IEICE TRANS. INF. & SYST., VOL.E94–D, NO.8 AUGUST 2011

PAPER

Constraints on the Neighborhood Size in LLE

Zhengming MA[†a)], ***Member***, **Jing CHEN**[†b)], ***and*** **Shuaibin LIAN**[†], ***Nonmembers***

SUMMARY Locally linear embedding (LLE) is a well-known method for nonlinear dimensionality reduction. The mathematical proof and experimental results presented in this paper show that the neighborhood sizes in LLE must be smaller than the dimensions of input data spaces, otherwise LLE would degenerate from a nonlinear method for dimensionality reduction into a linear method for dimensionality reduction. Furthermore, when the neighborhood sizes are larger than the dimensions of input data spaces, the solutions to LLE are not unique. In these cases, the addition of some regularization method is often proposed. The experimental results presented in this paper show that the regularization method is not robust. Too large or too small regularization parameters cannot unwrap S-curve. Although a moderate regularization parameters can unwrap S-curve, the relative distance in the input data will be distorted in unwrapping. Therefore, in order to make LLE play fully its advantage in nonlinear dimensionality reduction and avoid multiple solutions happening, the best way is to make sure that the neighborhood sizes are smaller than the dimensions of input data spaces.

key words: *locally linear embedding, nonlinear dimensionality reduction, manifold learning, principle component analysis*

1. Introduction

Nonlinearity is the main advantage that almost all manifold learning methods [1]–[5] claim to have and Locally Linear Embedding (LLE) [1], [6] is no exception. The mathematical proof and experimental results presented in this paper show that when the neighborhood sizes are larger than the dimensions of input data spaces, LLE will degenerate from a nonlinear method for dimensionality reduction into a linear method for dimensionality reduction. The mathematical relation between input data and output data as well as the experimental results are much similar to those of principal component analysis (PCA). Saul and Rowels, the authors of original LLE papers, have pointed out that the neighborhood sizes cannot be too large, otherwise the locally linear assumption would break down and the solutions to LLE are not unique. However, Saul and Rowels have never pointed out that too large neighborhood size would make LLE degenerate from a nonlinear method into a linear method, neither have other researchers. Wu et al. [7] have discussed the useful results LLE could lead to if LLE had been a linear method, but they did not specify the conditions under which LLE would become linear.

Manuscript received September 21, 2010.
Manuscript revised February 22, 2011.

[†]The authors are with the School of Information Science and Technology, Sun Yat-sun University, Guangzhou 510275, China.
a) E-mail: issmzm@mail.sysu.edu.cn
b) E-mail: flyichen@gmail.com

DOI: 10.1587/transinf.E94.D.1636

It is argued that the choices of neighborhood sizes less than the dimensions of input data spaces are most natural because for most real-world data the neighborhood sizes are always less than the dimensions of input data spaces. However, for most artificial data which are widely used to visualize the effectiveness of manifold learning algorithms, the neighborhood sizes are larger than the dimensions of input data spaces [8]–[10]. Therefore, it is meaningful to make clear the constraints on the selection of neighborhood sizes.

When the neighborhood sizes are larger than the dimensions of input data spaces, the solutions to LLE are not unique. In these cases, people often pay much attention to the uniqueness of solutions, but overlook or do not realize at all the degeneration of LLE from nonlinearity into linearity. Saul and Rowels [6] proposed that some regularization method must be added to get a unique solution. Another algorithm [11] for selecting a regularization parameter has been proposed. However, Wang et al. [12] and Hou et al. [13] have pointed out that the regularization method is not robust as the embedding results are quite sensitive to the regularization parameters. The experimental results presented in this paper show that the too large regularization parameter will make the regularized LLE far away from the genuine LLE, while too small regularization parameter makes the regularized LLE much similar to PCA. In theory, the employment of regularization violates the principle of minimum reconstruction error, a basic idea of LLE.

Therefore, in order to make LLE play fully its advantage in nonlinear dimensionality reduction and avoid multiple solutions happening, the best way is to make sure that the neighborhood sizes are smaller than the dimensions of input data spaces.

