Local linear transformation embedding

Chenping Hou\textsuperscript{a,}\textsuperscript{*}, Jing Wang\textsuperscript{b}, Yi Wu\textsuperscript{a}, Dongyun Yi\textsuperscript{a}

\textsuperscript{a} Department of Mathematics and System Science, National University of Defense Technology, Changsha 410073, China
\textsuperscript{b} School of Information Science and Technology, Huqiao University, Quanzhou 362021, China

1. Introduction

The problem of dimensionality reduction arises in many fields, such as machine learning, pattern recognition, scientific visualization, image processing, neural computation and multi-variable analysis [1]. There are mainly two different research interests, linear methods and nonlinear approaches. Linear methods, such as [2–5], are simple and efficient in many real applications. Nonlinear dimensionality reduction (NLDR), or recently manifold learning, aims at finding the meaningful nonlinear low dimensional structures hidden in the high dimensional data. In many cases of interest, the observed data are assumed to lie on a nonlinear low dimensional manifold. NLDR problem is also known as ‘manifold learning’ in this form. Mathematically, it can be formulated as follows. Given $N$ points $x_1, x_2, \ldots, x_N$ in the high dimensional subspace of $\mathbb{R}^D$, the goal is to find a suitable mapping

$$f: \mathbb{R}^D \rightarrow \mathbb{R}^d, y_i = f(x_i) \quad \text{for} \; i = 1, 2, \ldots, N,$$

where $y_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, N$ and $d$ is the dimensionality of embedding space. Due to the large complexity of approximating a function in high dimensional space, there is no request to express the mapping explicitly in applications. One only needs to find the low dimensional embeddings.

Recently, there have been advances in developing effective and efficient algorithms to perform nonlinear dimension reduction. It includes isometric mapping (Isomap) [6], locally linear embedding (LLE) [7,8] and its variations, manifold charting [9], hessian LLE [10], robust LLE (RLLE) [11], modified LLE (MLLE) [12], and local tangent space alignment (LTSA) [13,14]. All these algorithms cover two common steps: learning a local geometry around each data point and nonlinearly mapping the high dimensional data points into a lower dimensional space using the learned local information [12]. The performances of these algorithms are different both in learning the local information and in constructing the global embedding, though each of them solves an eigenvalue problem eventually. The effectiveness of the retrieved local geometry determines the efficiency of the methods.

Among all of these methods, LLE is the earliest and most widely used one. However, it has some intrinsic shortcomings. (1) If the sample density is low or the points are unevenly sampled, LLE is unavoidable to derive the nonuniform wraps and folds when the data are of low sample density or unevenly sampled. LLE would also fail when the data are contaminated by even small noises. We have analyzed the performance of LLE and pointed out the reason why LLE fails. An improved algorithm, local linear transformation embedding (LLTE), is then proposed. Local linear transformation is performed on nearby points. The ‘Three-stage LLTE’ is also provided when the data has outliers. Comparing with LLE and Local tangent space alignment (LTSA), LLTE could derive more practical embedding than LLE and has wider application prospect than LTSA. Meanwhile, it exploits the tight relations between LLE/LTSA and LTSA. Several experiments and numerical results demonstrate the potential of our algorithm.

* Corresponding author. Tel.: +86 10 627 96 872; fax: +86 10 627 86 911.
E-mail address: hcpnudt@hotmail.com (C. Hou).
dimensional nonlinear manifold. LTSA could derive prominent embeddings on synthetic data. However, for real data, the local tangent space structure is not so obvious. It is difficult and unsuitable to compute local tangent spaces.

This paper will focus on explaining the reason why LLE fails and then improve its performance. Generally, the major curse, which makes LLE failing, is that the local geometry, exploited by the reconstruction weights, cannot be well determined. Since the constrained least squares (LS) problem involved for determining the local weights may be ill-conditioned, especially when the data points are of low sample density or unevenly sampled, or contaminated by noises. Traditional method, such as Tikhonov regularization, may be not a good approximation to the exact solution since it is not stable to the regularization parameter.

The main idea, of this paper, is to improve LLE by constructing a new constrained LS problem instead of the ill-conditioned one. Since the local geometric structure determined by our algorithm is very stable, it could solve the above-mentioned problem of standard LLE and improve its performances. This approach can be considered as computing reconstruction weights approximately in the local tangent spaces. It can be regarded as another kind of alignment of local tangent spaces and has similar properties to LTSA both in measuring the linear dependence of neighborhood and in constructing the (sparse) matrix. It also exploits the tight relationships between LLE and LTSA. We name this approach ‘local linear transformation embedding (LLTE)’. When the data set has outliers in a different way.

2. Performance analysis of LLE

LLE and its variants have been used in many fields, such as face detection [15], denoising [16], feature extraction [17], spectra reconstruction and data visualization because of its simple geometric intuitions, straightforward implementation, and global optimization [8]. We have employed LLE to solve high dimensional corresponding problem [18]. However, it has been reported in [19] that LLE is not stable and may produce distorted embeddings, especially when the points are of low sample density or unevenly sampled, or contaminated by even small noises.

