Pattern Recognition 43 (2010) 1257–1268

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.de/locate/pr

Nonlinear embedding preserving multiple local-linearities

Jing Wang [a,b,*,1], Zhenyue Zhang [a,2,*]

[a] Department of Mathematics, Zhejiang University, Yuquan Campus, Hangzhou 310027, PR China
[b] School of Computer Science and Technology, Huaqiao University, Quanzhou 362021, PR China

ARTICLE INFO

Article history:
Received 12 December 2008
Received in revised form
4 September 2009
Accepted 8 September 2009

Keywords:
Manifold learning
Dimensionality reduction
Weight vector
Stability of algorithm

ABSTRACT

Locally linear embedding (LLE) is one of the effective and efficient algorithms for nonlinear dimensionality reduction. This paper discusses the stability of LLE, focusing on the optimal weights for extracting local linearity behind the considered manifold. It is proven that there are multiple sets of weights that are approximately optimal and can be used to improve the stability of LLE. A new algorithm using multiple weights is then proposed, together with techniques for constructing multiple weights. This algorithm is called as nonlinear embedding preserving multiple local-linearities (NEML). NEML improves the preservation of local linearity and is more stable than LLE. A short analysis for NEML is also given for isometric manifolds. NEML is compared with the local tangent space alignment (LTSA) in methodology since both of them adopt multiple local constraints. Numerical examples are given to show the improvement and efficiency of NEML.

1. Introduction

The task of nonlinear dimensionality reduction is to find meaningful low-dimensional structures hidden in high dimensional data. Recently, there have been advances in developing effective and efficient algorithms to perform such reduction. These algorithms include isometric mapping (Isomap) [16], locally linear embedding (LLE) [12] and its variations, Laplacian Eigenmaps (LE) [2], manifold charting [3], Hessian LLE [4], weighted LLE [11], and local tangent space alignment (LTSA) [19]. The following two strategies are shared in these algorithms: (1) exploiting the local geometry around each data point; (2) map the manifold nonlinearly to a lower dimensional space based upon the learned local geometric information [6]. Of course, these algorithms are different in the performance of local information extracting and global embedding. For example, ISOMAP estimates the pairwise geodesic distance based on Euclidean distances between neighbors and then maps the high-dimensional points into a lower dimensional Euclidean space by preserving the geodesic distances. LLE extracts the linearity by representing each point as a linear combination of its neighbors and then determines a low-dimensional embedding that preserves the locally linear combination structures. LE computes a nonlinear projection by penalizing the squared Euclidean distances of the projected points in a low dimensional space according to preset pairwise weights of neighbors. In LTSA, the local linear structures are retrieved by projecting neighbors of each point into its tangent space and the required low-dimensional embedding follows by aligning the local tangent coordinates globally.

Unquestionably, extraction of local linear structures plays a key role in nonlinear dimensionality reduction. The more interesting issue is to explore how the retrieved local structures affect the low-dimensional embedding. Recently, an eigenstructure analysis is presented in [19] to explore the behaviors and performance of local manifold learning algorithms, focusing on LTSA. In this paper, we continue the work and focus our interests on LLE.

In the literature, LLE has many applications such as image classification, image recognition, spectra reconstruction and data visualization because of its simple geometric intuitions, straightforward implementation, and global optimization [13,17,9]. However, it is also mentioned that LLE may not be stable and may produce distorted embedding if the manifold has dimension larger than one. One of the causes that make LLE fail is that the local geometry exploited by the reconstruction weights is not well-determined. This can be perceived through the ill-conditionness of the constrained least squares (LS) problem involved for determining the local weights for ideal data with small noise, or through the large combination errors for data with large noise. Tikhonov regularization is generally used for the ill conditioned LS problem. However, the regularized solution may not be a good approximation to the exact solution if the regularization

* Corresponding author at: Department of Mathematics, Zhejiang University, Yuquan Campus, Hangzhou 310027, PR China.
E-mail address: zyzhang@zju.edu.cn (Z. Zhang).

[1] The work of this author was supported in part by NSFC for Youth (Project 10901062).
[2] The work of this author was supported in part by NSFC (Project 10771194) and National Basic Research Program of China (973 Program) 2009CB320804.

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.09.014

parameter is not suitably selected. Meanwhile, a large combination error implies unacceptable linearity extracting.

The purpose of this paper is to stabilize LLE by making use of multiple local weight vectors. Our key observation is that only a set of linear reconstruction weights may not sufficiently determine the whole local linearity. The lack of full constraints on the local linearity may result in instability in numerical embedding. We will give a detailed analysis to clarify this statement. Practically, we will show the existence of linearly independent weight vectors that are approximately optimal. The stability of LLE can be significantly improved by imposing multiple local constraints on the projected coordinates via the multiple approximate optimal weight vectors. Based upon the analysis and the techniques of constructing multiple weights, we will propose a new algorithm using multiple weights. This new algorithm is called as nonlinear embedding preserving multiple local-linearities (NEML). A short analysis for an isometric manifold is also given to show that NEML can stably retrieve the ideal isometric embedding approximately. NEML is closely related to LTSA both on extracting the linearity and on constructing alignment matrix whose eigenvectors form the intended lower dimensional embedding. We will show their similarities. Numerical examples given in this paper show the improvement and efficiency of NEML.

