RL-Isomap, respectively, on the data to detect its intrinsic geometric structure, where $k = 10$. As shown in Figure 9, representative faces are shown next to circled points in different parts of the space where each coordinate axis of the embedding result corresponds with one degree of freedom underlying the initial data: up-down pose and left-right pose. The grayscale of a 5×64 square bar below each representative face represents lighting direction. In Figure 9(a), the embedding result of L-Isomap is not correct because some points representing left-pose are projected close to the points representing the right pose which is clearly caused by short-circuiting problem. In Figure 9(b), RL-Isomap correctly finds the 3-dimensional embedding of the input data.

5.4. Nonlinear Structure Analysis of Traffic Matrix by RL-Isomap. Internet traffic matrix has been a useful traffic data model to understand the Internet from the whole-network perspective, but Internet traffic matrix usually possesses high-dimensional attributes [21]. The experiments above have proved the effectiveness of RL-Isomap, so, in this part, we apply RL-Isomap to analyze the intrinsic nonlinear structures of the Internet traffic matrix.

5.4.1. Traffic Matrix Modelling. An Origin-Destination (OD) flow comprises all traffic originating from a given source and delivering to a given destination. Let N denote the number of all sources and destinations in a network. A traffic matrix is naturally represented by a three-dimensional,

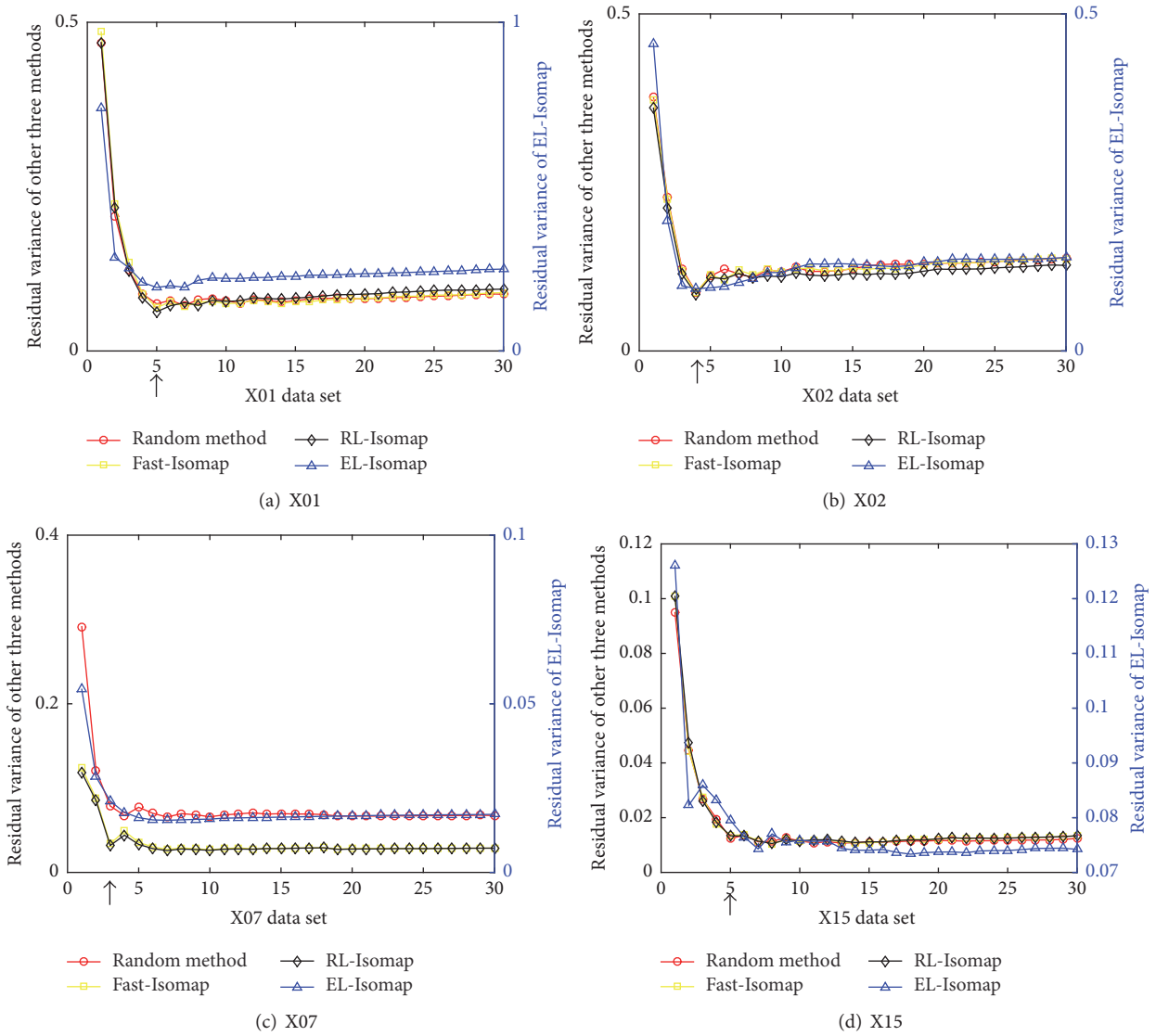

Figure 10: Dimension and residual variance.

nonnegative hypermatrix $\mathbf{X}(t)$, with i, jth entry $X_{i,j}(t)$. Each entry represents the traffic volume, measured in terms of bytes or packets, from source i to destination j in time interval $[t, t+\Delta t) \subset T$, the full measurement interval being denoted by T. It is intractable to monitor the traffic in real-time; thus the measurements are restricted to the average traffic in a discrete interval. The choice of Δt depends on the applications and available measurements [22]. In this paper, the data set we obtained was sampled from the backbone network (Abilene) over time intervals of 5 minutes.

