Manifold Learning with Local Geometry Preserving and Global Affine Transformation

Dong Huang and Xiaorong Pu
Computational Intelligence Laboratory
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, P. R. China.
E-mail: {donnyhuang, puxiaor}@uestc.edu.cn

***Abstract*—This paper proposes a new approach to learn the low dimensional manifold from high dimensional data space. The proposed approach deals with two problems in the previous algorithms. The first problem is local manifold distortion caused by the cost averaging of the global cost optimization during the manifold learning. The second problem results from the unit variance constraint generally used in those spectral embedding methods where global metric information is lost. The formulation of the proposed method is described in details. Experiments on both low dimensional data and real image data are performed to illustrate the theory.**

I. Introduction

Many high-dimensional data in real-world applications can be modelled as data points lying close to a low-dimensional nonlinear manifold. Discovering the structure of the manifold from a set of data points sampled from the manifold with noise is very challenging in the unsupervised learning. The key difficulty is that the data points are unorganized, i.e., no adjacency relationship among them are known beforehand. Otherwise, the learning problem becomes the well-researched nonlinear regression problem. Traditional dimension reduction techniques such as Principal Component Analysis (PCA) and Factor Analysis usually work well when the data points lie close to a linear (affine) subspace in the input space, while tend to fail to detect nonlinear structures of the data set.

Recently, many algorithms have been developed to deal with the nonlinear low-dimensional manifolds in high-dimensional noisy data samples. These algorithms vary in both motivations and final results. For example, nonlinear dimension reduction and manifold learning, which can respectively be traced back to Self-Organizing Maps (SOM) [1] and Principal Curves/Surfaces [2]. To ease the following discussion, we call the original high dimensional data space the input space, and the underling low dimensional space the feature space.

Nonlinear dimension reduction aims to project/embed the high dimensional data to their low (usually as low as possible) dimensional counterparts while preserves the certain geometric properties among neighboring data points. The geometric properties include the pairwise geodesic distances (ISOMAP [3] and MDS [4]), the local convexity (e.g. LLE [5][6]), local distances (e.g. MVU [7]) and angles between nearby points (e.g. Conformal Eigenmaps [8]). Many algorithms are formulated as convex optimization problems and provide ways to estimate the manifold dimension. However, it is generally assumed in these models that the low dimensional manifold is isometric to a convex subset of Euclidean space. These models may have difficulties with high curvature of the manifold and Out-of-sample extension for non-isometric manifold. For example, image sequence of the same 3-D object rotating through 360 degrees. On the other hand, the Local Manifold Learning tries to parameterize the underlying low-dimensional manifold based on locally linear approximation. In LTSA [9], Non-Local Manifold Tangent Learning [10] and LSML [11], the nonlinear (possibly non-Isometric) manifold of the input data set are modelled as a set of overlapping local tangent planes. The Local Manifold Learning is especially suitable for recovering the structure of the manifold in sparsely populated regions and beyond the support of the provided data. These methods provide no ways to estimate manifold dimension (except LSTA) and compute the explicit embedding. Besides, these models prone to local minima since their optimization involve Alternating Least Squares solution for more than one separated variables.

Our approach draws inspiration from and improves upon the pioneering work. Instead of considering the dimension reduction and manifold learning in isolation, the proposed method construct a nonlinear mapping that avoiding the local minima in optimization problems. The nonlinear mapping is realized by modelling the Local Geometry and a Global Affine transformation in the input space. The proposed method, referred to as LGGA, features on two aspects. One of them is avoiding the local manifold distortion caused by the cost averaging of the global cost optimization during the manifold learning. The other is recovering the global metric information lost in those spectral embedding methods using the unit variance constraints. Moreover, the proposed method can estimate the underlying dimension and is robust to the number of neighbors.

The rest of this paper is organized as follows. In section II, the proposed method are formulated in two steps: modelling the local manifold and recovering the global metric information. Section III presents the simulation results and discussions. Finally, the paper is concluded in Section IV.

978–1–4244–1674–5/08/$25.00 © 2008 IEEE

II. Formulation of the LGGA Manifold Learning

Assume that a smooth m-dimensional manifold F lying in the d-dimensional input data space ($d \gg m$). The input data set $X = \{x_j \in R^d, j = 1, \cdots, N\}$ is sampled possibly with noise from the manifold. We seek to learn the nonlinear function F that transform a point on the manifold into its neighboring points on the manifold, capturing all the modes of variation of the data. Manifold learning is conducted by constructing a global coordinate system $\tau \in R^m$, i.e. the feature space.