LLE characterizes local geometry structure at each point \( x_i \) by constructing \( x_i \) using its nearest \( k \) neighbors \( N_i = \{x_j \mid j \in J_i \} \). For convenience, we set \( k_i = K \) for \( i = 1, 2, \ldots, N \). The optimal weights are determined by solving the constrained LS problem of following form:

\[
\min_{\mathbf{w}_i} \left\| x_i - \sum_{j \in J_i} \mathbf{w}_{ij} x_j \right\|^2 \quad \text{s.t.} \quad \sum_{j \in J_i} \mathbf{w}_{ij} = 1. \tag{2}
\]

After repeated solving LS problem in Eq. (2) for all \( N \) points, the obtained reconstruction weights form a weight matrix \( \mathbf{W} = [\mathbf{w}_i]_{i=1}^{N} \). LLE attempts to find \( N \) points \( \{y_1, y_2, \ldots, y_N\} \) in lower dimensional space \( \mathbb{R}^d (d < D) \) that can preserve these local combination weights best. That is to minimize

\[
\min_{\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}} \sum_{i} \left\| y_i - \sum_{j \in J_i} \mathbf{w}_{ij} y_j \right\|^2 \quad \text{s.t.} \quad \mathbf{Y}^T \mathbf{Y} = \mathbf{I}. \tag{3}
\]

Actually [8], the low dimensional embeddings can be derived by finding the eigenvectors with the smallest (nonzero) eigenvalues of \( \mathbf{M} = (\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W}) \).

Obviously, the efficiency of LLE tightly depends on these weights. Let us formulate the weight vector \( w_i \) by the local weights \( \mathbf{w}_{ij} \) and denote the matrix \( \mathbf{G}_i = [x_i, y_1 - x_i, \ldots, y_K - x_i] \). Since \( \sum_{j \in J_i} \mathbf{w}_{ij} = 1 \), we can reformulate the combination error as

\[
x_i - \sum_{j \in J_i} \mathbf{w}_{ij} y_j = \mathbf{G}_i \mathbf{w}_i. \quad \text{Hence, Eq. (2) changes to}
\]

\[
\min \left\| \mathbf{G}_i \mathbf{w}_i \right\|^2 \
\text{s.t.} \quad \mathbf{w}_i^T \mathbf{1}_K = 1, \tag{4}
\]

where \( \mathbf{1}_K \) denotes the \( K \)-dimensional column vector of all ones.

Problem in Eq. (4) is a constrained LS problem. Theoretically, there are two different cases:

- If there is a null vector of \( \mathbf{G}_i \), which is not orthogonal to \( \mathbf{1}_K \), i.e., we can find a certain \( \mathbf{w}_i \) that satisfies \( \|\mathbf{G}_i \mathbf{w}_i\|^2 = 0 \) and \( \mathbf{w}_i^T \mathbf{1}_K = 0 \). The solution can be then normalized \( \mathbf{w}_i \rightarrow \mathbf{w}_i/(\mathbf{w}_i^T \mathbf{w}_i) \).
- Otherwise, for any null vector \( \mathbf{w}_i \) of \( \mathbf{G}_i \), i.e., \( \|\mathbf{G}_i \mathbf{w}_i\|^2 = 0 \), we all have \( \mathbf{w}_i^T \mathbf{1}_K = 0 \). As pointed in [8], the problem in Eq. (4) can be solved by adding a Lagrange multiplier \( \lambda \) and minimizing the following equation:

\[
\mathbf{L}(\mathbf{w}_i, \lambda) : = \|\mathbf{G}_i \mathbf{w}_i\|^2 - \lambda (\mathbf{1}_K^T \mathbf{w}_i - 1) = \mathbf{w}_i^T \mathbf{G}_i^T \mathbf{G}_i \mathbf{w}_i - \lambda (\mathbf{1}_K^T \mathbf{w}_i - 1). \tag{5}
\]

By calculating the following derivations:

\[
\frac{\partial \mathbf{L}(\mathbf{w}_i, \lambda)}{\partial \mathbf{w}_i} = \mathbf{G}_i^T \mathbf{G}_i \mathbf{w}_i - \lambda \mathbf{1}_K = 0,
\]

\[
\frac{\partial \mathbf{L}(\mathbf{w}_i, \lambda)}{\partial \lambda} = \mathbf{1}_K^T \mathbf{w}_i - 1 = 0. \tag{6}
\]

The optimal weights are given by \( \mathbf{w}_i = t_\lambda/(\mathbf{1}_K^T t_\lambda) \), where \( t_\lambda \) is the LS solution to \( \mathbf{G}_i^T \mathbf{G}_i \mathbf{t}_\lambda = \mathbf{1}_K \).

In summary, these results, which have been proposed in [12], are listed as follows. Denote by \( t_0 \) the orthogonal projection of \( \mathbf{1}_K \) to the null space of \( \mathbf{G}_i \) and \( t_1 = (\mathbf{G}_i^T \mathbf{G}_i)^{-1} \mathbf{1}_K \). Then the vector

\[
\mathbf{w}_i^* = \frac{t_\lambda}{t_1^T t_\lambda}, \quad t_\lambda = \begin{cases} t_0 & \text{if } t_0 \neq 0, \\ t_1 & \text{if } t_0 = 0. \end{cases} \tag{7}
\]

is an optimal solution to the problem in Eq. (4). Here \( t_\gamma \) represents the Moore–Penrose generalized inverse.

Based on these analysis, we now explain why LLE fails in these two conditions. Please note that, the local intrinsic dimensionality of the original manifold is \( d \). \( \mathbf{G}_i^T \mathbf{G}_i \) only has \( d \) relatively large eigenvalues. The rest are either very small or zeros. When the data are of low sample density or unevenly sampled or \( K > D \), the solution given by Eq. (7) is not stable since \( \mathbf{G}_i^T \mathbf{G}_i \) is singular or nearly singular. It is suggested in [8] to regularize this problem. One should solve the regularized problem replaced.

\[
(\mathbf{G}_i^T \mathbf{G}_i + \gamma \|\mathbf{G}_i\| \mathbf{I}) \mathbf{t}_\gamma = \mathbf{1}_K, \quad \mathbf{w}_i = t_\lambda/(\mathbf{1}_K^T t_\lambda). \tag{8}
\]

Here \( \gamma \) is a small positive number.

