An Improved Isomap Algorithm Based onLP-Centers

W.W. Chen, W.J. Mao
State Key Laboratory of Industrial Control Technology
Institute of Cyber-Systems and Control
Zhejiang University
Hangzhou, P.R.China

***Abstract*—The isometric feature mapping (Isomap) is one of the widely used low-dimensional embedding methods. However, Isomap has three bottlenecks: sensitivity to outliers, high computational complexity and topological instability caused by short-circuit edges. To address above problems, by introducing the concepts of linear patch (LP) and k-nearest center (k-NC), a LP-center based Isomap (LPC-Isomap) approach is proposed in this paper. Firstly, before constructing a neighborhood graph, outliers are eliminated by LPC-Isomap. Then LP-centers are selected reasonably as landmark points by LPC-Isomap to reduce computational complexity. We compute the intersections between k-NC of each pair of points. If an intersection is empty, the edge connecting the corresponding pair of points is considered as a short-circuit edge. Lastly, the useful behavior of LPC-Isomap is verified through numerical examples with several data sets.**

Keywords-Isomap; linear patch (LP); k-nearest center (k-NC); LPC-Isomap

I. INTRODUCTION

Isomap is one of representative algorithms [1-3] in manifold learning community, which extends classical multidimensional scaling (MDS) by exploiting geodesic distances instead of Euclidean distances. In contrast to previous algorithms for nonlinear dimensionality reduction, Isomap computes a globally optimal solution and has lower residual variance [1]. Despite of the superiority of global optimization, there also exist some problems about Isomap.

The first problem of Isomap is its sensitivity to outliers, which may collapse the total manifold structure [4]. A lot of algorithms have been developed to detect outliers [5-7]. However, they also caused mis-deletion of normal points when neighborhood size is large. In order to overcome this situation, BORDER algorithm [8] is adopted in this paper, which detects the outliers with their reverse k-nearest neighbor (RkNN) number.

The second problem is the high computational complexity of Isomap. Isomap with landmark points (L-Isomap) was developed to solve this bottleneck [9]. In L-Isomap algorithm, some points are selected to be landmark points and it only calculates the distances from each data point to each landmark point. But it is a hard work to pick out the appropriate landmark points. Fast Isomap based on minimum set coverage is a feasible road to solve the problem [10]. Unfortunately, it only applies to the situation of low curvature. The centers of the linear patches (LP-centers) are taken as the landmark points to solve this problem.

The third problem is the topological instability caused by short-circuit edges which may devastate global manifold structure. An easy way to avoid this short-circuit edge is to decrease the neighborhood size. Several methods of selecting neighbourhood size were presented [11], [12], and an adaptive neighbourhood selection algorithm proposed [13] has achieved a better result. In addition, some algorithms select neighborhood according to the dissimilarity costs between each pair of points [14], [15]. Another feasible method is to find out all the short-circuit edges and remove them from the neighborhood graph. Obviously, the latter is better because of the free choice of the neighborhood size. To realize the latter idea, we compute the LP-centers, and find out the k-nearest center (k-NC) for each data point. An edge will be regarded as a short-circuit edge if there is no intersection between the k-NC of the two points it connects.

The rest of this paper is organized as follows. Next section introduces relevant definitions proposed in this paper. Section 3 develops our new algorithm, LPC-Isomap. Numerical experiments are shown in Section 4, in order to demonstrate the better performance of LPC-Isomap. Conclusions are summarized in Section 5.

II. RELEVANT CONCEPTS

Before introducing the LPC-Isomap algorithm, the relevant definitions and formulas must be put forward. The first is the definition of a linear patch:

Definition 1 (*d-Dimensional Linear Patch*): Given a D-dimensional $\boldsymbol{X}$, and a user defined threshold ε_0, $\boldsymbol{X}$ is a d-dimensional linear patch ($d \ll D$), denoted as $dLP(\boldsymbol{X})$, if we project $\boldsymbol{X}$ to d-dimensional linear space by linear analysis, *e.g.* principal component analysis (PCA), with a projection error smaller than ε_0.