The rest of the paper is organized as follows. In Section 2, we illustrate the instability of LLE resulted from the uncertain local weights and discuss some properties of the weight vector. In Section 3, we show the existence of linearly independent weight vectors that are approximately optimal. Several schemes are given for constructing the multiple weights. The new algorithm NEML will be proposed in Section 4. We give a short analysis of NEML for isometric manifolds in Section 5. In Section 6, we compare NEML with LTSA, especially in linear dependence of neighbors and alignment matrices. Numerical examples will be given in Section 7.

2. Stabilities of the local reconstruction weights

We start with a brief description of LLE [12] for the self-containness of this paper. Let $\{x_1,\dots,x_N\}$ be a given data set of N points in $\mathcal{R}^m$, sampled from a d-dimensional manifold ($d \ll m$) with noise. For a point x_i, we assume that its neighbor set $\mathcal{N}_i=\{x_j, j\in J_i\}$ is well selected so that its linear property dominates the local structure of the manifold. In general, the number $k_i=|J_i|$ of the neighbors is larger than d. LLE learns the local linear structure of the neighbor set $\mathcal{N}_i$ by representing x_i using a linear combination of its neighbors,

$$x_i=\sum_{j\in J_i} w_{ji}x_j+\eta_i$$

with small reconstruction error η_i. The optimal reconstruction weights w_{ij} are determined by solving the optimization problem under the normalization constraint $\sum_{j\in J_i} w_{ji}=1$,

$$\min_{\{w_{ji}, j\in J_i\}}\left\|x_i-\sum_{j\in J_i} w_{ji}x_j\right\| \quad \text{s.t.} \sum_{j\in J_i} w_{ji}=1. \tag{2.1}$$

Once all the reconstruction weights $\{w_{ji}, j\in J_i\}$, $i=1,\dots,N$, are computed, LLE maps the set $\{x_1,\dots,x_N\}$ to $\{t_1,\dots,t_N\}$ in a lower dimensional space $\mathcal{R}^d$ by solving the following optimization problem:

$$\min_{T=[t_1,\dots,t_N]}\sum_i\left\|t_i-\sum_{j\in J_i} w_{ji}t_j\right\|^2 \quad \text{s.t. } TT^T=I, \tag{2.2}$$

which tries to preserve the local reconstruction properties totally as much as possible. Since the objection function in (2.2) can be rewritten in the trace of $T(I-W)(I-W)^TT^T$ with $W=(w_{ji})_{N\times N}$, the optimal solution is given by the eigenvectors of the symmetric positive semi-definite matrix $(I-W)(I-W)^T$ corresponding to the second to $(d+1)$ st smallest eigenvalues. Note that W is a sparse matrix since $w_{ji}=0$ if $j\notin J_i$ for all i.

The stability of LLE heavily depends on the construction of the local weights. As pointed out in [19], it also depends on the gap between the $(d+1)$ st and $(d+2)$ nd eigenvalues of $(I-W)(I-W)^T$ in increasing order if data noise can be ignored. It is to be expected that the resulting embeddings can be quite different if different local weights are used, or if the eigen-space of $(I-W)(I-W)^T$ corresponding to its $d+1$ smallest eigenvalues is blurred by other eigenvectors because of large errors. We do not intend to touch on the latter issue since it needs a complicate analysis as that given in [19] for LTSA. We only focus our analysis on the dependence of the stability on local weights.

Our key observation is that if a manifold has dimension $d>1$, a single set of reconstruction weights may not able to determine the whole local linearity. The lack of control on the local linearity may result in instability in numerical embedding. To see the reason, let us first recall the construction of the optimal weights.

For simplicity, we write the local weights as a vector $w_i=(\dots,w_{ji},\dots)^T_{j\in J_i}$ of length $k_i=|J_i|$. Let $G_i=[\dots,x_j-x_i,\dots]_{j\in J_i}$ and $\mathbf{1}_{k_i}$ be the k_i- dimensional column vector of all 1's. The optimal w_i in LLE is determined by solving the constrained LS problem

$$\min_{w_i}\|G_iw_i\| \quad \text{s.t. } w_i^T\mathbf{1}_{k_i}=1. \tag{2.3}$$

If G_i is of full column-rank numerically, the unique optimal weight vector is given by $w_i^*=y_i/\mathbf{1}_k^Ty_i$, where y_i a solution to the linear system

$$G_i^TG_iy_i=\mathbf{1}_{k_i}. \tag{2.4}$$

Otherwise, it is suggested in [13] to solve the regularized linear system with a regularization constant $\gamma\ll 1$,

$$(G_i^TG_i+\gamma\|G_i\|_F^2I)y_i(\gamma)=\mathbf{1}_{k_i}, \quad w_i(\gamma)=y_i(\gamma)/\mathbf{1}_{k_i}^T. \tag{2.5}$$

Note that $w_i(\gamma)$ depends on the parameter γ.