However, the solution to this problem is sensitive to \( \gamma \). Let \( t_\gamma(\gamma) \) be the unique solution to the regularized problem. It has been shown in [12] that \( \mathbf{w}_i(\gamma) = t_\gamma(\gamma)/(\mathbf{1}_K^T t_\gamma(\gamma)) \) converges to \( \mathbf{w}_i^* \) as \( \gamma \rightarrow 0 \). Since \( \mathbf{G}_i^T \mathbf{G}_i \) has relatively small nonzero eigenvalues and \( (\mathbf{G}_i^T \mathbf{G}_i)^{-1} \) is largely determined by these small eigenvalues, \( (\mathbf{G}_i^T \mathbf{G}_i)^{-1} \) is not.
stable and, hence, the regularized solution \( w_i(\gamma) \) of Eq. (8) is very sensitive to \( \gamma \). Intuitively, when \( \gamma \) is large, local weights cannot be well approximated by \( w_i(\gamma) \). When \( \gamma \) is comparative to small eigenvalues of \( C_i^T G_i \), a little vibration of \( \gamma \) would make large changes of \( (C_i^T G_i) \).

For example, if \( t_0 \neq 0 \) is small, \( w_i(\gamma) \) trends to \( \mu = t_0/(1_t^T t_1) \) at first and then turns to the limit value \( w_i^c \) eventually. Please note that \( \mu \) and \( w_i^c \) are orthogonal to each other. There must be large gaps among \( w_i(\gamma) \)s of different small \( \gamma > 0 \) and, hence, \( w_i^c \) cannot be well approximated by \( w_i(\gamma) \). Zhang et al. have drawn some numerical results to show this metaphase phenomenon [12].

In summary, since \( C_i^T G_i \) has relatively small nonzero eigenvalues, the solutions to problems in Eqs. (7) and (8) cannot represent the true solution. It is not practical to utilize the regularized solution to approximate the problem in Eq. (7), we can assume that

\[
\Delta_i = A_i^T \Delta A_i = I,
\]

where \( A_i \) has relatively small nonzero eigenvalues. \( \Delta A_i \) is not stable and tends to be very large. Meanwhile, a small \( \Delta A_i \) would cause a large variance between \( (A_i + \Delta A_i)^T A_i \) and \( A_i^T A_i \). It implies that there is a large gap between \( \hat{t} \) and \( t \). In other words, a small noise could make the result unsatisfied. If we assume that the elements of \( e \) are independently normally distributed with the mean zero, the variance \( \sigma^2 \) and meanwhile, the rows of \( \Delta A_i \) are independent, normal random variables with the mean zero and the variance \( \Sigma = t + A_i^T(e - \Delta A_i t) \) would have the variance

\[
E(\text{var}(e)) = (A_i^T A_i)^T \sigma^2 + t^T \Sigma t.
\]

Obviously, based on Eq. (14), the variance of \( \hat{t} \) is very large when the matrix \( A_i^T A_i \) (or equivalently, \( A_i \)) is near singular. This can also partly explain the instability of LLE to noises.

We have also done some experiments to show that LLE is sensitive to noises. The 300 points, shown in Fig. 1(e), are uniformly sampled from a two-dimensional noisy helical line. These noises are independently normally distributed with the mean zero and the variances 0.01 and 0.09, respectively. We draw the expected embeddings in Fig. 1(f). LLE embeddings, with different noise variances, are shown in Fig. 1(g) and (h). The x-axis in (f), (g) and (h) represent the ideal one-dimensional coordinates.

Based on above analysis, we can see that both of the embeddings are far from the ideal one. Even the small perturbation of the original data would cause the failure of LLE.

If the data have noises, i.e., \( \hat{x}_i = x_i + \eta_i \). Here \( \hat{x}_i \) is the clean data and \( \eta_i \) is the noise. We denote \( A_i = C_i^T G_i \), \( \Delta G_i = \Delta \), \( \eta_i = \Delta \), then Eq. (9) changes to

\[
(A_i + \Delta A_i) t + e - \Delta A_i t = 1_r.
\]

Since \( (A_i + \Delta A_i)^T (A_i + \Delta A_i) = I \), it follows that the solution to the problem in Eq. (10) is

\[
\hat{t} = t + A_i^T(e - \Delta A_i t,
\]

where \( F^T \) is the nonlinear part of \( (A_i + \Delta A_i)^T \). From this, we see that

\[
V(\hat{t}) = (A_i + \Delta A_i)(\sigma^2 + t^T \Sigma t).
\]

Obviously, based on Eq. (14), the variance of \( \hat{t} \) is very large when the matrix \( A_i^T A_i \) or equivalently, \( A_i \) is near singular. This can also partly explain the instability of LLE to noises.

We have also done some experiments to show that LLE is sensitive to noises. The 300 points, shown in Fig. 1(e), are uniformly sampled from a two-dimensional noisy helical line. These noises are independently normally distributed with the mean zero and the variances 0.01 and 0.09, respectively. We draw the expected embeddings in Fig. 1(f). LLE embeddings, with different noise variances, are shown in Fig. 1(g) and (h). The x-axis in (f), (g) and (h) represent the ideal one-dimensional coordinates.