The projection error ε of data set $\boldsymbol{X}$is measured by the form:

$$\varepsilon = \sum_{i=d+1}^{D} \lambda_i \Big/ \sum_{i=1}^{D} \lambda_i \quad (1)$$

where $\boldsymbol{X}= [\boldsymbol{x}_1,\boldsymbol{x}_2, \ldots,\boldsymbol{x}_N] \in \boldsymbol{\mathcal{R}}^{D\times N}$, λ_i is the eigenvalues of the covariance matrix of $\boldsymbol{X}$, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D$.

A linear patch $\boldsymbol{LP}$ can be considered as a class, which employs Gaussian distribution as its discriminant function[16], denoted as $N(\boldsymbol{x})$. We can split a D-dimensional sample data $\boldsymbol{S}$ into several linear patches $\boldsymbol{LP}_1$, $\boldsymbol{LP}_2$,..., when take appropriate values for projection error thresholdε_0 and destination dimensionality$d$. A center of $\boldsymbol{LP}_i$, denoted as $\boldsymbol{\mu}_i$, refers to the D-dimensional mean vector of $\boldsymbol{LP}_i$. For $\forall \boldsymbol{x} \in \boldsymbol{S}$, the k-nearest center (k-NC)of $\boldsymbol{x}$ can be defined as follows:

Definition 2 (*k-Nearest Center*): Given a set of linear patches $\boldsymbol{LP}_i$ with centers$\boldsymbol{\mu}_i$ and subordinate degree functions $N_i(\bullet)$, and a data point $\boldsymbol{x}$, k nearest centers of $\boldsymbol{x}$ are a set of centers$\boldsymbol{\mu}_{i_j}$where $N_{i_j}(\boldsymbol{x})$ is the first k largest values of $N_i(\boldsymbol{x})(1\leq j\leq k)$.

III. MAIN RESULTS

In this section, LPC-Isomap will be proposed based on the concepts of section 2. BORDER algorithm is originally designed to find boundary points and outliers are extraordinary boundary points, being away from densely distributed data. The data points whose RkNN number equals a user defined threshold η or less are output as outliers. Generally, η is set to 1.

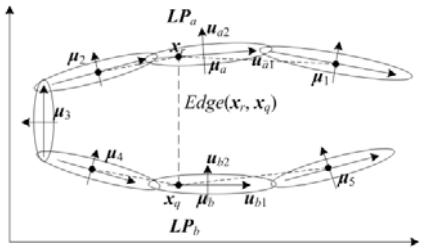


FIGURE I. THE ELLIPSES SHOW THE CONTOURS OF LINEAR PATCHES, AND $\boldsymbol{M}_I$(I =A, B, 1, 2,…) ARE LP-CENTERS. $\boldsymbol{U}_{AI}$ AND $\boldsymbol{U}_{BI}$(I = 1, 2) ARE ORTHONORMALEIGENVECTORS OF COVARIANCE MATRICES OF $\boldsymbol{LP}_A$ AND $\boldsymbol{LP}_B$, RESPECTIVELY.$EDGE(\boldsymbol{X}_R, \boldsymbol{X}_Q)$ IS A SHORT-CIRCUIT EDGE CONNECTING $\boldsymbol{X}_R$ AND $\boldsymbol{X}_Q$. K-NC SIZE K_CIS 3. $\boldsymbol{M}_A$, $\boldsymbol{M}_1$ AND $\boldsymbol{M}_2$ ARE K-NC OF $\boldsymbol{X}_R$. $\boldsymbol{M}_B$,$\boldsymbol{M}_4$ AND $\boldsymbol{M}_5$ ARE K-NC OF $\boldsymbol{X}_Q$