However, the nonsingular matrix $G_i^TG_i$ is a perturbation of a singular matrix with an error matrix due to the approximate linearity of the neighborhood. The neighbors $x_j\in\mathcal{N}_i$ of x_i and x_i itself have the approximate linear representation

$$x_j=c_i+U_it_j+\varepsilon_j, \quad j\in J_i \text{ or } j=i, \tag{2.6}$$

where c_i is the center of the neighborhood of x_i, U_i is an orthonormal matrix of d columns, t_j is the d-dimensional vector of local coordinates of x_j, and ε_j denotes the error, mainly resulted by data noise, which is orthogonal to U_i. Thus, writing $F_i=[\dots,t_j-t_i,\dots]_{j\in J_i}$ and $E_i=[\dots,\varepsilon_j-\varepsilon_i,\dots]_{j\in J_i}$, we see that

$$G_i=U_iF_i+E_i, \quad G_i^TG_i=F_i^TF_i+E_i^TE_i.$$

Thus, (2.3) or (2.5) can be viewed as a regularized version of the singular system

$$\min_{w_i}\|F_iw_i\| \quad \text{s.t. } w_i^T\mathbf{1}_{k_i}=1 \tag{2.7}$$

using the regularization form

$$(F_i^TF_i+\Delta_i)y_i=\mathbf{1}_{k_i} \tag{2.8}$$

with $\Delta_i=E_i^TE_i$ in (2.3) or $\Delta_i=\gamma(\|F_i\|^2+\|E_i\|^2)I+E_i^TE_i$ in (2.5). Notice that the matrix Δ_i here does not provide any information of the true manifold behind the data. Since the $k_i\times k_i$ matrix $F_i^TF_i$ is of rank at most d and is singular since $k_i>d$ generally, F_i has a null vector. For an optimal weight vector w_i, it is easy to construct an approximately optimal weight vector $\hat{w}_i$ such that $\|\hat{w}_i-w_i\|=1$,

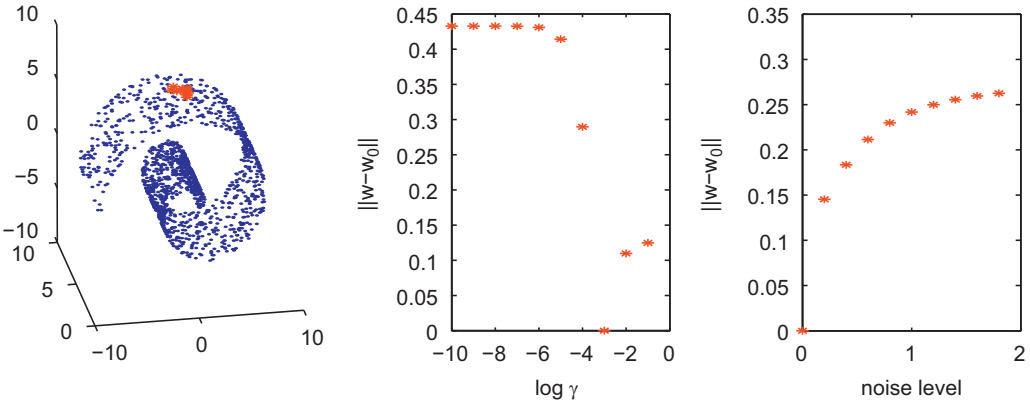

Fig. 1. A fixed neighbor set of the swiss-roll data (left) and the error between $w_0 = w(10^{-3})$ and $w(\gamma)$ depending on γ (middle panel) and noise levels (right).

but $\|G_i\hat{w}_i\| \le \|G_i w_i\| + \|E_i\|$. Such a $\hat{w}_i$ exists in the form $\hat{w}_i = w_i + z$ with a null vector z of $[\begin{smallmatrix} F_i \\ \mathbf{1}_k^T \end{smallmatrix}]$ satisfying $\|z\| = 1$, provided $k_i > d+1$. Thus the resulting weight vector w_i is obvious unstable in the sense: different noise level or different parameter may result in quite different weights. Fig. 1 shows the phenomena for a fixed neighbor set of swiss-roll data with different noise level and parameters. The original data points are generated as follows (in MATLAB notation):

$t = (3 * \pi/2) * (1 + 2 * \text{rand}(1, N)),$
$s = 21 * \text{rand}(1, N),$
$X_{true} = [t. * \cos(t); s; t. * \sin(t)],$

with $N = 1500$. On the left of Fig. 1, we plot the original data set and the fixed neighbor set with $k = 10$. To show the sensitivity of weights to the parameter γ, we compute the distance between $w_0 = w(10^{-3})$ and $w(\gamma)$ with γ varying from 10^{-10} to 10^{-1}. As shown in the middle of Fig. 1, the weights are sensitive to γ when γ is not very small. Then different noise levels are added to the original data set with

$X = X_{true} + noiselevel * (1 - 2 * \text{rand}(3, N)).$

The regularized local weights $w_i(\gamma)$ of the fixed neighbor set under different noise levels are computed with $\gamma = 10^{-3}$. On the right of Fig. 1, we compute the distance between w_0 and those computed local weights under different noise levels. As can been seen, the local weights are very sensitive to noise even if the noise level is not large. Note that different sets of weights for a neighbor set may result in variant embeddings, although they minimize the LLE cost function approximately.

The uncertainty of optimal weights can be further discerned from the behavior of the regularized weights of (2.8) with the simple regularization form $\Delta = \gamma I$.[3] To do this, let us first formulate the optimal weight vector without regularization.