Denote $F(x, \varepsilon)$ the transformation of x, with $\varepsilon \in R^m$ acting on the degrees of freedom of the transformation according to the formula $M : F(x, \varepsilon) = M(\tau + \varepsilon)$, where $\tau = M^{-1}(x)$. Taking the first order approximation of in the neighborhood of τ given small enough ε: $F(x, \varepsilon) \approx x + H(x)\varepsilon$, where each column of the matrix $H(x)$ is the partial derivative of M w.r.t. $\tau_k : H_k(x) = \frac{\partial}{\partial \tau} M(\tau)$. Thus our goal can be restated as learning a function $H(\cdot) : R^d \to R^{d \times m}$ and ε_{tj} for each data points x_t with a neighbor x_j sampled from the manifold. In this case, $F(x_t, \varepsilon_{tj}) \approx x_j$ in the least square sense, or equivalently speaking: $H(x_t)\varepsilon_{tj} \approx x_j - x_t$.

Following this idea we can estimate an embedding of the unknown lower dimensional feature vectors τ_js from the x_js preserving the nonlinear transformations in the input space. Noting the manifold learning results in dimension reduction and denoising of the data points.

A. Modelling the Local Manifold

Let $X^{(t)} = [x_1^{(t)}, \cdots, x_k^{(t)}] \in R^{d \times k}$ be the k neighbors of the data point $x_t \in X$ $(t = 1, \cdots, N)$. One nature way to learn the nonlinearity is to construct local unit around each data point and minimize the global cost functions. For example, LLE is based on linear combination, LSTA fits the local unit by the tangent plane. However, since the global cost optimization tends to average the cost among local units, both methods yield incorrect overlapping in the areas of large deformation or low sampling rate. In the proposed method, the relative scale of each local unit is removed from the global optimization.

We begin modelling the manifold variations in the local region. The local unit in the neighborhood of x_t is modelled as

$$\overline{X^{(t)}} = (X^{(t)} - x_t) \approx U^{(t)} \Sigma^{(t)} (V^{(t)})^T,$$

where $U^{(t)} \in R^{d \times m}$ and $V^{(t)} \in R^{k \times m}$ are composed of eigenvectors, and $\Sigma^{(t)} \in R^{m \times m}$ is the diagonal matrix of singular values. The underlying assumption is that the numbers of neighbors is no less than the dimension of the underlying manifold, i.e. $k > m$. Then the local eigen-space at x_t can be approximate by the local subspace spanned by the eigenvectors of the centered matrix. It follows that

$$U^{(t)} \Sigma^{(t)} (V_{j\cdot}^{(t)})^T = H(x_t)\varepsilon_{tj} \approx x_j^{(t)} - x_t, (j = 1, \cdots, k)$$

or equivalently

$$U^{(t)} \Sigma^{(t)} (V^{(t)})^T \approx (X^{(t)} - x_t),$$

where $V_{j\cdot}^{(t)}$ is the jth row of $V^{(t)}$. Thus for the tth data point x_t, the error function in the input space is

$$E_t^{(I)} = \|U^{(t)} \Sigma^{(t)} (V^{(t)})^T - (X^{(t)} - x_t)\|^2. \qquad (1)$$

We then try to find the global coordinates $\tau_t \in R^m, (t = 1, \cdots, N)$ in the feature space can be constructed using the local information learned in the input space. Minimizing the reconstruction error in the feature space:

$$E_t^{(F)} = \|W^{(t)} \Sigma^{(t)} (V^{(t)})^T - (T^{(t)} - \tau_t)\|_F^2, \qquad (2)$$

where $T = [\tau_1, \cdots, \tau_N] \in R^{m \times N}$. Let the matrix $S^{(t)}$ and S_t be the 0-1 selection matrix such that $TS^{(t)} = T^{(t)}$ and $TS_t = \tau_t$. The optimal alignment matrix $W^{(t)} \in R^{m \times m}$ that minimizes the reconstruction error $E_t^{(F)}$ can be computed as

$$W^{(t)} = T(S^{(t)} - S_t) \left(\Sigma^{(t)} (V^{(t)})^T \right)^+. \qquad (3)$$