Based on above analysis, we can see that both of the embeddings are far from the ideal one. Even the small perturbation of the original data would cause the failure of LLE.

---

**Fig. 1.** Toy examples. (a) Original S-curve; (b) expected embedding of S-curve; (c) LLE embedding of S-curve with regularization parameter \( \gamma = \text{tol} = 1e - 4 \); (d) LLE embedding of S-curve with regularization parameter \( \gamma = \text{tol} = 1e - 5 \); (e) noisy helical line with Sigma = 0.3; (f) expected embedding of noisy helical line; (g) LLE embedding of noisy helical line with Sigma = 0.1; and (h) LLE embedding of noisy helical line with Sigma = 0.3.
Generally, the instability of LLE can be attributed to the bad determination of the local coefficients. Especially, when the constrained LS problem involved for determining the local weights has small eigenvalues. Fortunately, it also implies that the performance of LLE can be improved if this problem can be effectively solved.

3. Our methods

In this section, we will discuss two different cases. When the noise is small, we propose LLTE to improve the performance of LLE. When the data has outliers (large noises), we provide ‘Three-stage LLTE’, which combines advantages of RLLE and LLTE. Since it can be regarded as the extension of LLTE, we will mainly focus on LLTE and omit some details of ‘Three-stage LLTE’.

3.1. Local linear transformation embedding

In this section, we will employ linear local transformations to eliminate the small eigenvalues of $G_i^T G_i$. Therefore, local construction weights can be stably computed and the performance of LLE is improved. First, we make a suitable local transformation to improve its spectral property. Then, the local construction weight matrix $W$ is computed based on these new local points. Finally, low dimensional embeddings can be derived by finding the eigenvectors corresponding to the smallest (nonzero) eigenvalues of $M = (I - W)^T (I - W)$.

Assume that the SVD decomposition of $G_i$ is $G_i = U_i G_i V_i$. An expected local transformation $f_i$ should hold the following characters:

- The new matrix $\tilde{G}_i^T \tilde{G}_i$, formed by the modified data, should not have very small nonzero single values. This would assure that the improved algorithm avoids solving the ill-conditioned LS problem.
- This transformation is able to maintain the original character as much as possible. It would guarantee that LLTE could derive the intrinsic structure of original data.
- This transformation should be easy to realize and the additional computation should not be too large.

Considering these requirements, we choose the local linear transformation. Since local manifolds can be well approximated by the $d$-dimensional patches [8], we can assume that $P_i \in \mathbb{R}^{d \times d}$ is the transformation matrix for points in $\mathbb{R}^d_i = \{ x_j : G_i \in I \} \cup \{ x_i \}$. That is

$$z_i^j = f_i(x_i) = P_i^T x_j, \quad \forall x_j \in \mathbb{R}^d_i. \quad (15)$$

Obviously, $\tilde{G}_i^T[z_{i1}^j, \ldots, z_{id}^j] = P_i^T G_i$. Note that $G_i = U_i G_i V_i$. Hence

$$G_i G_i^T = U_i G_i^T V_i U_i G_i V_i = U_i G_i^T U_i G_i V_i = U_i G_i^T G_i V_i, \quad (16)$$

$$\tilde{G}_i^T \tilde{G}_i = V_i^T I_i U_i G_i^T G_i V_i = V_i^T G_i^T G_i V_i, \quad (17)$$

$$\tilde{G}_i^T \tilde{G}_i = C_i^T P_i P_i^T G_i = V_i^T I_i U_i G_i^T P_i P_i^T G_i V_i. \quad (18)$$

For convenience, we assume that $U_i$ has been arranged according to the eigenvalues of $G_i C_i^T$ (or equivalently, $C_i G_i^T$). The eigenvector, whose eigenvalue is the largest, is in the first column.

Considering the above requirements and the expression of $C_i^T \tilde{G}_i$, we can take the first $d$ columns of $U_i$ as $P_i$. This is because

- It has been pointed that the manifold can be locally approximated by the patches with dimensionality $d$. The first $d$ columns of $U_i$ corresponds to the $d$ directions in which the elements of $G_i$ have the $d$ largest variances. Thus, $P_i^T x_j, \forall x_j \in \mathbb{R}^d_i$ is enough to hold the most significant local structure information of the original data.
- $P_i$ has a significant background. It can be approximately regarded as the basis of local tangent spaces.
- Through this kind of transformation, the instability of the regularized LS problem is reduced.

In fact, if we decompose $U_i = [P_i, Q_i]$, where $Q_i \in \mathbb{R}^{d \times (d - d)}$. Then

$$\tilde{G}_i^T \tilde{G}_i = V_i^T I_i U_i^T P_i P_i^T U_i G_i^T G_i V_i = V_i^T I_i \left[ \begin{array}{cc} P_i^T P_i & P_i^T Q_i \\ Q_i^T P_i & Q_i^T Q_i \end{array} \right] \left[ \begin{array}{c} P_i^T P_i \\ Q_i^T Q_i \end{array} \right] I_i^T G_i V_i. \quad (19)$$

Since any two column vectors of $U_i$ are orthogonal to each other, it follows that

$$\tilde{G}_i^T \tilde{G}_i = V_i^T I_i \left[ \begin{array}{cc} P_i^T P_i & \frac{P_i^T Q_i}{Q_i^T Q_i} \\ \frac{Q_i^T P_i}{Q_i^T Q_i} & Q_i^T Q_i \end{array} \right] \left[ \begin{array}{c} P_i^T P_i \\ \frac{P_i^T Q_i}{Q_i^T Q_i} \end{array} \right] I_i^T G_i V_i$$