The data set in this paper was downloaded from http://www.cs.utexas.edu/~yzhang/research/AbileneTM/, the publicly available set of traffic matrix [23]. The experimental environment was like a typical network with 12 PoPs; hence there are 144 PoP pairs and 144 OD flows. Thus, the traffic matrix has dimensionality of 144.

5.5. Intrinsic Dimensionality Analysis. The original traffic matrix data possesses 144 dimensions; thus it is very difficult to analyze them directly. Through the analysis above, RL-Isomap gives a simple way to analyze the high-dimensional input data and find their low-dimensional structures. In this section, we try to explore the underlying nonlinear structure hidden behind the high-dimensional traffic matrix data using RL-Isomap. We apply the algorithm to the 24 data sets, X01, X02, ..., X24 to learn the relations between the residual variance and the dimensions, and the results are illustrated in Figure 10 (Due to limited space, we only present 4 data sets in this paper.). As mentioned above, residual variance is used to characterize how well the low-dimensional Euclidean embedding preserves the geodesic distances estimated from the neighbourhood graph. The intrinsic dimension of the data can be estimated by looking for the "elbow" where the curve of the residual variance stops decreasing significantly with added dimensions. The intrinsic dimensions of these 24 data sets are shown in Figure 11. It is clear that the intrinsic dimensions of these data sets fluctuate from 3 to 8, which are far less than 144.

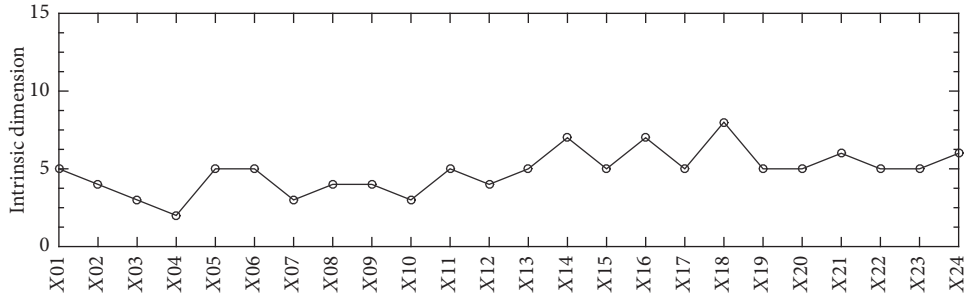

Figure 11: The intrinsic dimension of the 24 data sets.

Why does low dimensionality in OD traffic matrix exist? Two causes bring about this kind of low dimensionality. First, if the magnitude of variation among dimensions in the original data differs greatly, then the data may have low effective dimension for that reason alone. This is the case if variation along a small set of dimensions in the original data is dominant [24]. Second, it is caused by the spatial correlation of Internet traffic. Internet consists of core network and edge network. The traffic flows through different ingresses and egresses of the core network may originate from the same edge network, such that the traffic from different OD pairs shares some common patterns or trends.