After BORDER, we will divide the sample data set $\boldsymbol{S}$in to several linear patches $\boldsymbol{LP}_1$, $\boldsymbol{LP}_2$,…, with the user defined projection error threshold ε_0 and destination dimensionality$d$. Firstly, we calculate the projection error ε_i of each point $\boldsymbol{x}_i\in\boldsymbol{S}$ and its k-NN, and search the point $\boldsymbol{x}_m$ with minimum projection error $\boldsymbol{x}_m$ satisfying $\varepsilon_m<\varepsilon_0$. Then $\boldsymbol{x}_m$ and its k-NN are marked as a new linear patch $\boldsymbol{LP}_j$, and points not in $\boldsymbol{LP}_j$ nearest to the LP-center $\boldsymbol{\mu}_j$ are added to $\boldsymbol{LP}_j$ continuously until projection error ε^{P_j} of $\boldsymbol{LP}_j$ $\geq\varepsilon_0$. The above procedures are repeated until all points are added to the linear patches or there is no point $\boldsymbol{x}$ satisfying $\varepsilon<\varepsilon_0$.

Considering the flexibility of our proposed algorithm, we design an overlapping factorα($0\ \leq\alpha\leq1$). According the properties of Gaussian distribution, the contour of a linear patch $\boldsymbol{LP}$ is a hyper-ellipsoid, with the center at $\boldsymbol{\mu}$ and axes oriented along the eigenvectors of covariance matrix of$\boldsymbol{LP}$, and with scaling factors in the directions of the axes givenby corresponding eigenvalues. When we shrink the contour of$\boldsymbol{LP}_j$, the point $\boldsymbol{x}_i$ belongs to overlapping region if it satisfies:

$$M_j(\boldsymbol{x}_i) > (1-\alpha)\max_{\boldsymbol{y}\in\boldsymbol{LP}_j} M_j(\boldsymbol{y}) \quad (2)$$

where $M_j(\boldsymbol{x})$ is the Mahalanobis distance [16].The points not covered by linear patches are considered as outliers.

For $\forall\boldsymbol{x}_r\in\boldsymbol{LP}_a,\boldsymbol{x}_q\in\boldsymbol{LP}_b$, if $Edge(\boldsymbol{x}_r,\boldsymbol{x}_q)$ is a short-circuit edge connecting $\boldsymbol{x}_r$ and $\boldsymbol{x}_q$, $Edge(\boldsymbol{x}_r,\boldsymbol{x}_q)$ passes through a region containing few data points and the variances of $\boldsymbol{LP}_a$ and $\boldsymbol{LP}_b$ along the direction of $Edge(\boldsymbol{x}_r,\boldsymbol{x}_q)$ are very small. Therefore, both $N_a(\boldsymbol{x}_q)$ and $N_b(\boldsymbol{x}_r)$ are so small that the inequality $N_j(\boldsymbol{x}_r) > N_b(\boldsymbol{x}_r)$ or $N_j(\boldsymbol{x}_q) > N_a(\boldsymbol{x}_q)$holds for most of the linear patches $\boldsymbol{LP}_j$ in sample data. Therefore, there is no intersection between k-nearest center(k-NC) of $\boldsymbol{x}_r$ and $\boldsymbol{x}_q$. An edge will be marked as a short-circuit edge if there is no intersection between k-NC of the two points it connects(see fig. 1).

The principle of selecting landmark points is that the information of intrinsic manifold structure can be recoveredas much as possible according to the landmark points. In other words, more landmark points are required in the regionwith relatively large curvature. More linear patches will be generated in the place of large curvature because of containing fewer points. Therefore, the LP-centers can be exploited as landmark points. For convenience, we select a point which is nearest to the LP-center and in the same linear patch as a landmark point instead.

TABLE I. LPC-ISOMAP ALGORITHM

Input: D-dimensional data set $\boldsymbol{S}$composed of data points $\boldsymbol{x}_i$; Projection error threshold ε_0, Overlapping factor α and k-NC size k_c;
Neighborhood size k, Destination dimensionality d.
Output: A set of outliers $\boldsymbol{OT}$ and A set of landmarks $\boldsymbol{LM}$;
A set of short-circuit edges $\boldsymbol{SC}$ and An embedding result.