Since any two column vectors of $U_i$ are orthogonal to each other, it follows that

$$\tilde{G}_i^T \tilde{G}_i = V_i^T I_i \left[ \begin{array}{cc} P_i^T P_i & P_i^T Q_i \\ Q_i^T P_i & Q_i^T Q_i \end{array} \right] \left[ \begin{array}{c} P_i^T P_i \\ Q_i^T Q_i \end{array} \right] I_i^T G_i V_i \quad (20)$$

$$\tilde{G}_i^T \tilde{G}_i = V_i^T I_i \left[ \begin{array}{cc} P_i^T P_i & \frac{P_i^T Q_i}{Q_i^T Q_i} \\ \frac{Q_i^T P_i}{Q_i^T Q_i} & Q_i^T Q_i \end{array} \right] \left[ \begin{array}{c} P_i^T P_i \\ \frac{P_i^T Q_i}{Q_i^T Q_i} \end{array} \right] I_i^T G_i V_i \quad \text{where} \ A_i = K \times K \text{ diagonal matrix and its elements are arranged according to the eigenvalues of } G_i C_i^T, \text{ we may write}$$

$$\tilde{G}_i^T \tilde{G}_i = V_i^T I_i \left[ \begin{array}{cc} P_i^T P_i & \frac{P_i^T Q_i}{Q_i^T Q_i} \\ \frac{Q_i^T P_i}{Q_i^T Q_i} & Q_i^T Q_i \end{array} \right] \left[ \begin{array}{c} P_i^T P_i \\ \frac{P_i^T Q_i}{Q_i^T Q_i} \end{array} \right] I_i^T G_i V_i = V_i^T A_i V_i, \quad (21)$$

where $A_i$ is a $K \times K$ diagonal matrix and formulated by the largest $r$ nonzero eigenvalues of $G_i C_i^T$ (or $C_i^T G_i$), and $r = \min(K, d)$.

Because it is commonly assumed that $K > d$, $r = \min(K, d)$. In this situation, the regularized solution should be employed to approximate the real solution.

Comparing to the formulations of $G_i C_i^T$ and $\tilde{G}_i C_i^T$, which are shown in Eqs. (17) and (21) respectively, we can conclude that $\tilde{G}_i C_i^T$, which truncates the largest $d$ eigenvalues of $G_i C_i^T$, has better spectral properties than $G_i C_i^T$. We will compare them in two different cases.

- When $K > D > d$. This usually occurs on the synthetic data, such as the S-curve and helical line in Fig. 1. Assume $\sigma_1 > \sigma_2, \ldots, > \sigma_d$ are the $N$ nonzero eigenvalues of $G_i C_i^T$. Since the intrinsic dimensionality is $d$, $\sigma_d \gg \sigma_{d+1}$. The eigenvalues of $G_i C_i^T$ are $\sigma_1, \sigma_2, \ldots, \sigma_d$ and $K - d$ zero elements. Correspondingly, the eigenvalues of $C_i^T G_i$ are $\sigma_1, \sigma_2, \ldots, \sigma_d$ and $K - d$ zeros. Since the intrinsic dimensionality of local patches is $d$, $G_i C_i^T$ has nearly $D - d$ relatively small nonzero eigenvalues. They will lead to the instability of $w_i(\gamma)$. Since the small eigenvalues of $G_i C_i^T$ are eliminated, the $M - P$ inverse of $G_i C_i^T + \gamma \| G_i \|_F I = 0$ is stable to $\gamma$. Let $\tilde{v}_i(\gamma)$ be the unique solution to the following regularized problem:

$$\tilde{G}_i C_i^T + \gamma \| G_i \|_F I = \tilde{v}_i(\gamma). \quad (22)$$

The regularized result $\tilde{w}_i(\gamma) = \tilde{v}_i(\gamma)/(\| \tilde{v}_i(\gamma) \|)$ has well convergence behavior. It could converge to the approximate optimal solution as $\gamma \to 0$. More importantly, there is no metastable phenomenon. This would guarantee that the optimal solution could be well estimated by $\tilde{w}_i(\gamma)$ as $\gamma \to 0$.

- When $D > K > d$. This usually occurs in real applications where dimensionality of the original data is very high. As mentioned in the above case, the eigenvalues of $G_i C_i^T$ are $\sigma_1, \sigma_2, \ldots, \sigma_K$. However, the eigenvalues of $C_i G_i^T$ are $\sigma_1, \sigma_2, \ldots, \sigma_d$ and $K - d$ zeros. Since $G_i C_i^T$ only has $d$ significant eigenvalues, the convergence behavior of $w_i(\gamma)$ is quite...
uncertain. By contrast, $\mathcal{C}_i^2 G_i$ compels these small eigenvalues to zero. Hence, the solution to the regularized problem is also stable to $\gamma$ and the optimal solution can be well approximated.

When the data are contaminated by noises, we have shown that the major reason for failure of LLE is the instability of the inverse of $G_i^2 G_i$. If we employ this transformation and formulate the new matrix $\mathcal{C}_i^2 G_i$, the performance of LLE is also improved. Actually, when the data are contaminated by small noises, the significant eigenvalues of $G_i^2 G_i$ do not change too much. Therefore, the $M$–$P$ inverse of $G_i^2 G_i + \gamma |G_i|_2 I$ is stable and, hence, our approach is not sensitive to small noises.

We summarize LLTE algorithm in Table 1.