Thus, the overall reconstruction error $E^{(F)}$ is given by

$$\begin{aligned} E^{(F)} &= \sum_{t=1}^{N} E_t^{(F)} \\ &= \sum_{t=1}^{N} \|W^{(t)} \Sigma^{(t)} (V^{(t)})^T - (T^{(t)} - \tau_t)\|_F^2 \\ &= \sum_{t=1}^{N} \left\| T(S^{(t)} - S_t)\Theta_t \right\|_F^2 \\ &= \|TS\Theta\|_F^2, \end{aligned}$$

where

$$S = \left[(S^{(1)} - S_1), (S^{(2)} - S_2), \cdots, (S^{(N)} - S_N) \right],$$

$$\begin{aligned} \Theta_t &= \left(I - \left(\Sigma^{(t)} (V^{(t)})^T \right)^+ \left(\Sigma^{(t)} (V^{(t)})^T \right) \right) \\ &= \left(I - V^{(t)} (V^{(t)})^T \right), \end{aligned} \qquad (4)$$

$$\Theta = diag\{\Theta_1, \cdots, \Theta_N\}.$$

Here, the unit variance constrain $TT^T = I$ is imposed on T, where $I \in R^{m \times m}$ is the unitary matrix. It follows that T can be computed as the eigenvectors corresponding to the 2nd to $(m + 1)$th smallest eigenvalues of the matrix

$$B = S\Theta\Theta^T S^T,$$

Note in (4), the relative scale of each local unit is lost. The distortion resulted from the global cost optimization can be reduced for areas of sparse distribution or large curvature.

Also note that $\Sigma^{(t)} (V^{(t)})^T$ in the input space are obtained with respect to an orthonormal basis (1), therefore it seems quite natural to preserve the orthogonality of $W^{(t)}$ in the low-dimensional feature space as well. This idea may lead to additional alternating least square problem [9]. In our method, this is amended using a global affine transformation in the sense of least square solution.

B. Recovering The Global Metric Information

In the spectral embedding methods, i.e. LTSA, LLE, the global metric information is totally lost due to the unit variance constrain. We want to find a global affine transformation $z_t = L\tau_t$, $L \in R^{k\times k}$, such that x_t and z_t have similar local Gram matrix:

$$(X^{(t)} - x_t)^T(X^{(t)} - x_t) \approx (Z^{(t)} - z_t)^T(Z^{(t)} - z_t),$$

where $Z = [z_1, \cdots, z_N] \in R^{k\times N}$. Denote $\overline{Z^{(t)}} = Z^{(t)} - z_t$ and $\overline{T^{(t)}} = T^{(t)} - \tau_t$. It follows that

$$\overline{Z^{(t)}}^T\overline{Z^{(t)}} = \overline{T^{(t)}}^T L^T L\overline{T^{(t)}}$$

Therefore, the cost function to be minimized is given by

$$\begin{aligned} & E_t^{(F)} \\ = \ & \left\|(X^{(t)} - x_t)^T(X^{(t)} - x_t) - \overline{Z^{(t)}}^T\overline{Z^{(t)}}\right\|_F^2 \\ = \ & \left\|(X^{(t)} - x_t)^T(X^{(t)} - x_t) - \overline{T^{(t)}}^T L^T L\overline{T^{(t)}}\right\|_F^2. \end{aligned}$$

Let $P = L^T L$, then P is a positive semi-definite matrix, i.e. $P \succeq 0$. The above problem is then formulated as

$$\begin{aligned} \min \quad & \sum_{t=1}^{N}\left\|(X^{(t)} - x_t)^T(X^{(t)} - x_t) - \overline{T^{(t)}}^T P\overline{T^{(t)}}\right\|_F^2 \\ s.t. \quad & \quad P \succeq 0 \end{aligned}$$