The rest of this paper is organized as follows. LLE will be formulated in the next section. In Sect. 3, we will prove that when the neighborhood size is larger than the dimension of input data space, LLE will become a linear method for dimensionality reduction. The experimental results are given in Sect. 4. Finally, some concluding remarks will be given in Sect. 5.

2. Locally Linear Embedding

Given a set of N points $X = \{x_1, x_2, \cdots, x_N\}$ in input data space R^D, the data points are assumed to lie on or near a nonlinear manifold of intrinsic dimensionality d. Let us denote the corresponding set of N points in output data space R^d by $Y = \{y_1, y_2, \cdots, y_N\}$. LLE consists of three steps, and

we formulate them as follows.

2.1 Step I: Neighborhood Selection

For each data point x_i, find its k neighbors $x_{i_1}, \ldots, x_{i_k}$, where $1 \le i_p \le N$, $1 \le p \le k$. The k-nearest neighbors is widely used due to its simplicity and ease of implementation. There are a lot of other ways to find the neighbors [8]–[10].

2.2 Step II: Local Representation

For each data x_i and its k neighbors $x_{i_1}, \ldots, x_{i_k}$, find k reconstruction weights by solving the constrained least square problem of following form:

$$\begin{aligned} &\arg\min_{w_{ij}} \left\| x_i - \sum_{p=1}^{k} w_{i_p i} x_{i_p} \right\|^2 \\ &s.t. \ \sum_{p=1}^{k} w_{i_p i} = 1. \end{aligned} \tag{1}$$

This is what local linearity means. The above problem can be also restated as follows: find an element in the following convex set (not subspace described in) which is closest to x_i:

$$\left\{ \sum_{p=1}^{k} w_{i_p i} x_{i_p} \middle| w_{i_p i} \in R, \ p = 1, \ldots, k; \ \sum_{p=1}^{k} w_{i_p i} = 1 \right\}. \tag{2}$$

Note that the above convex set is not complete and therefore the element closest to x_i can not be unique.

When the weights for each data point have been found, the reconstruction weight matrix $\boldsymbol{\Omega}$ can be constructed as follows:

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1N} \\ \vdots & \ddots & \vdots \\ \omega_{N1} & \cdots & \omega_{NN} \end{bmatrix}, \tag{3}$$

where

$$\begin{aligned} \omega_{ji} = \begin{cases} 1 & j = i \\ 0 & x_j \text{ is not a neighbor of } x_i \\ -w_{ji} & x_j \text{ is a neighbor of } x_i \end{cases}, \\ i, j = 1, \ldots, N. \end{aligned} \tag{4}$$

Then, Eq. (1) becomes:

$$\begin{aligned} &\arg\min_{\boldsymbol{\Omega}} \|X\boldsymbol{\Omega}\|^2 \\ &s.t. \ \mathbf{1}^T \boldsymbol{\Omega} = \mathbf{0}^T, \end{aligned} \tag{5}$$

Note that generally $X\boldsymbol{\Omega} \neq \mathbf{0}$.

2.3 Step III: Global Embedding

Find a matrix Y to minimize the following objective function:

$$\begin{aligned} &\arg\min_{Y} \|Y\boldsymbol{\Omega}\|^2 \\ &s.t. \ Y\mathbf{1} = \mathbf{0}, \ YY^T = I \end{aligned} \tag{6}$$

The column vectors of Y are the output data after dimensionality reduction.

3. Constraints on the Size of Neighborhood in LLE

In this section, we will firstly analyze the LLE in the case where $k > D$ by two steps in a strictly mathematical way and then summarize several insights to get the constraints on k.

3.1 Optimal Local Representation

The objective function in (1) can be rewritten as:

$$\begin{aligned} &\arg\min_{W_i} \|G_i W_i\|^2 \\ &s.t. \ \mathbf{1}^T W_i = 1, \end{aligned} \tag{7}$$

where $G_i = [x_{i_1} - x_i, \ldots, x_{i_k} - x_i]$, and $W_i = [w_{i_1 i}, \ldots, w_{i_k i}]^T$ is the reconstruction weight vector.

Definition 1. If a vector W_i satisfies

$$\|G_i W_i\|^2 = 0, \ s.t. \ \mathbf{1}^T W_i = 1, \tag{8}$$

then the W_i is said to be the optimal local representation of x_i.