1. $\boldsymbol{OT}$={Outliers identified by BORDER}, $\boldsymbol{S}=\boldsymbol{S}-\boldsymbol{OT}$;
2. For $\forall\boldsymbol{x}_i\in\boldsymbol{S}$ do
2.1. Calculate $\boldsymbol{U}$ nearest neighbors $KNN_U(\boldsymbol{x}_i)$ of $\boldsymbol{x}_i$($U = d + 1$);
2.2. Obtain the projection error ε_i of $\{\boldsymbol{x}_i\}\cup KNN_U(\boldsymbol{x}_i)$ when project it to $\mathcal{R}^d$;
3. The set of overlapping region $\boldsymbol{OR}$ = ∅;
4. While $\boldsymbol{S}\neq\emptyset$ do
4.1. Find the point $\boldsymbol{x}_m$ with minimum ε_m;
4.2. If $\varepsilon_m>\varepsilon_0$ then break;
4.3. Mark the set $\{\boldsymbol{x}_m\}\cup KNN_U(\boldsymbol{x}_m)$ as a new linear block$\boldsymbol{LP}_j$;
4.4. While projection error ε_j^P of $\boldsymbol{LP}_j<\varepsilon_0$
4.4.1. Find the point $\boldsymbol{x}_n\in\boldsymbol{S}\cup\boldsymbol{OR}-\boldsymbol{LP}_j$ nearest to center $\boldsymbol{\mu}_j$;
4.4.2. Add $\boldsymbol{x}_n$ to $\boldsymbol{LP}_j$ and recalculate ε_j^P ;
4.5. Remove the last added point $\boldsymbol{x}_l$ from $\boldsymbol{LP}_j$ and $\boldsymbol{S}=\boldsymbol{S}-\boldsymbol{LP}_j$;
4.6.$\boldsymbol{OR}=\boldsymbol{OR}\cup\{\boldsymbol{x}\in\boldsymbol{LP}_j | M_j(\boldsymbol{x}) > (1-\alpha)\max_{\boldsymbol{y}\in\boldsymbol{LP}_j} M_j(\boldsymbol{y})\}$;
5. $\boldsymbol{OT}=\boldsymbol{OT}\cup\boldsymbol{S}$, $\boldsymbol{LM}=\bigcup_j\{$The point in $\boldsymbol{LP}_j$ nearest to $\boldsymbol{\mu}_j\}$;
6. For $\forall\boldsymbol{x}_i\in\bigcup_j\boldsymbol{LP}_j$do
6.1. Calculate k_c nearest centers of $\boldsymbol{x}_i$, denoted as $KNC_{k_c}(\boldsymbol{x}_i)$;
7. $\boldsymbol{SC}$ = ∅;
8. For $\forall\boldsymbol{x}_r,\boldsymbol{x}_q\in\boldsymbol{LP}_j$ do
8.1. If $KNC_{k_c}(\boldsymbol{x}_r)\cap KNC_{k_c}(\boldsymbol{x}_q)==\emptyset$;
8.1.1. $\boldsymbol{SC}=\boldsymbol{SC}\cup\{Edge(\boldsymbol{x}_r,\boldsymbol{x}_q)\}$;
9. Construct the neighborhood graph $\boldsymbol{G}$ with neighborhood size k;
10. $\boldsymbol{G}=\boldsymbol{G}-\boldsymbol{SC}$, then compute shortest path graph between $\boldsymbol{LM}$ and $\bigcup_j\boldsymbol{LP}_j$;
11. Make a d-dimensional embedding by means of MDS.

Based on above analysis, we summarize the proposed LPC-Isomap algorithm in Table 1. Furthermore, an incremental PCA algorithm can be utilized to reduce the computational complexity of recalculating ε_j^P in step 4.4.2 of LPC-Isomap[17]. When the dimensionality D of data set is very high and the data set size N is insufficient, we only calculate the first p largest

eigenvalues and the corresponding orthonormal eigenvectors, where $d<p\ll\min(N,D)$. Eqn (1) is transformed into:

$$\varepsilon = \sum_{i=d+1}^{p} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i \qquad (3)$$