Let

$$Y_t\Big|_{ij} = \left\{\frac{1}{2}[(\tau_i - \tau_t)(\tau_j - \tau_t)^T + (\tau_j - \tau_t)(\tau_i - \tau_t)^T]\right\},$$

and

$$g_t\Big|_{ij} = (x_i - x_t)^T(x_j - x_t),$$

where $Y_t\Big|_{ij} \in R^{m\times m}$ and $g_t\Big|_{ij}$ is a scaler, we have

$$\begin{aligned} & \sum_{t=1}^{N} E_t^{(F)} \\ = \ & \sum_{t=1}^{N}\left\|(X^{(t)} - x_t)^T(X^{(t)} - x_t) - \overline{T^{(t)}}^T P\overline{T^{(t)}}\right\|_F^2 \\ = \ & \sum_{t=1}^{N}\sum_{i,j}\left[vec(P)^T vec\left(Y_t\Big|_{ij}\right) - g_t\Big|_{ij}\right]^2 \\ = \ & vec(P)^T\sum_{t=1}^{N}\sum_{i,j} vec\left(Y_t\Big|_{ij}\right) vec\left(Y_t\Big|_{ij}\right)^T vec(P) \\ & -2vec(P)^T\sum_{t=1}^{N}\sum_{i,j}\left(g_t\Big|_{ij}\right)\cdot vec\left(Y_t\Big|_{ij}\right) \\ & +\sum_{t=1}^{N}\sum_{i,j}\left(g_t\Big|_{ij}\right)^2 \end{aligned}$$

It follows that, the above problem can be transformed into

$$\begin{aligned} Min \quad & \|A\cdot vec(P) - B\|^2 \\ s.t. \quad & \quad P \succeq 0, \end{aligned}$$

where

$$A = \left(\sum_{t=1}^{N}\sum_{i,j} vec\left(Y_t\Big|_{ij}\right) vec\left(Y_t\Big|_{ij}\right)^T\right)^{1/2} \in R^{m^2\times m^2}$$

$$B = (A^T)^+\sum_{t=1}^{N}\sum_{i,j}\left(g_t\Big|_{ij}\right)\cdot vec\left(Y_t\Big|_{ij}\right) \in R^{m^2}.$$

Using Schur complement, the problem above can be solved as a Semi-Definite Programming (SDP) problem

$$\begin{aligned} Min \quad & h \\ s.t. \quad & \begin{pmatrix} I & (A\cdot vec(P) - B) \\ (A\cdot vec(P) - B)^T & h \end{pmatrix} \succeq 0, \\ & P \succeq 0. \end{aligned}$$

The left-hand side of the first constrain depends on the vector $vec(P)$. It can be expressed as

$$F(vec(P)) = F_0 + vec(P)_1\cdot F_1 + \cdots + vec(P)_{m^2}\cdot F_{m^2} \succeq 0$$

where

$$F_0 = \begin{pmatrix} I & -B \\ -B^T & h \end{pmatrix},\ F_i = \begin{pmatrix} 0 & A_i \\ A_i^T & 0 \end{pmatrix}, i = 1, \cdots, m^2$$

This problem can be easily solved using YALMIP [12] and CSDP [13]. Finally the global affine transformation matrix is obtained by $L = P^{1/2}$. The affine transformation $z_t = L\tau_t$ recovers both the absolute and relative scales in each direction of the underling manifold.

To this end, the nonlinear mapping we construct for manifold learning can be explicitly summarized as follows. Approximating data point x_j by the local unit it belongs to (at x_t), we have $\left(U^{(t)}\right)^T(x_j^{(t)} - x_t) \approx W^{(t)}(\tau_j^{(t)} - \tau_t)$. Then the nonlinear mapping from the input space to the feature space can be represented as:

$$M^{-1}(x) : LW^{(t)}\left(U^{(t)}\right)^T(x_j^{(t)} - x_t) \rightarrow z_j^{(t)} - z_t. \qquad (5)$$

Finally, we call the proposed manifold learning method as LGGA considering its Local Geometry preserving and Global Affine transformation.

One advantage of LGGA is that we can potentially detect the intrinsic dimension of the underlying manifold by analyzing the local subspace structure. In particular, we can examine the distribution of singular values of the centered data matrix $\overline{X^{(t)}}(t = 1, \cdots, N)$ for the neighborhood of each data point x_t. Let the vector $\sigma^{(t)} \in R^k$ contains the singular values of the centered matrix $\overline{X^{(t)}}$. Define the summation $\rho = \sum_{t=1}^{N}\left(\sigma^{(t)}\right)^2$. In Fig.5 it is clearly that the underlying dimension of the feature space can be easily inferred from ρ.

III. Results and Discussion

We tested the proposed framework on four synthetic data sets (Swiss roll, Swiss hole, punctured sphere and twin peaks) and real image data set (the Columbia Object Image Library (COIL-20) [14]). All experiments are conducted by running uncompiled Matlab codes on a $2.8GHz$ Pentium IV PC.