Because $rank(G_i G_i^T) = rank(G_i^T G_i) = r$ and $r \le \min\{D, k\}$, G_i can be decomposed by SVD decomposition as

$$G_i = U_i \Sigma_i V_i^T, \tag{9}$$

where U_i and V_i are column-orthogonal matrices of size $D \times r$ and $k \times r$ respectively, the column vectors of them are the orthonormal eigenvectors of $G_i G_i^T$ and $G_i^T G_i$ respectively corresponding to the nonzero eigenvalues, Σ_i is a diagonal matrix of size $r \times r$ whose elements are the square roots of the nonzero eigenvalues of the matrix $G_i G_i^T$ or the matrix $G_i^T G_i$.

If $k > D$, then $r < k$. This means that the matrix $G_i^T G_i$ has $k - r$ orthonormal eigenvectors corresponding to the zero eigenvalue. Let V_{i0} be a matrix of size $k \times (k - r)$ whose column vectors are the orthonormal eigenvectors corresponding to the zero eigenvalues of $G_i^T G_i$, then

$$V_i^T V_{i0} = \mathbf{0}. \tag{10}$$

If let

$$W_i = \frac{V_{i0} A}{\mathbf{1}^T V_{i0} A}, \tag{11}$$

where $A \in R^{k-r}$ and $A \neq \mathbf{0}$, then by (10) we have

$$\|G_i W_i\|^2 = \left\| U_i \Sigma_i V_i^T \frac{V_{i0} A}{\mathbf{1}^T V_{i0} A} \right\|^2 = 0, \ \mathbf{1}^T W_i = 1. \tag{12}$$

It means that (11) gives the optimal local representation of

x_i. Note that A is a degree of freedom.

The above derivation shows that, if $k > D$, no matter how to choose neighbors for each data point, the optimal local representation can be found. Consider all data points, in this case, and we have

$$X\boldsymbol{\Omega} = \mathbf{0},\ \mathbf{1}^T\boldsymbol{\Omega} = \mathbf{0}^T. \tag{13}$$

Also mention that $k > D$ is a sufficient but not necessary condition to obtain the optimal local representations of data points. When k neighbors for a data point are excessively linear dependent, the Eq. (8) will be established.

3.2 Optimal Global Embedding

Definition 2. If a matrix Y satisfies

$$\|Y\boldsymbol{\Omega}\|^2 = \mathbf{0},\ Y\mathbf{1} = \mathbf{0},\ YY^T = I, \tag{14}$$

the Y is said to be the optimal global embedding.

If $rank(XX^T) = rank(X^TX) = \gamma$, by SVD decomposition, X can be expressed as

$$X = U\Sigma V^T. \tag{15}$$

Where U and V are column-orthogonal matrices of size $D \times \gamma$ and $N \times \gamma$ respectively, the column vectors of them are the orthonormal eigenvectors of XX^T and X^TX respectively corresponding to the nonzero eigenvalues, Σ is a diagonal matrix of size $\gamma \times \gamma$ whose elements are the square roots of the nonzero eigenvalues of the matrix XX^T or the matrix X^TX. It is evident that

$$V^TV = I,\ V^T\mathbf{1} = \mathbf{0}. \tag{16}$$

When $k > D$, by (13) and (15) we have

$$U\Sigma V^T\boldsymbol{\Omega} = \mathbf{0}. \tag{17}$$

Then we obtain

$$V^T\boldsymbol{\Omega} = \mathbf{0}. \tag{18}$$

Now let

$$Y = PV^T, \tag{19}$$

where P is a matrix of size $d \times \gamma$ which satisfies

$$PP^T = I \tag{20}$$

and in fact defines a linear mapping from γ-dimensional space to d-dimensional space, then by (16), (18), (19) and (20) we have

$$\|Y\boldsymbol{\Omega}\|^2 = \|PV^T\boldsymbol{\Omega}\|^2 = \mathbf{0},\ Y\mathbf{1} = \mathbf{0},\ YY^T = I. \tag{21}$$

This means that (19) gives the optimal global embedding. Note that P is a degree of freedom.

3.3 Discussions on the Neighborhood Size in LLE

Based on the above derivations, several insights are presented as follows.