As discussed above, we know that there truly exists nonlinear manifold structure hidden behind the original traffic matrix data. The low-dimensional structure can help analyze the flow characteristic of the network, including the traffic variation trend and traffic anomaly. These thoughts drive us to explore more of the nonlinear structure of the traffic matrix in the future.

6. Conclusion

In this paper, we present RL-Isomap algorithm. Two methods are given in RL-Isomap that, respectively, address the landmark selection problem and stable instability problem in L-Isomap. The experiments on synthetic data sets and physical data sets demonstrate the effectiveness of RL-Isomap.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grants nos. 61233003, 61202285, 61572463, and 61673361, in part by Research Fund for the Doctoral Program of Higher Education of China under Grant no. 20123402110029, in part by Natural Science Research Program of the Anhui High Education Bureau of China under Grant no. KJ2012A286, in part by the grant of the Open Project Program of the State Key Lab of CAD & CG under Grant A1709, Zhejiang University, and in part by the grant of the Shanghai Key Laboratory of Intelligent Information Processing, China under Grant IIPL-2016-003.

References

[1] D. W. Scott, "The curse of dimensionality and dimension reduction," in *Multivariate Density Estimation: Theory, Practice, and Visualization*, pp. 195–217, 2008.

[2] I. T. Jolliffe, *Principal Component Analysis*, Wiley Online Library, 2nd edition, 2002.

[3] M. A. A. Cox. and T. F. Cox, *Multidimensional scaling*, vol. 59 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London, UK, 1994.

[4] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[5] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[6] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.

[7] V. D. Silva and J. B. Tenebaum, "Global versus local methods in nonlinear dimensionality reduction," *Advances in Neural Information Processing Systems 15*, pp. 705–712, 2003.

[8] Y.-K. Lei, Z.-H. You, T. Dong, Y.-X. Jiang, and J.-A. Yang, "Increasing reliability of protein interactome by fast manifold embedding," *Pattern Recognition Letters*, vol. 34, no. 4, pp. 372–379, 2013.

[9] D. Liang, C. Qiao, and Z. Xu, "Enhancing both efficiency and representational capability of isomap by extensive landmark selection," *Mathematical Problems in Engineering*, vol. 2015, 18 pages, Article ID 241436, 2015.

[10] M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum et al., "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, Article 7, 2002.

[11] J. A. Lee and M. Verleysen, "Nonlinear dimensionality reduction of data manifolds with essential loops," *Neurocomputing*, vol. 67, pp. 29–53, 2005.

[12] H. Choi and S. Choi, "Robust kernel Isomap," *Pattern Recognition*, vol. 40, no. 3, pp. 853–862, 2007.

[13] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[14] R. W. Floyd, "Algorithm 97: shortest path," *Communications of the ACM*, vol. 5, no. 6, article 345, 1962.

[15] D. S. Johnson, "Approximation algorithms for combinatorial problems," *Journal of Computer and System Sciences*, vol. 9, pp. 256–278, 1974.

[16] V. Chvatal, "A greedy heuristic for the set-covering problem," *Mathematics of Operations Research*, vol. 4, no. 3, pp. 233–235, 1979.

[17] H. Shi, B. Yin, Y. Bao, and Y. Lei, "A novel landmark point selection method for L-ISOMAP," in *12th IEEE International Conference on Control and Automation ICCA '16*, pp. 621–625, Nepal, June 2016.

[18] M. P. Wand and M. C. Jones, "Kernel smoothing," Crc Press.

[19] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, UK, 1986.

[20] B. A. Turlach, *Bandwidth Selection in Kernel Density Estimation: A Review*, Universit catholique de Louvain, Louvain-la-Neuve, Belgium, 1993.

[21] H. Shi, B. Yin, X. Zhang, Y. Kang, and Y. Lei, "A landmark selection method for L-Isomap based on greedy algorithm and its application," in *54th IEEE CDC 2015 Conference on Decision and Control*, pp. 7371–7376, Japan, December 2015.

[22] P. Tune and M. Roughan, "Internet traffic matrices: a primer," *Recent Advances in Networking*, vol. 1, pp. 1–41, 2013.

[23] S. Uhlig, B. Quoitin, J. Lepropre, and S. Balon, "Providing public intradomain traffic matrices to the research community," *ACM SIGCOMM Computer Communication Review*, vol. 36, pp. 83–86, 2006.

[24] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," *SIGMETRICS Performance Evaluation Review*, vol. 32, pp. 61–72, 2004.