A. *synthetic data sets*

To evaluate the performance of the proposed algorithm, several competing algorithms, i.e. linear PCA, ISOMAP, LLE and LTSA are compared on the four sets of synthetic data. The objective is to map each data set, originally in 3D space, onto a 2D plane. These synthetic data provide a standard benchmark to evaluate the embedding performance, because both input and output data are low-dimensional and thus can be easily visualized. We set the neighbor number $k = 8$ throughout the experiments. In the following, we compare the results of each data set (Fig. 1-4) in detail:

1. Punctured sphere data in Fig. 1. This data set is sampled from a punctured sphere, rather than a complete sphere. The reason is that a sphere is not homeomorphic to a 2D patch. To embed a sphere onto a 2D space, the sphere must be segmented into multiple patches or be punctured. The Linear PCA and ISOMAP produce incorrect mixed point clouds. LTSA produces satisfactory results in the central area, but the boundary area shrinks and overlaps. The proposed algorithm LGGA yields good result, even in the boundary area that undergoes large deformation. In addition, the anisotropic distribution is preserved.

2. Twin peaks data in Fig. 2. Most algorithms produce satisfactory results, except that Linear PCA and diffusion maps yield incorrect mixed point clouds.

3. Swiss roll data in Fig. 3. LGGA produces an ideal output, showing that both local geodesic distances and global scale are preserved almost perfectly. Linear PCA fails since linear projection methods cannot unfold curved structures. ISOMAP attempts to preserve all shortest path distances, but its output is far from satisfactory. LLE yield irregular 2D embedding. In contrast, LTSA yield more faithful results than ISOMAP does. However, as the outputs of LTSA are compressed into a square region, both scale information and aspect ratio are lost.

4. Swiss hole data in Fig. 4. LGGA also works perfectly by preserving the geometry around the hole. Linear PCA produces an incorrect mixed point cloud. ISOMAP and LLE yield distorted shapes around the hole. LTSA can maintain the shape around the hole that is close to LGGA while loses the relative scale information.

B. *Image data sets*

To evaluate the proposed LGGA algorithm on real world data, the Columbia Object Image Library (COIL-20) are used to perform dimensionality reduction. This database contains 128×128 gray-scale images of 20 objects. For each object, 72 views at the intervals of 5 degrees are sequentially obtained by rotating through 360 degrees. The images is then clipped out from the black background using a rectangular bounding box. The bounding box is resized (zoomed) to 128×128 using interpolation decimation filters to minimize aliasing. These settings result in very complex features in the underling manifold. See Fig. 5 and Fig. 6 for the results on the 72 "Duck" images. According to the dimensional estimation method in the previous section, the original high-dimensional ($128 \times 128 = 16384$) data set can be embedded into the 2D space (See Fig. 5). Here we still set the neighbor number $k = 5, 8$.

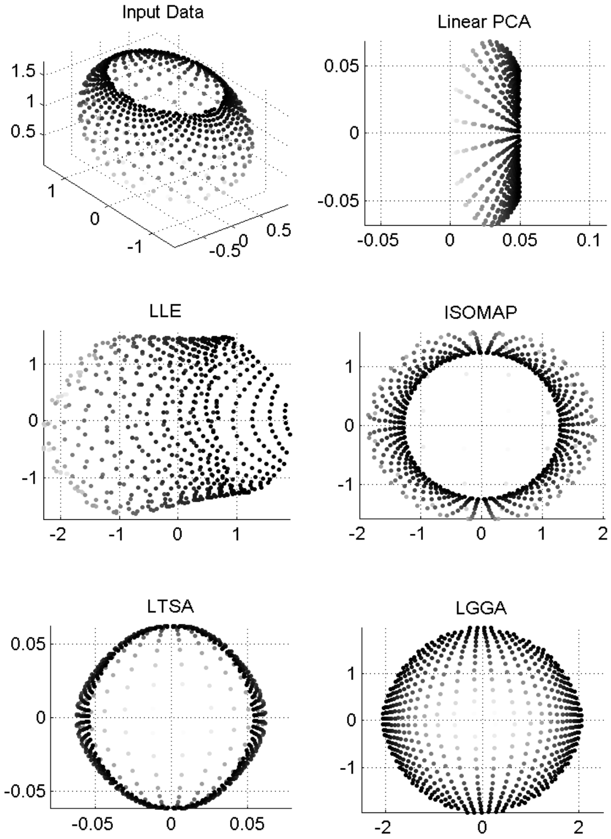

Fig. 1. Comparison on the Punctured Sphere data set.