By (15) we have

$$V^T = \Sigma^{-1}U^TX. \tag{22}$$

Substituting (22) into (19), we rewrite the optimal global embedding Y as

$$Y = P\Sigma^{-1}U^TX = CX, \tag{23}$$

where $C = P\Sigma^{-1}U^T$. This means that Y can be expressed as the linear combination of X and LLE becomes a method of linear dimensionality reduction if $k > D$. Note that U^TX in (23) is the PCA of X. PCA is a commonly-used linear method for dimensionality reduction.

Some degrees of freedom for the LLE algorithm are provided by the matrix A in (11) and the matrix P in (19) which can be reasonably used to get better embedding results. For example, set

$$A = \arg\min_A(\|W_i\|^2) = \arg\min_A\left(\left\|\frac{V_{i0}A}{\mathbf{1}^TV_{i0}A}\right\|^2\right) \tag{24}$$

to pick the weight vector which minimize the sum of the squared weights.

To some extent, this solution is similar to the using of regularization in [6]. The intention of employing regularization in [6] is originally to obtain unique local reconstruction weights in the unusual case where $k > D$. In this case, regularization is employed and (7) changes to

$$\begin{aligned} &\arg\min_{w_{ij}}(\|G_iW_i\|^2 + \varepsilon\|W_i\|^2), \\ &s.t.\ \mathbf{1}^TW_i = 1. \end{aligned} \tag{25}$$

Set $\varepsilon = \Delta^2/(k \cdot trace(G^TG))$, where $\Delta^2 \ll 1$ is regularization parameter. Not that (7) is what local linearity means.

The employment of regularization prevents the LLE becoming globally linear. At the same time, it violates the locally linear assumption, the basic idea of LLE. In addition, regularization can be problematic for the following reasons. On one hand, too small regularization parameter makes regularization useless (see Fig. 2 in Sect. 4). On the other hand, when the regularization parameter is not small enough, it was shown in [12] that the correct vectors cannot be well approximated by (25). Moreover, when the regularization parameter is relatively high, it produces weight vectors that tend towards the uniform vector $W_i = (1/k, \ldots, 1/k)$. Consequently, the solution for LLE with a large regularization parameter does not reflect a solution based on reconstruction weight vectors. The regularization parameter must be tuned carefully, since LLE can yield completely different embeddings for different values of this parameter [14]. Therefore, it is not practical to utilize the regularized solution to approximate the true solution.

In conclusion, only set the neighborhood size smaller than the dimension of input data space can LLE play its advantage of nonlinearity.

4. Experimental Results

We draw some experimental results in Fig. 1, Fig. 2 and Fig. 3 for illustration. The 1000 points are randomly sam-

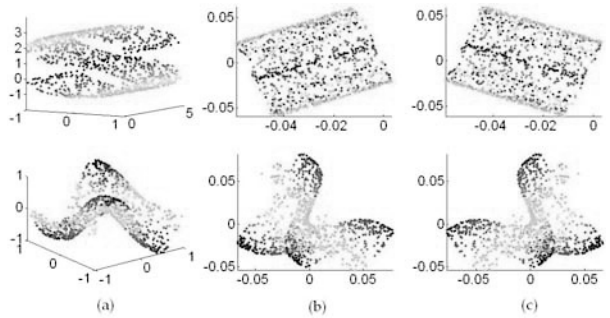

Fig. 1 Dimensionality reduction results by LLE and PCA. (a) Three data sampled from two dimensional manifold S-curve and Twin Peaks respectively; (b) LLE embeddings; (c) PCA embeddings.

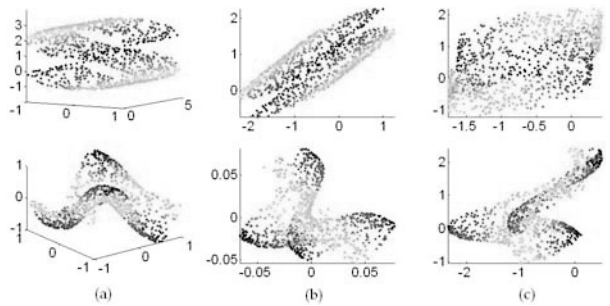

Fig. 2 Dimensionality reduction results by our LLE and LLE using small regularization parameter. (a) Sampled S-curve and Twin Peaks; (b) Our LLE embeddings, A in (11) is set to satisfy (24); (c) LLE embedding with $\Delta^2 = 1e-9$.