The Mahalanobis distance $M(x)$ of $\boldsymbol{LP}$ is reformulated as:

$$M(\boldsymbol{x}) = \sqrt{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 / \sigma + \sum_{i=1}^{p} \left(1/\lambda_i - 1/\sigma\right)\left(\boldsymbol{u}_i^T(\boldsymbol{x}-\boldsymbol{\mu})\right)^2} \qquad (4)$$

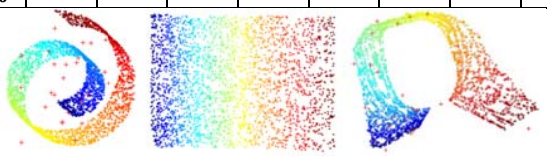
where $\sigma \in \mathcal{R}^+$, and $\lambda_i (1 \le i \le p)$ is the positive eigenvalues of the covariance matrix satisfying $\lambda_1 \ge \ldots \ge \lambda_p \ge \sigma > 0$,with corresponding orthonormal eigenvectors $\boldsymbol{u}_i$ $(1 \le i \le p)$.

IV. Numerical Experiments

The first experiment is parameter setting of threshold $\varepsilon_0$of projection error and overlapping factorα. $\varepsilon_0$determines the approximate scope of center quantity where as α just trims the quantity. Table 2 shows the ability to identify outliers. Different values of k_c determine different thresholds of short-circuit edges. We can record the change process of residual variance of the embedding result by increasing k_c(neighborhood size is large enough), and a marked rise of residual variance indicates the appearance of short-circuit edges.

TABLE II. Ability of LPC-Isomap to Identify Outliers in Noisy Swiss-Rolldata (%)

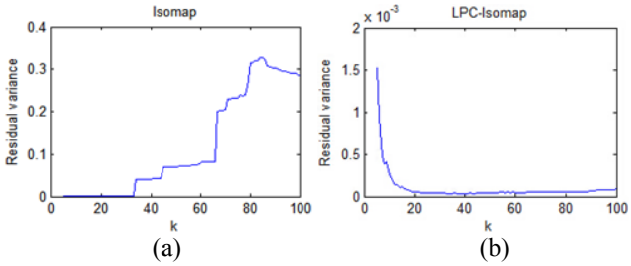
Size of data set	BORDER		ε_0 = 0.03 α = 0.5		ε_0 = 0.003 α = 0.5		ε_0 = 0.003 α = 0.8	
noise	FRR	FAR	FRR	FAR	FRR	FAR	FRR	FAR
1000 + 10	50	0.35	93.8	0.54	32.7	3.1	18	0.92
3000 + 10	28	0.24	100	0.20	46	0.80	47.7	0.15
3000 + 30	32	0.22	98.7	0.22	31	0.93	42.6	0.16



(a)(b)(c)

FIGURE II. (A) NOISY SWISS-ROLL DATA. (B) RESULT OF LPC-ISOMAP.(C) RESULT OF ISOMAP

Then we verify the behavior of LPC-Isomap based on noisy Swiss-Roll data. The size of the data set is 3030, including 30 noisy points (see fig.2a). The parameters are set to ε_0 = 0.003, α= 0.7, and k_c = 5,the neighborhood size k of k-NN is set to 10. As shown in fig. 2b, most outliers are eliminated and the intrinsic manifold structure of sample data set is worked out by LPC-Isomap, whereas Isomap fails to learn the intrinsic structure in fig.2c.To verify effectiveness of deleting the short-circuit edges, we increase the neighborhood size k from 5 to 100 based on the Swiss-Roll data of 3000 points without noise (seefig.3). Table 3 shows the comparison on computational complexity between LPC-Isomap and Isomap based on Swiss-Roll data.



(a) (b)

FIGURE III. THE CHANGE OF RESIDUAL VARIANCE WITH NEIGHBORHOOD SIZE *K* BEING INCREASED FROM 5 TO 100

TABLE III. Running Time of Isomap and LPC-Isomap in Swiss-Roll Data (s)

Data set size	1000	1500	2000	2500	3000	3500
Isomap	2.611	5.525	9.285	13.887	20.330	30.131
LPC-Isomap	1.187	1.970	2.884	4.104	5.458	7.012

V. Conclusion

We have utilized BORDER algorithm and linear patches to eliminate the outliers from noisy data set and select landmark points. Moreover, most of the short-circuit edges have been removed from the neighborhood graph by analyzing intersection of k-nearest center (k-NC) between any two points. These three contributions led to the LPC-Isomap. Experiment results based on Swiss-Roll data and noisy sculpture face data illustrate that our new algorithm achieves better precision and speed than that of Isomap.