In the 2D representation (Fig. 6), data point distribution clearly reveals the image variations of both zooming and view angles. This can be easily visualized by the image sequence (A)-(D) correspond to the circled points on the chain (A)-(D) in the 2D embedding. According to the directions notion in Fig. 6, the 2-D points on the chain (B) and (D) are around the 0 and 180 degree in view angles, and are nearly symmetrically distributed with respect to the axis along view angle 90 or 270 degrees. Meanwhile, because of the beak of the duck, the drastic variation of images on (C) are clearly distinguished from that of (A). This difference is also revealed by the 2-D embedding. In addition, due to the extra strong zooming effects of the fixed bounding box, the embedded points between the chains (A)-(D) tend to be more spare than those points on (A)-(D). In sum, the proposed methods produce embeddings that are easy to understand and interpret.

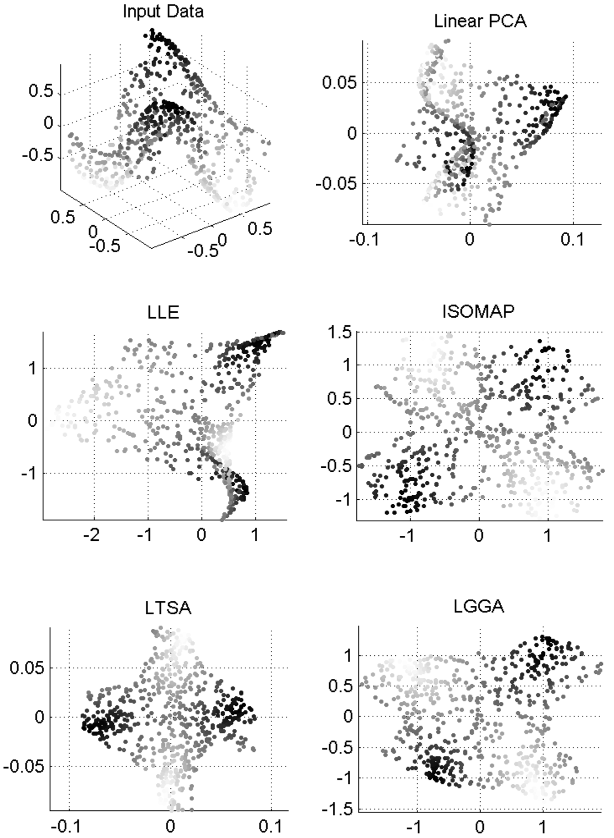

Fig. 2. Comparison on the Twin Peaks data set.

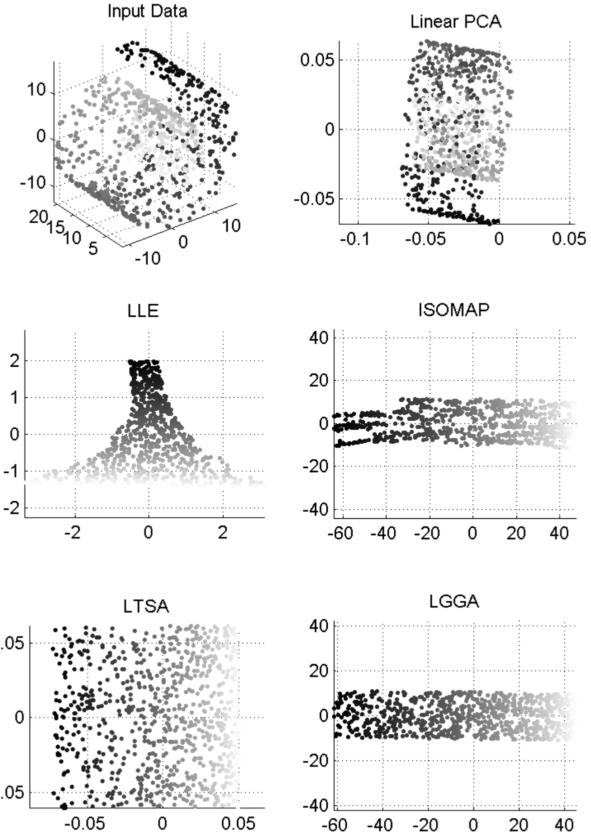

Fig. 3. Comparison on the Swiss Roll data set.