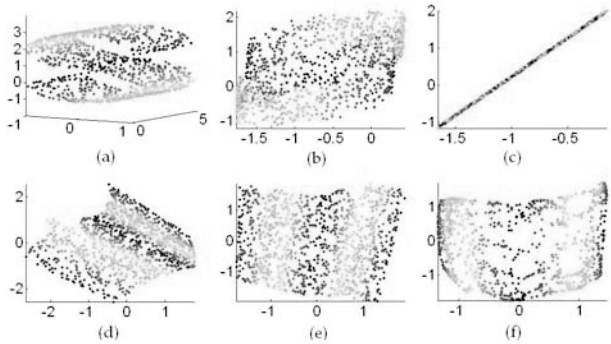

Fig. 3 LLE embeddings with different regularization parameters (a) Sampled S-curve; (b) $\Delta^2 = 1e-9$; (c) $\Delta^2 = 1e-6$; (d) $\Delta^2 = 1e-3$; (e) $\Delta^2 = 1e-2$; (f) $\Delta^2 = 1$.

pled from the three-dimensional S-curve and Twin Peaks. Parameters are set to be $d = 2$, $k = 8$.

Firstly, we give an experiment to illustrate when $k > D$, LLE is not a globally nonlinear but a globally linear method of dimensionality reduction. As evident by Fig. 1, the results by our LLE and PCA are almost similar. As all know, PCA is a linear method for dimensionality reduction. In this experiment, the LLE embeddings are obtained by let $P = [I_{2\times2}, \mathbf{0}_{2\times1}]$ in (19).

Figure 2 illustrates that the LLE embeddings with $\Delta^2 = 1e-9$ are similar to the results of LLE. Here the results of our LLE are obtained by set the matrix A in (11) to satisfy (24). This experimental results show that too small regularization parameter makes regularization useless. We can see from Fig. 1 (b) and Fig. 2 (b) that different LLE embedding results can be obtained if the A or P is set differently.

Figure 3 shows that LLE employing regularization can yield completely different embeddings for different regularization parameters. Too small or too large regularization parameters make the LLE break down. Acceptable result can be obtained only when $\Delta^2 = 1e-2$. Obviously, the so-called acceptable result drawn in Fig. 3 (e) is still different from the ideal embedding. The results of LLE are closely related to the regularization parameter. It implies that the regularized problem is not stable to Δ^2. Therefore, it is not practical to utilize the regularized solution to approximate the true solution.

5. Conclusions

LLE has found wide applications and demonstrated excellent performance [15]–[19] since it was proposed in 2000. In the application to head pose recognition implemented by us, LLE outperformed PCA greatly. The superiority of LLE over PCA has also been reported in many previous literatures. In all these cases, the neighborhood sizes are smaller than the dimensions of input data spaces. If the neighborhood sizes are larger than the dimensions of input data spaces, as shown in this paper, LLE will become a kind of PCA.

When the neighborhood sizes are larger than the dimensions of input data spaces, some regularization method is often added to get a unique solution. However, the regularization method is not robust. S-curve can not been unwrapped if using too large or too small regularization parameters. Even if S-curve can been unwrapped as using a moderate regularization parameter, the relative distance between the input data will be distorted in unwrapping. Therefore, the best way to apply LLE is to make sure the neighborhood sizes smaller than the dimensions of input data spaces.

It should be mentioned that the neighborhood sizes smaller than the dimensions of input data spaces is a necessary but not sufficient condition for LLE becoming a method for globally nonlinearly dimensionality reduction. Saul and Rowels have mentioned that LLE could only be expected to recover embeddings whose dimensionality is strictly less than the neighborhood size, k, because the k neighbors span a space of dimensionality at most $k-1$ and some margin between the intrinsic dimensionality and the neighborhood size is generally necessary to obtain a topology-preserving embedding. Finding the exact relation between the neighborhood size and the faithfulness of the resulting embedding is our future work.