Obviously, $\gamma$ should not be too large or too small. When $\gamma$ is too small, it could not play the role of regularization and the problem of $G_i G_i$ becomes ill-posed. Fortunately, when the original data are normalized, the regularized solution is far from the truth. Additionally, the differences in scale will cause the difference in determining the neighbor set of $x_i$, we can approximately regard $x_i \approx \bar{x}_i$. Hence $ar{X}_i \approx G_i$, $Q_i \approx P_i$. Linear mapping in LLTE can be considered as projecting nearby points into an approximated local tangent space. In other words, LLTE can be regarded as a MLE, which computes construction weights in local tangent space. Meanwhile, it can also be considered as a new kind of LTSA.

Generally, LLTE is not only a MLE, but also a new kind of LTSA. It tightly relates LLE and LTSA and provides us a new way to view these two methods.

### 3.2. Three-stage LLTE

Another concern for application is that the noise level in the previous data is not high. When the data contain outliers (large noises), LLTE, as most of other manifold learning methods, would fail. However, RLLE can detect and eliminate outliers. Nevertheless, it cannot deal with this problem, which is of low level or located closely to the original data.

A direct way to solve this problem is to employ LLTE in three stages. (1) Apply the outlier eliminating strategy of RLLE to eliminate outliers and form the ‘clean data’. (2) Perform LLTE on the clean data. (3) Derive embeddings of the outliers based on representations of the clean data. Since RLLE has been introduced in [11], we would like to omit some details of the first step.

Assume $X_C = \{x_1^{(C)}, x_2^{(C)}, \ldots, x_9^{(C)}\}$ are ‘clean’, $X_N = \{x_1^{(N)}, x_2^{(N)}, \ldots, x_9^{(N)}\}$ are detected outliers. Embedding $Y_C = \{y_1^{(C)}, y_2^{(C)}, \ldots, y_9^{(C)}\}$ of $X_C$ is derived by LLTE on $X_C$ directly. Finally, in the similarly way shown in the first step of LLE, we approximate each point in $X_N$ by its nearby points in $X_C$ and formulate an approximation matrix $L$. The embedding $Y_N$ of $X_N$ can be derived by $Y_N = Y_C L$. We would like to omit some details and list each step of ‘Three-stage LLTE’ in Table 2. We have also done an experiment for illustration in Section 4, see [11] for more discussions.

The ‘Three-stage LLTE’ has obvious advantages as follows:

- Since LLTE performs better than RLLE when the data has small noises, ‘Three-stage LLTE’ could not only eliminate and map the outliers in a practical way as RLLE, but also map the clean data in a more practical way than RLLE. Therefore, the total results of ‘Three-stage LLTE’ are more reasonable than that of RLLE. We have done an experiment shown in Fig. 4 for illustration.
- Moreover, ‘Three-stage LLTE’ is not sensitive to the effectiveness of outlier detection since LLTE is robust to small noises. The performance of RLLE, however, heavily depends on the effectiveness of the outlier detection. If points with small noises are not detected, RLLE would fail. Intuitively, one can see the comparisons between RLLE and LLTE shown in the following section.

Generally, LLTE, together with ‘Three-stage LLTE’ is simple and efficient, they are both of high application potentials.
4. Experiments

In this section, we present several examples to show the performances of LTSA algorithm. The data sets include both the simulated data and the real-world examples. We would like first to introduce the evaluation matrix.

4.1. Standard diversity

In order to measure the diversity of two embeddings, we introduce a particularly designed measurement—standard diversity (SD). It mainly focuses on the shape of manifolds and is unchanged to translation, scale change and symmetrical deformation.

Assume that $Y^{(1)} = \{y_1^{(1)}, y_2^{(1)}, \ldots, y_N^{(1)}\}$ denote the expected embeddings and the derived representations are $Y^{(2)} = \{y_1^{(2)}, y_2^{(2)}, \ldots, y_N^{(2)}\}$, where $y_i^{(2)} \in \mathbb{R}^d$ for $i = 1, 2, \ldots, N$ and $j = 1, 2$. We would like to find a linear transformation from $y_j^{(2)}$ to $y_j^{(1)}$, through which $Y^{(1)}$ can be approximated best. In other words, we expect to find $c \in \mathbb{R}^d$ and $T \in \mathbb{R}^{d \times d}$ that minimize

$$\sum_{i=1}^{N} ||y_i^{(1)} - c - Ty_i^{(2)}||^2.$$  \hfill (24)

The optimal solution to this problem is

$$c^* = \frac{1}{N} \sum_{i=1}^{N} y_i^{(1)}, \quad T^* = (Y^{(1)} - c^*1_N)(Y^{(2)})^T. \hfill (25)$$

The SD from $Y^{(2)}$ to $Y^{(1)}$ is defined as the mean of fitting errors, i.e.,

$$\text{SD}(Y^{(2)} \rightarrow Y^{(1)}) = \frac{1}{N} \sum_{i=1}^{N} ||y_i^{(1)} - c^* - Ty_i^{(2)}||^2. \hfill (26)$$

It is obvious that SD is a reasonable measurement of the diversity and consistent with the intuition. We would like to omit some details.

4.2. Experiments on the synthetic data

The first experiment is performed to evaluate LLE, RLLE, MLLE, LTSA and LLTE on the data that are of low density or unevenly sampled. We utilize the S-curve data, which are shown in Fig. 1(a).

There are totally $N = 1500$ randomly sampled points. We run these algorithms with $K = 8$. On the top left of Fig. 2, we draw the expected embeddings.