Theorem 2.1. *Let F be a given matrix of k column vectors, $\mathcal{N}(F)$ the null space of F. Then w is an optimal solution to $\min_{\mathbf{1}_k^T w = 1} \|Fw\|$ if and only if*

(1) *w is a null vector of F satisfying $w^T\mathbf{1}_k = 1$ when $\mathbf{1}_k$ is not orthogonal to $\mathcal{N}(F)$, or*
(2) *$w = z + w^*$ for a null vector z of F when $\mathbf{1}_k \perp \mathcal{N}(F)$, where $w^* = (1/\mathbf{1}_k^T y_1) y_1$ and $y_1 = (F^TF)^\dagger \mathbf{1}_k$.*

Proof. See Appendix.

[3] For a positive semidefinite perturbation Δ, $F^TF + \Delta$ can be rewritten in the form $\hat{F}^T\hat{F} + \Lambda$ with a positive semidefinite diagonal matrix and $\|\Lambda\| = \|\Delta\|$. The analysis for general Δ is similar as for the spacial $\Delta = \gamma I$ but complicated a bit.

Let $w(\gamma)$ be the regularized optimal weight vector corresponding to $\Delta = \gamma I$, that is, $w(\gamma) = y(\gamma)/(\mathbf{1}_k^T y(\gamma))$ with $y(\gamma)$ solving $(F^TF + \gamma I)y = \mathbf{1}_k$. We consider the behavior of $w(\gamma)$ as $\gamma \to 0$. Let

$$F = [U, U_0] \begin{bmatrix} \Sigma & \\ & 0 \end{bmatrix} [V, V_0]^T = U\Sigma V^T \tag{2.9}$$

be the singular value decomposition (SVD) of F, where Σ is the diagonal matrix of nonzero singular values of F, and $[U, U_0]$ and $[V, V_0]$ are two orthonormal matrices. Obviously, $\mathcal{N}(F)$ is the column space of V_0. Then

$$y(\gamma) = (F^TF + \gamma I)^{-1}\mathbf{1}_k = V(\Sigma^2 + \gamma I)^{-1}V^T\mathbf{1}_k + \gamma^{-1}V_0V_0^T\mathbf{1}_k.$$

If γ is much smaller than the smallest diagonal of Σ^2,

$$y(\gamma) \approx V\Sigma^{-2}V^T\mathbf{1}_k + \gamma^{-1}V_0V_0^T\mathbf{1}_k = y_1 + \gamma^{-1}y_0,$$

where $y_1 = V\Sigma^{-2}V^T\mathbf{1}_k$ and $y_0 = V_0V_0^T\mathbf{1}_k$. $y(\gamma)$ is very stable for small γ if $y_0 = 0$. Otherwise, $w^* = y_0/\mathbf{1}_k^T y_0$ is an optimal weight vector and the behavior depends on the magnitude of $\|y_0\|$. If $\|y_0\|$ is not small, the second term $\gamma^{-1}y_0$ dominates and $w(\gamma) = y(\gamma)/\mathbf{1}_k^T y(\gamma)$ tends to w^* quickly as $\gamma \to 0$. If $\|y_0\|$ is small, $w(\gamma)$ tends to $w_1 = y_1/\mathbf{1}^T y_1$ first and then turns back to w^* eventually, meaning that $w(\gamma)$ may not approximate to the optimal w^* within an acceptable accuracy, if γ is not small enough. Fig. 2 plots two error curves $\|w(\gamma) - w^*\|$ (solid line) and $\|w(\gamma) - w_1\|$ (dotted line) of three neighbor sets of the original swiss-roll data. As can be seen, this kind of variation will be more evident if $\|y_0\|$ is smaller.

The uncertainty from local weights may occur in LLE embedding. Here is a small illustrative example with $N = 20$ points sampled from a two-dimensional space. See the points marked by red-circles in Fig. 3. At each x_i, we choose $k = 4$ neighbors and the matrix G_i is now rank-deficient and $\dim(\mathcal{N}(G_i)) = 2$. We consider three sequences of weight vectors: $\{w_i^{(1)}\}$, $\{w_i^{(2)}\}$, and $\{w_i^{(3)}\}$, where $w_i^{(1)}$ and $w_i^{(2)}$ are two linear independent optimal weight vectors corresponding to G_i with zero residues, and $w_i^{(3)} = w_i(\gamma)$ is a regular weight vector with $\gamma = 10^{-4}$. However, the three two-dimensional embeddings $T^{(j)}$, $j = 1, 2, 3$, obtained by LLE are quite different to each other even within affine transformation. In the left of Fig. 3, we plot three affinely transformed coordinates $Y^{(j)} = c\mathbf{1}^T + LT^{(j)}$ of $T^{(j)}$ such that the error $\|X - Y^{(j)}\|$ is minimized. None of them can retrieve the ideal X.