Acknowledgements

The work was supported by the Zhejiang Provincial Natural Science Foundation of China (No. LR12F03002) and the National Natural Science Foundation of China (No. 61473252).

References

[1] J. B. Tenenbaum, V. de Silva, and J. C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science, vol. 290, pp. 2319–2323, 2000.

[2] S. T. Roweis and L. K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science, vol. 290, pp. 2323–2326, 2000.

[3] R. Wang, S. Shan, X. Chen, J. Chen, and W. Gao, *Maximal linear embedding for dimensionality reduction*, Pattern Analysis and Machine Intelligence, vol. 33, no. 9, pp. 1776–1792, Sept 2011.

[4] M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum, V. de Silva, and J. C. Langford, *The Isomap algorithm and topological stability*, Science, vol. 295, no. 5500, p. 7a, January 2002.

[5] Z. G. Tang, J. Yang, and B. R. Yang, *A new outlier detection algorithm based on manifold learning*, in Control and Decision Conference (CCDC), 2010 Chinese, May 2010, pp. 452–457.

[6] P. Feng, L. Li, H. Zhang, and G. Qian, *Outliers detection in ICA*, in Communications, Circuits and Systems (ICCCAS), 2013 International Conference on, vol. 2, Nov 2013, pp. 328–330.

[7] B. Liu, Y. Xiao, P. Yu, Z. Hao, and L. Cao, *An efficient approach for outlier detection with imperfect data labels*, Knowledge and Data Engineering, IEEE Transactions on, vol. 26, no. 7, pp. 1602–1616, July 2014.

[8] C. Y. Xia, W. Hsu, M. L. Lee, and B. C. Oo, *BORDER: Efficient computation of boundary points*, IEEE Trans. Knowl. Data Eng., vol. 18,

no. 3, pp. 289–303, 2006.

[9] J. B. Tenenbaum and V. de Silva, *Global versus local methods in nonlinear dimensionality reduction*, Advances in Neural Information Processing Systems, vol. 15, pp. 721–728, 2003.

[10] Y. K. Lei, Y. M. Xu, *Fast Isomap based on minimum set coverage*, Lecture Notes in Artificial Intelligence, vol. 6216, pp. 173–179, 2010.

[11] L. Feng, C. K. Gao, T. Sun*, and H. Wu, *A neighborhood selection algorithm for manifold learning*, International Conference On Computer Design And Appliations, vol. 2, pp. 339–342, 2010.

[12] Q. S. Jiang, Y. P. Lu, and Z. K. Hong, *A neighborhood parameter optimization method of lle based on topology preservation*, in Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on, vol. 8, August 2011, pp. 4231–4234.

[13] Y. B. Zhan, J. P. Yin, and X. W. Liu, *Adaptive neighborhood selection based on local linearity for nonlinear dimensionality reduction*, Lecture Notes in Computer Science Volume, vol. 5821, pp. 337–348, 2009.

[14] T. Wang, T. Xia, and X. M. Hu, *Minimax distance metric-based neighborhood selection algorithm for Isomap*, in Education Technology and Computer, 2010 2nd International Conference on, vol. 3, 2010, pp. 342–346.

[15] V. Premachandran and R. Kakarala, *Consensus of k-nns for robust neighborhood selection on graph-based manifolds*, in Computer Vision and Pattern Recognition, 2013 IEEE Conference on, 2013, pp.1594–1601.

[16] C. M. Bishop, Ed., *Pattern recognition and machine learning*. New York: Springer, 2006.

[17] D. Huang, Z. Yi, and X. R. Pu, *A new incremental PCA algorithm with application to visual learning and recognition*, Neural Process Letters, vol. 30, pp. 171–185, 2009.