IV. Conclusion

This paper proposes a new manifold learning algorithm, i.e. LGGA. The proposed method deals with both local manifold distortion and the global metric information lost problem generally existed in those spectral embedding methods. Experiments on synthesis data sets and real images show that our model give better performance than closely related models (Linear PCA, ISOMAP, LLE and LSTA) on these problems. In addition, it is not our intention to convince the reader that the proposed algorithm offers an optimal solution to any dimensionality reduction problem. In fact, all existing algorithms are derived from different motivations, and have their own strength and weakness. This leads to interesting directions of our future works, which include automatic parameter estimation, out-of-sample extension and applications to classification problems.

Acknowledgment

This work was supported by Chinese 863 High-Tech Program under Grant 2007AA01Z321.

References

[1] T. Kohonen. *Self-organizing Maps*, 3rd Edition . Springer-Verlag 2000.

[2] T. Hastie and W. Stuetzle. Principal curves, *Journal of The American Statistical Association*. 84: 502-516, 1988.

[3] J.B. Tenenbaum, V. de Silva and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction, *Science*. 290: 2319-2323, 2000.

[4] T. Cox and M. Cox. Multidimensional Scaling, *Chapman Hall*. London, 1994.

[5] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding, *Science*. 290: 2323-2326, 2000.

[6] L. K. Saul and S. T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds, *Journal of Machine Learning Research*. 4:119-155, 2003.

[7] K. Weinberger and L. Saul. Unsupervised Learning of Image Manifolds by Semidefinite Programming, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 2:988-995, 2004.

[8] F. Sha and L.K. Saul. Analysis and Extension of Spectral Methods for Nonlinear Dimensionality Reduction, *Proc. Int'l Conf. Machine Learning*. 785-792, 2005.

[9] Z. Zhang and H. Zha. Principal Manifolds and Nonlinear Dimension Reduction via Tangent Space Alignment, *SIAM Journal of Scientific Computing*. 26 (1): 313-338, 2004.

[10] Y. Bengio, M. Monperrus and H. Larochelle. Nonlocal Estimation of Manifold Structure, *Nueral Computation*. 18(10): 2509-2528, 2006.

[11] P. Dollár, V. Rabaud and S. Belongie. Non-Isometric Manifold Learning: Analysis and an Algorithm, *In Proceedings of International Conference on Machine Learning*. 241-248, 2007.

[12] J. Löfberg. YALMIP: A Toolbox for Modeling and Optimization in MATLAB, *In Proceedings of the CACSD Conference*. Taipei, Taiwan, 2004.

[13] D. B. Borchers. CSDP, a C library for semidefinite programming, *Optimization Methods and Software*. 11(1): 613-623, 1999.

[14] S. A. Nene, S. K. Nayar and H. Murase. Columbia Object Image Library (COIL-20), *Technical Report CUCS-005-96*. February 1996.

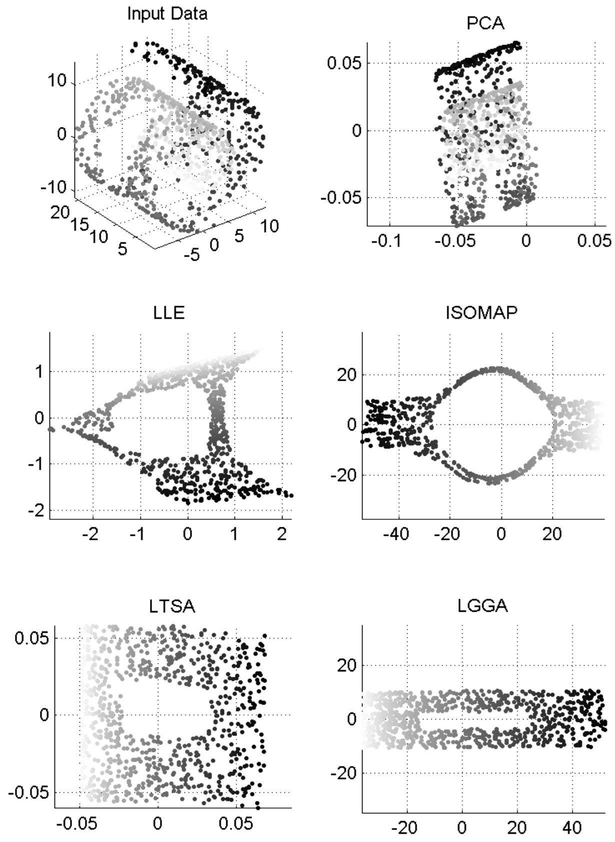

Fig. 4. Comparison on the Swiss Roll with Hole data set.