Finally, we point out that the approximation (2.6) also implies an equivalence between an approximate linear reconstruction of x_i by its neighbors $\{x_j\}$ and an approximate linear reconstruction for t_i,

$$x_i = \sum_{j \in J_i} w_{ji}x_j + \eta_i \iff t_i = \sum_{j \in J_i} w_{ji}t_j + \delta_i. \tag{2.10}$$

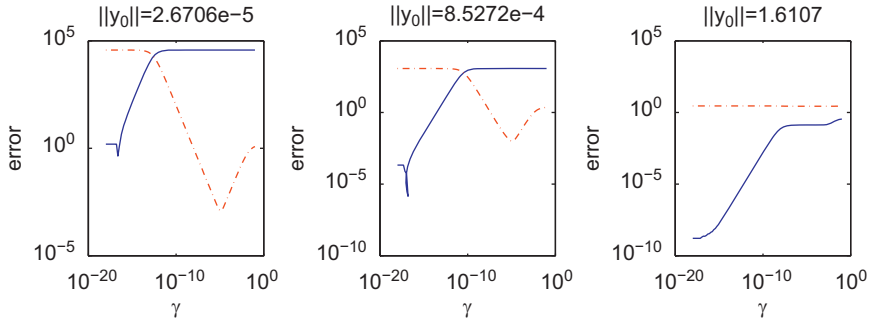

Fig. 2. The behavior of the weight vector $w(\gamma)$ for three fixed points of a swiss-roll data. The plots show the errors $\|w(\gamma)-w^*\|$ (solid line) and $\|w(\gamma)-w_1\|$ (dotted line). If y_0 is small, $w(\gamma)$ tends to w_1 first and then turns and approximates w^*, which will not happen if y_0 is not small.

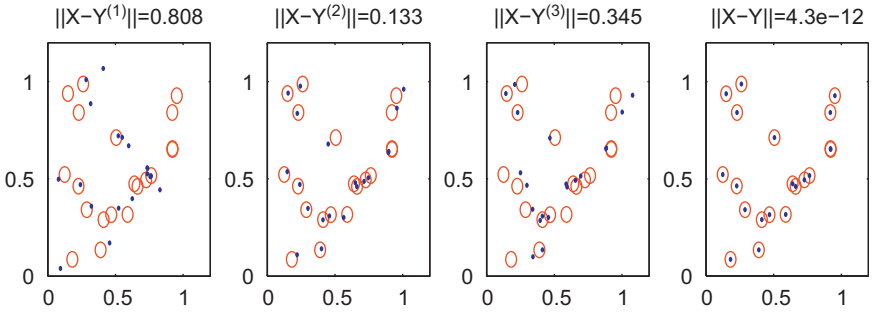

Fig. 3. The left three figures plot the LLE embeddings (under optimal affine transformation) based on two sets of exact local weight vectors and a set of regularization weights, respectively. The right one shows a perfect improvement by using multiple weight vectors.

Substituting (2.6) into (2.10), then we have

$$\eta_i = x_i - \sum_{j\in J_i} w_{ji}x_j = U_i t_i + \varepsilon_i - U_i\sum_{j\in J_i} w_{ji}t_j - \sum_{j\in J_i} w_{ji}\varepsilon_j$$
$$= U_i\left(t_i - \sum_{j\in J_i} w_{ji}t_j\right) + \varepsilon_i - \sum_{j\in J_i} w_{ji}\varepsilon_j = U_i\delta_i + \varepsilon_i - \sum_{j\in J_i} w_{ji}\varepsilon_j.$$

Note that ε_i and ε_j are orthogonal to U_i, and it follows

$$\delta_i = U_i^T\eta_i. \tag{2.11}$$

Obviously,

$$\|\delta_i\| \le \|\eta_i\|, \quad \|\eta_i\| \le \|\delta_i\| + \left\|\varepsilon_i - \sum_{j\in J_i} w_{ji}\varepsilon_j\right\|.$$

Eq. (2.7) is just for solving optimal weights for t_i. The existence of multiple weights for t_i implies the existence of multiple weights for x_i. In the next section, we will give a practical implementation of construction the required multiple weight sets.

3. Multiple local weight vectors

Let $X=[x_1,\dots,x_N]$ be the matrix of the data points $\{x_i\}$ sampled from a d-dimensional manifold with noise. Consider a particular point x_i with k nearest neighbors x_j, $j\in J_i$. According to the local linear model (2.6),[4]

$$x_j = c + Ut_j + \varepsilon_j, \quad j\in J_i \text{ or } j=i.$$

As before, the shifted matrix $G=[\dots,x_j-x_i,\dots]_{j\in J_i}$ has the form $G=UF+E$ with $F=[\dots,t_j-t_i,\dots]_{j\in J}$ and $E=[\dots,\varepsilon_j-\varepsilon_i,\dots]_{j\in J_i}$. Theoretically, F has a null space of dimension $s=k-d$, provided the data dimension $k>d$. Let Z_0 be an orthonormal basis matrix of the null space of F. We see that $\|GZ_0\|=\|EZ_0\|\le\|E\|$.[5] That is, the independent column vectors of Z_0 have small residues in magnitude of noise. So suitably normalizing these columns yields multiple independent and approximately optimal weight vectors.

In practical, approximately optimal weight vectors can be directly constructed by left singular vectors of G corresponding to the smallest singular values. Let $G=U\Sigma V^T$ be the SVD of G with the diagonal matrix Σ of the singular values $\sigma_1\ge\cdots\ge\sigma_k$. Assume that the r largest singular values are much larger than the remainders. Then the matrix V_0 consisting of the right singular vectors corresponding to the $s=k-r$ smallest singular values play a role similar to Z_0 on constructing approximately optimal weight vectors. To show it, let us partition $\Sigma=\mathrm{diag}(\Sigma_1,\Sigma_0)$ with Σ_1 consisting of the first r largest singular values. Conformably, $U=[U_1,U_0]$ and $V=[V_1,V_0]$. We see that

$$\|GV_0\| = \|U_0\Sigma_0\| = \sigma_{r+1} \le \|E\|. \tag{3.12}$$

The last inequality holds since $\|GV_0\|$ is the minimum of $\|GZ\|$ in the set of orthonormal matrices of s columns.