Acknowledgments

In this study, the authors used the Matlab codes published on the homepage of LLE.

References

[1] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol.290, pp.2323–2326, 2000.
[2] J. B Tenenbaum, V. De Silva, and J.C. Langford, "A global geometric framework for nonlinear dimension reduction," Science, vol.290, pp.2319–2323, 2000.
[3] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and Laplacian eigenmaps for dimensionality reduction and data representation," Neural Comput., vol.15, pp.1373–1396, 2003.
[4] D. Donoho and C. Grimes, "Hessian Eigenmaps: New tools for nonlinear dimensionality reduction," Proc. National Academy of Science, vol.100, pp.5591–5596, 2003.
[5] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," SIAM J. Scientific Computing, vol.26, pp.313–338, 2005.
[6] L.K. Saul and S.T. Rowels, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," J. Machine Learning Research, vol.4, pp.119–155, 2004.
[7] F. Wu and Z.Y. Hu, "The LLE and a linear mapping," Pattern Recognit., vol.39, pp.1799–1804, 2006.
[8] Y. Pan, S.S. Ge, and A. Al Mamun, "Weighted locally linear embedding for dimension reduction," Pattern Recognit., vol.42, pp.798–811, 2009.
[9] C. Varini, A. Degenhard, and T.W. Nattkemper, "ISOLLE: LLE with geodesic distance," Neurocomputing, vol.69, pp.1768–1771, 2006.
[10] H. Chang and D. Yeung, "Robust locally linear embedding," Pattern Recognit., vol.39, pp.1053–1065, 2006.
[11] R. Karbauskaite, G. Dzemyda, and V. Marcinkevicius, "Selecting a regularisation parameter in the locally linear embedding algorithm," 20th EURO Mini Conference2008, pp.59–64, 2008.
[12] J. Wang and Z. Zhang, "Nonlinear embedding preserving multiple local-linearities," Pattern Recognit., vol.43, pp.1257–1268, 2010.
[13] C. Hou, J. Wang, Y. Wu, and D. Yi, "Local linear transformation embedding," Neurocomputing, vol.72, pp.2368–2378, 2009.
[14] J.A. Lee and M. Verleysen, Nonlinear Dimensionality Reduction, Springer, Berlin, 2007.
[15] Y. Wang and Y. Wu, "Complete neighborhood preserving embedding for face recognition," Pattern Recognit., vol.43, pp.1008–1015, 2010.
[16] B. Li, C. Zheng, and D. Huang, "Locally linear discriminant embedding: An efficient method for face recognition," Pattern Recognit., vol.41, pp.3813–3821, 2008.
[17] X. Li, S. Lin, S. Yan, and D. Xu, "Discriminant locally linear embedding with high-order tensor data," IEEE Trans. Syst., Man Cybern., B, Cybern., vol.38, pp.342–352, 2008.
[18] Y. Yan and Y. Zhang, "Discriminant projection embedding for face and palmprint recognition," Neurocomputing, vol.71, pp.3534–3543, 2008.
[19] S. Kadoury and M.D. Levine, "Face detection in gray scale images using locally linear embeddings," Computer Vision and Image Understanding, vol.105, pp.1–20, 2007.

Zhenming Ma was born in 1957. He received the B.Sc. and M.Sc. degrees in electronic and communication system from South China University of Technology and Ph.D. degree from Tsinghua University, Chian. He is now a professor at the School of Informaton Science and Technology, Sun Yat-sen University, Guangzhou, China. His research interests include machine learning, pattern recognition and image processing.

Jing Chen received the B.Sc. and M.Sc. degrees in communication and information system from Sun Yat-sen University, Guangzhou, China, in 2001 and 2005, respectively. Currently she is an instructor at Guangdong University of Technology and a Ph.D student of the School of Information Science and Technology at Sun Yat-sen University. Her research interests include manifold learning and pattern recognition.

Shuaibin Lian received Received the B.E degree in Information and Computing Science from Henan University in 2002. Currently he is pursuing the M.S degree in Electronics and Communication Engineering at the School of Information Science and Technology in Sun Yat-sen University. His research interests include machine learning and pattern recognition.