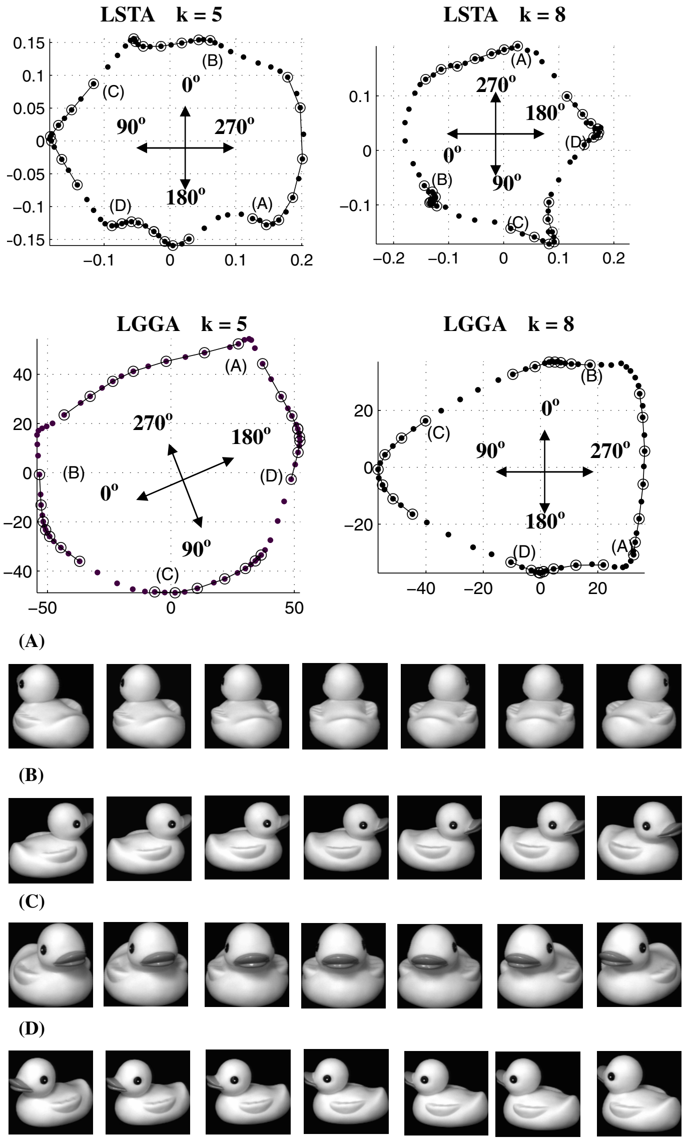

Fig. 6. Image manifold extracted from 72 "Duck" images with neighbor number k=5, 8. The first row shows the results of LSTA. The second row shows the results of LGGA. The image sequence (A)-(D) correspond to the circled points on the chain (A)-(D) in the 2D embedding.

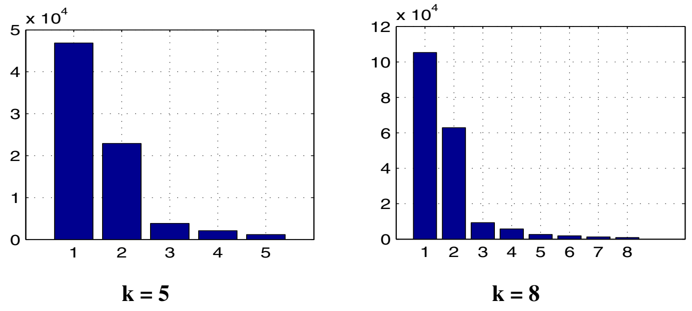

Fig. 5. Dimension estimation of Duck images with neighbor number k=5, 8.