The approximately optimal weight vectors $\{w^{(1)},\dots,w^{(s)}\}$ constructed by the columns of V_0 are not unique, depending on the ways of normalization. We show below four normalization approaches.

The simplest way is to divide V_0e_ℓ by α_ℓ with $\alpha_\ell=\mathbf{1}_k^TV_0e_\ell$, i.e., $w^{(\ell)}=(1/\alpha_\ell)V_0e_\ell$, where e_ℓ is the ℓ- th column of the identity matrix. The resulting $w^{(1)},\dots,w^{(s)}$ are also orthogonal to each others but the corresponding residues are changed with a factor of

[4] For simplicity we delete the index i of c_i and U_i, and also for the others.

[5] In the paper, $\|\cdot\|$ denotes the 2-norm of a vector or matrix.

$1/\alpha_\ell$. This approach works well if all α_ℓ' s are not small. However, a small α_ℓ may obviously increase the reconstruction error.

The above risk can be reduced by averaging all α_ℓ' s. It can be done by orthogonally transforming the columns of V_0 to $\hat{V}_0 = V_0 H$ with an orthogonal matrix H such that the corresponding $\hat{\alpha}_\ell = \mathbf{1}_k^T \hat{V}_0 e_\ell$ are equal with $\hat{\alpha}_\ell = \alpha$. H can be easily determined as follows. Let $y_0 = (\alpha_1, \ldots, \alpha_s)^T = V_0^T \mathbf{1}_k$. We see that $H^T v_0 = H^T V_0^T \mathbf{1}_k = \hat{V}_0^T \mathbf{1}_k = \alpha \mathbf{1}_s$. So H can be a symmetric Householder matrix $H = I - 2uu^T$ [5] that transforms v_0 to $\alpha \mathbf{1}_s$. Here u can be constructed easily: $u = 0$ if $v_0 = \alpha \mathbf{1}_s$, otherwise, $u = (v_0 - \alpha \mathbf{1}_s)/\|v_0 - \alpha \mathbf{1}_s\|$. So

$$w^{(\ell)} = \frac{1}{\alpha} \hat{V}_0 e_\ell, \quad \mathbf{1}_k^T w^{(\ell)} = 1, \quad \ell = 1, \ldots, s. \tag{3.13}$$

The residue $\|Gw^{(\ell)}\|$ has the magnitude of noise if α is not small:

$$\|Gw^{(\ell)}\| = \frac{\|G\hat{V}_0 e_\ell\|}{\alpha} = \frac{\|GV_0 H e_\ell\|}{\alpha} \le \frac{\sigma_{r+1}}{\alpha}. \tag{3.14}$$

Otherwise the residue may be enlarged by small α.

The third approach of normalizing $V_0 e_\ell$ is to transform it to $(1-\alpha_\ell)w(\gamma) + V_0 e_\ell$. For consistency, we use the columns of $\hat{V}_0$, rather than the columns of V_0. This leads to the normalization

$$w^{(\ell)} = (1-\alpha)w(\gamma) + \hat{V}_0 e_\ell, \quad \ell = 1, \ldots, s. \tag{3.15}$$

It guarantees that $w^{(\ell)}$ in (3.15) has small residue,

$$\|Gw^{(\ell)}\| \le |1-\alpha| \|Gw(\gamma)\| + \sigma_{r+1}. \tag{3.16}$$

Note that $w^{(1)}, \ldots, w^{(s)}$ are no longer orthogonal but are still linearly independent.

More flexibly, one can set $w^{(\ell)}$ to be a linear combination of (3.13) and (3.15) to balance the linear independence of the weight vectors and the residue size,

$$w^{(\ell)} = \left(1 + \frac{\lambda(1-\alpha)}{\alpha}\right) \hat{V}_0 e_\ell + (1-\lambda)(1-\alpha) w(\gamma). \tag{3.17}$$

Obviously, $\lambda = 1$ gives the best independence (the weight vectors are orthogonal to each others) but possible large residue (only in the case when α is small). If $\lambda = 0$, it is (3.15) which is as optimal as $w(\gamma)$ if $\|E\|$ is negligible compared with $\|Gw(\gamma)\|$. The smallest residue $\|Gw(\gamma)\|$ can be achieved if $\lambda = \alpha/(\alpha - 1)$, but the independence of the weight vectors is lost (all the $w^{(\ell)}$ are equal to each others). We usually set $\lambda = \alpha$ to guarantee that $w^{(\ell)}$ also has small residue for small α as follows:

$$\|Gw^{(\ell)}\| \le (1-\alpha)^2 \|Gw(\gamma)\| + |2-\alpha| \sigma_{r+1}. \tag{3.18}$$

In general, the linear independence of the weight vectors $w^{(1)}, \ldots, w^{(s)}$ can be measured by the condition number $\mathrm{cond}(W)$ of the weight matrix

$$W = [w^{(1)}, \ldots, w^{(s)}] = \left(1 + \frac{\lambda(1-\alpha)}{\alpha}\right) V_0 H + (1-\lambda)(1-\alpha) w(\gamma) \mathbf{1}_s^T.$$

It is defined by the ratio of the largest singular value and the smallest singular value of W,

$$\mathrm{cond}(W) = \frac{\sigma_{\max}(W)}{\sigma_{\min}(W)}.$$

Obviously, $\mathrm{cond}(W) \ge 1$. The smaller the condition number is, the stronger the linear independence of the columns. The following theorem shows an upper bound of the condition number. Its proof is given in the Appendix.