Since RLLE only tries to detect and remove outliers, its embedding, which is shown in Fig. 2(c), is similar to the results of LLE (Fig. 2(b)). As seen from the bottom row of Fig. 2, MLLE, LTSA and LLTE all perform well.

With different number of points, we also compute SDs between the expected embedding and that derived by LLE, RLLE, MLLE, LTSA and LLTE, respectively. Obviously, with the increase of $N$, the sample density turns to high and the points tend to be evenly sampled. With each fixed $N$, 10 experiments have been done for each method. The mean and standard derivation (within bracket) of these 10 SDs are shown in Table 3. The parameters are the same as previous.

As expected, comparing with LLE and RLLE, MLLE, LTSA and LLTE generate relatively small SDs in all cases. It implies that MLLE, LTSA and LLTE could derive more ‘faithful’ embeddings. Meanwhile, with the increase of $N$, since the sample tends to be with large density and evenly selected, SDs of all methods are decreased. Even with a few number of points, LLTE could achieve satisfied results. This implies that LTTE is of extensive application scopes, especially when the data are difficult to obtain.

The second experiment is employed to compare the performances of these methods on the data with small noises. The 300 points are uniformly sampled from the noisy helical line (shown in Fig. 1(e)). These noises are independently normally distributed with the mean zero and the variance 0.09. We set

$$\text{Mean and Standard derivation of SDs between expected embedding and computed coordinates by LLE, RLLE, MLLE, LTSA and LLTE of the S-curve data with different number of points.}$$

<table>
<thead>
<tr>
<th>$N = 500$</th>
<th>$N = 1000$</th>
<th>$N = 1500$</th>
<th>$N = 2000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLE</td>
<td>3.5269(0.6352)</td>
<td>2.0762(0.4615)</td>
<td>1.3441(0.3205)</td>
</tr>
<tr>
<td>RLLE</td>
<td>1.6634(0.4090)</td>
<td>1.0035(0.2441)</td>
<td>0.9832(0.1015)</td>
</tr>
<tr>
<td>MLLE</td>
<td>0.0332(0.0126)</td>
<td>0.0301(0.0115)</td>
<td>0.0239(0.0091)</td>
</tr>
<tr>
<td>LTSA</td>
<td>0.0319(0.0176)</td>
<td>0.0301(0.0104)</td>
<td>0.0237(0.0091)</td>
</tr>
<tr>
<td>LLTE</td>
<td>0.0337(0.0162)</td>
<td>0.0310(0.0106)</td>
<td>0.0259(0.0096)</td>
</tr>
</tbody>
</table>

Fig. 2. Dimensionality reduction results of the S-curve data. (a) Expected embedding; (b) LLE embedding; (c) RLLE embedding; (d) MLLE embedding; (e) LTSA embedding; and (f) LLTE embedding.
$K = 8$ for all the algorithms. Similarly, expected embeddings and representations derived by these methods are drawn in Fig. 3.

As seen from Fig. 3, it is clear that LLE and RLLE fail since they cannot avoid solving the ill-conditioned problem. MLLE, which uses multiple weights instead of a single coefficient to avoid solving the ill-conditioned LS problem, achieves satisfied results. LTSA could approximately unfold the manifold with small perturbations. Among all of them, LLTE performs best.

Similarly, with different noise levels (with variance 0.01, 0.04 and 0.09, respectively), we compute SDs between the expected embeddings and representations derived by these methods. With each fixed noise standard deviation, 10 experiments have been performed and the mean of 10 results are shown in Table 4.

The forth experiment is to show that LLTE is not sensitive to the regularization parameter $\gamma$. We have set $\gamma = 1 \cdot e^{-5}$ for all the experiments. Actually, since LLTE eliminates the small nonzero eigenvalues of $G_i^T G_i$, it is stable to this parameter. To exclude the influence of noises, we have employed LLE and LLTE on an identical noisy helical line data with various $\gamma$ (ranging from $1 \cdot e^{-6}$ to $1 \cdot e^{-3}$). These computed coordinates are shown in Fig. 5.

As seen from the first row of Fig. 5, LLE is very sensitive to $\gamma$. Comparing with LLE, there are no apparent variances among the embeddings that are derived by LLTE with different $\gamma$. The variance of $\gamma$ would not heavily affect the result of LLTE.

We have also proposed some numerical comparisons. With each fixed $\gamma$, to exclude the influence of noises, we apply LLE and LLTE on the same data. After 10 times of repetition, we compute the mean and standard deviations of 10 SDs for each method. They are shown in Table 5. Other parameters are the same as that in the previous experiments.

Obviously, since the small eigenvalues of $G_i^T G_i$ are eliminated, LLTE has smaller means and Stds in all situations. LLE, however, is sensitive to $\gamma$. Moreover, LLE achieves much larger means and Stds in all cases than LLTE. It implies that LLTE is not only stable to $\gamma$ but also derives more faithful embeddings than LLE.

4.3. Experiments on the Translated Lenna data

In order to compare the performances of LLE, MLLE, LTSA and LLTE on real data sets, we generate the Translated Lenna data set. It consists of many different images of the Lenna. Examples are generated by translating the Lenna image across a larger background of random noises. These noises are independent from one example to another. We constrain the gray level of the noisy background from 50 to 60. The input consists of $N = 900$ grayscale images, with each image containing a $30 \times 30$ Lenna image
superimposed on a 61 × 61 background of noises. The intrinsic dimensionality of this data set is two.