Theorem 3.1. *Let W be the matrix of vectors $w^{(\ell)}$ in (3.17). Let $b = \max(|1 + \lambda(1-\alpha)/\alpha|, 1)$. Then*

$$\mathrm{cond}(W) \le \frac{(b + |(1-\lambda)(1-\alpha)|\sqrt{k}\|w(\gamma)\|)^2}{\left|1 + \frac{\lambda(1-\alpha)}{\alpha}\right|}.$$

Since $\|w(\gamma)\|$ is usually small, we approximately have

$$\mathrm{cond}(W) \le \frac{\max\left(\left|1 + \frac{\lambda(1-\alpha)}{\alpha}\right|, 1\right)^2}{\left|1 + \frac{\lambda(1-\alpha)}{\alpha}\right|}$$

and $\mathrm{cond}(W) \le \max(|2-\alpha|, 1)^2/|2-\alpha|$ if $\lambda = \alpha$. In out numerical experiments, we used the normalization (3.17) with $\lambda = \alpha$. The resulting W is very well-conditioned since α is closed to one or smaller. We mention that the parameter γ is not sensitive to the weight vectors $w^{(\ell)}$ in (3.17), since according to (3.16) and Theorem 3.1, both the residual norms and the linear independence depend only slightly on γ.

4. NEML: nonlinear embedding preserving multiple local-linearities

The existence of multiple weight vectors that are approximately optimal implies multiple linear constraints on the local linear structure of the manifold. We will use the multiple weight vectors to solidify the learned linearity with linear combinations. For robustness, we may adopt less s_i number of weight vectors for each neighbor set with $s_i \le k - d$, taking into account of big noise. We will show how to set the value of s_i later.

Consider the neighbor set $\mathcal{N}_i$ with k_i neighbors of x_i. Let $w_i^{(1)}, \ldots, w_i^{(s_i)}$ be s_i linearly independent column vectors in the weight matrix (3.17) with $\lambda = \alpha_i$ and $G = G_i$. We have

$$w_i^{(\ell)} = (1-\alpha_i)^2 w_i(\gamma) + (2-\alpha_i)\hat{V}_i h_\ell^{(i)}, \quad \ell = 1, \ldots, s_i. \tag{4.1}$$

Here $w_i(\gamma)$ is the regularized solution of (2.3), $\hat{V}_i$ is the right singular vector matrix of G_i corresponding to s_i smallest singular values of G_i, $\alpha_i = (1/\sqrt{s_i})\|v_i\|$ with $v_i = \hat{V}_i^T \mathbf{1}_{k_i}$, and $H_i = I - 2u_i u_i^T$ with u_i constructed as follows:

$$u_i = \begin{cases} \dfrac{u_{i0}}{\|u_{i0}\|} & \text{if } u_{i0} \ne 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{where } u_{i0} = \alpha_i \mathbf{1}_{s_i} - v_i.$$

We look for a d-dimensional embedding $\{t_1, \ldots, t_N\}$, $t_i \in \mathcal{R}^d$, that minimizes the sum of all reconstruction errors under the sequence of learned weight vectors among their neighbors, i.e., minimizes the embedding cost function

$$E(T) = \sum_{i=1}^{N} \sum_{\ell=1}^{s_i} \left\| t_i - \sum_{j \in J_i} w_{ji}^{(\ell)} t_j \right\|^2 \tag{4.2}$$

with the constraint $TT^T = I$. Let $W_i = [w_i^{(1)}, \ldots, w_i^{(s_i)}]$ with $w_i^{(\ell)}$ in (4.1),

$$W_i = (1-\alpha_i)^2 w_i(\gamma) \mathbf{1}_{s_i}^T + (2-\alpha_i)\hat{V}_i H_i, \tag{4.3}$$

and let $\hat{W}_i \in \mathcal{R}^{N \times s_i}$ be the embedded matrix of W_i into the N-dimensional space such that

$$\hat{W}_i(J_i, :) = W_i, \quad \hat{W}(i, :) = -\mathbf{1}_{s_i}^T, \quad \hat{W}(j, :) = 0, \quad j \notin I_i = J_i \cup \{i\}.$$