We apply LLE, MLLE, LTSA and LLTE with $K = 6$. Computed coordinates by LLE, MLLE, LTSA and LLTE are shown in Fig. 6(a), (b), (c) and (d) respectively. Four peak points, accompanied with corresponding real images, are marked by circles. Since it is difficult to compute local tangent spaces accurately, LTSA fails. MLLE also fails in this case. LLE and LLTE could reveal the generating structure of these images. Therefore, LLTE could be used in more real cases than LTSA and MLLE.

4.4. Experiments on the USPS data

To compare the performances of LLE, MLLE and LLTE on real noisy dataset, the USPS data [21] is employed. We consider a subset containing $N = 1500$ handwritten digits (‘2’, ‘3’, ‘6’) with first 500 examples in each class. The gray images of handwritten numerals are at 16 × 16 resolutions and converted to $D = 256$ dimensional vectors. We randomly select 70 (about 30%) pixels of each image and change their pix values to 255 (white). They can be regarded as noises permeated on these images.

We have applied LLE, MLLE and LLTE on the original data and the noisy data with $K = 8$, respectively. Two-dimensional embeddings, accompanied with several real images, are shown in Fig. 7.
LLTE achieves better results than LLE and MLLE, especially when the data are contaminated by noises. Since the intrinsic dimensionality of this data is unknown, MLLE also has poor performance in this case. On the contrary, most of the digits are well clustered in the embedding of LLTE. Representations of LLE and RLLE have larger overlaps between different classes. Therefore, LLTE can be used in more areas.

4.5. Experiments on the Breast cancer data

Comparing with LLE, another advantage of LLTE is that it is not sensitive to $K$, the size of neighborhood. In fact, we have set the smallest $K-d$ eigenvalues of $G_t^T G_t$ to be zeros. The major variance has been concentrated to the first $d$ significant eigenvalues. Therefore, the solution to the regularized LS problem in Eq. (22) is stable to $K$. LLE, however, keeps the largest eigenvalues of $G_t^T G_t$ if $D < K$. The solution to $G_t^T G_t t = 1_K$ has close relationship with the choice of $K$.

We employ the 'Wisconsin Diagnostic Breast Cancer' data [22] for illustration. It consists of 569 instances, with 32 attributes of each case. This data set can be divided into two categories: malignant and benign. Excluding the first attribute (ID) and the second attribute (diagnosis), we choose the first 98 instances, which contain 64 malignant cases and 34 benign cases. They are mapped to a two-dimensional subspace by LLE and LLTE with $K = 15$ and 25, respectively. We plot the results in Fig. 8. Red "*" represents the malignant instance and black dot represents the benign one.

As seen from Fig. 8, when $K = 15$, embeddings of LLE and LLTE could represent the class information of the original data, i.e., they are approximately separable. When $K = 25$, however, the computed coordinates of LLE of two classes are heavily overlapped with each other. Fortunately, LLTE can also achieve the satisfied
results. Representations derived by LLTE are also separable in two-dimensional space.

5. Discussions

In this section, we provide some discussions about LLTE, especially about the questions rising from real applications.

First, we can also analyze LLTE in the kernel perspective. Since LLTE is an improved LLE, it has the similar kernel properties to LLE [19,23], except for small change in the formulation of kernels.

Second, the linearization of LLTE is also interesting. LLE has been modified to its linear version, neighborhood preserving embedding (NPE) [24], for face recognition. LLTE can also be linearized through the same way.

Finally, there are also some shortcomings of LLTE. (1) It assumes that the intrinsic dimensionality is known in prior, this is not real in applications. (2) The parameters are determined experimentally. There is no rule or criterion. (3) It also lacks theoretical justification.

6. Conclusions and future works

This paper proposes a new manifold learning approach, LLTE. It can be considered as not only an improved LLE, but also a new kind of local tangent space alignment. Its extension, i.e., the ‘Three-stage LLTE’, is also provided when the data has outliers. LLTE integrates the advantages of both LLE and LTSA. It can derive more faithful embeddings than LLE and has more extensively application scope than LTSA. Experiments on both the synthetic and the real data sets have shown its potential.

Further work includes theoretical justification of LLTE. We will consider more real-world applications, within and beyond areas in computer vision, image processing, and image recognition. We will still research the linearization of LLTE for face recognition. Making a rule to determine parameters is also our research interests.

Acknowledgments

Thanks to two anonymous reviewers for their constructive suggestions. We would like to thank the creator of USPS datasets providing us with useful data. Thanks to creators of ‘Wisconsin Diagnostic Breast Cancer’ database. Thanks to the National Basic Research Program of China under Grant no. 2005CB321800, National Natural Science Foundation of China under Grant no. 60673090 and Graduate School of National University of Defense Technology for their support (B070201).

References

Chenping Hou received his BS degree in applied mathematics from the National University of Defense Technology, Changsha, China, in 2000. He is now a PhD candidate in the Department of Mathematics and System Science. His current research interests include dimensionality reduction, manifold learning, and computer vision.

Jing Wang was born in Fujian, China, in January 1981. He received a PhD in Department of Mathematics, Zhejiang University in 2006. He is now the lecturer of School of Information Science and Technology, Huazhong University. His research interests include manifold learning, data mining, and matrix computation.

Yi Wu is a professor in the Department of Mathematics and System Science at the National University of Defense Technology in Changsha, China. He earned bachelor's and master's degrees in applied mathematics at the National University of Defense Technology in 1981 and 1988. He worked as a visiting researcher at New York State University in 1999. His research interests include applied mathematics, statistics, and data processing.

Dongyun Yi is a professor in the Department of Mathematics and System Science at the National University of Defense Technology in Changsha, China. He has been working as a visiting researcher of the University of Warwick.