The cost function (4.2) can be rewritten as

$$E(T) = \sum_i \|T\hat{W}_i\|_F^2 = \mathrm{trace}\left(T \sum_i \hat{W}_i \hat{W}_i^T T^T\right) = \mathrm{trace}(T\Phi T^T), \tag{4.4}$$

where

$$\Phi = \sum_i \hat{W}_i \hat{W}_i^T. \tag{4.5}$$

Φ can be constructed in the updating way $\Phi xlarr; \Phi + \hat{W}_i \hat{W}_i^T$, $i = 1, \ldots, N$, starting with $\Phi = 0$. The i-th updating changes only

the rows and columns with index $j \in I_i = \{i\} \cup J_i$ by

$$\begin{cases} \Phi(i,i) \leftarrow \Phi(i,i) + s_i, \\ \Phi(J_i,J_i) \leftarrow \Phi(J_i,J_i) + W_i W_i^T, \\ \Phi(J_i,i) \leftarrow \Phi(J_i,i) - W_i \mathbf{1}_{s_i}, \\ \Phi(i,J_i) \leftarrow \Phi(J_i,i)^T. \end{cases} \tag{4.6}$$

As in LLE, the minimizer of $E(T)$ is given by the matrix $T = [u_2, \ldots, u_{d+1}]^T$ of the d eigenvectors of Φ corresponding to the 2nd to $d+1$ st smallest eigenvalues.

Now we consider how to determine the number s_i of approximate optimal weight vectors. Basically, s_i should be selected as large as possible such that $\sigma_{k_i - s_i + 1}(G_i)$ is relatively small. In general, if the data points are sampled from a d-dimensional manifold and the neighbor set is well selected, then $\sigma_d(G_i) \gg \sigma_{d+1}(G_i)$. So we can set $s_i = k_i - d$. However, it is possible that $\sigma_{d+1}(G_i)$ is not significantly smaller than $\sigma_d(G_i)$ if the neighbors are not well-selected or there is relative large data noise. In that case, let $r_i \geq d$ be the smallest integer such that $\sigma_{r_i+1}(G_i)$ is acceptably small compared with $\sigma_{r_i}(G_i)$, or equivalently, the ratio $\sum_{j=r_i+1}^{k_i} \lambda_j^{(i)} / \sum_{j=1}^{r_i} \lambda_j^{(i)}$ is small where $\lambda_j^{(i)} = \sigma_j^2(G_i)$ are the eigenvalues of $G_i^T G_i$. Then one can set $s_i = k_i - r_i$. Equivalently, we set

$$s_i = \max_{\ell} \left\{ \ell | \ell \leq k_i - d, \ \frac{\sum_{j=k_i-\ell+1}^{k_i} \lambda_j^{(i)}}{\sum_{j=1}^{k_i-\ell} \lambda_j^{(i)}} < \eta \right\} \tag{4.7}$$

for a given threshold $\eta > 0$. This scheme permits an overestimated dimensionality d of the hidden manifold. The accuracy parameter η can be set beforehand, according to the noise level. We suggest adaptively set η to be the mid-value of

$$\rho_i = \frac{\sum_{j=d+1}^{k_i} \lambda_j^{(i)}}{\sum_{j=1}^{d} \lambda_j^{(i)}}, \quad i = 1, \ldots, N,$$

in the nondecreasing order, i.e., $\eta = \rho_{\pi_{\lceil N/2 \rceil}}$ with $\rho_{\pi_1} \leq \cdots \leq \rho_{\pi_N}$.

Now we are ready to summarize the above method into the following algorithm which is called as nonlinear embedding preserving multiple local-linearities or NEML for short.

Algorithm. NEML (nonlinear embedding preserving multiple local-linearities). Given a date set of N points $x_i \in \mathcal{R}^m$, this algorithm produces a matrix $T \in \mathcal{R}^{d \times N}$ of N d-dimensional coordinates $t_i \in \mathcal{R}^d$ for the dimension reduction of x_i' s.

1. For each $i = 1, \ldots, N$,
   1.1 Determine $\mathcal{N}_i = \{x_j, j \in J_i\}$ of x_i and set $G_i = [\ldots, x_j - x_i, \ldots]_{j \in J_i}$.
   1.2 Compute the regularized solution $w_i(\gamma)$ by (2.5) with a small $\gamma > 0$.
   1.3 Compute the eigenvalues $\lambda_1^{(i)}, \ldots, \lambda_{k_i}^{(i)}$ and eigenvectors $v_1^{(i)}, \ldots, v_{k_i}^{(i)}$ of $G_i^T G_i$. Set $\rho_i = \frac{\sum_{j=d+1}^{k_i} \lambda_j^{(i)}}{\sum_{j=1}^{d} \lambda_j^{(i)}}$.
2. Set η to be the mid-value of $\{\rho_i\}$ and $\{s_i\}$ by (4.7).
3. Construct Φ by (4.24) and (4.21) starting with $\Phi = 0$.
4. Compute the $d+1$ eigenvectors $u_2, \ldots, u_{d+1}$ of Φ corresponding to the 2nd to $d+1$ st smallest eigenvalues, and set $T = [u_2, \ldots, u_{d+1}]^T$.

The computational cost of NEML is almost the same as that of LLE. The additional flops of NEML for computing the eigendecomposition of $G_i^T G_i$ is $O(k_i^3)$. Totally NEML has more $O(k^3 N)$ flops than LLE with $k = \max_i k_i$. Note that the most expensive steps in both LLE and NEML are to select neighbor sets and compute the eigenvectors of Φ corresponding to $d+1$ small eigenvalues which cost $O(mN^2)$ and $O(dN^2)$, respectively [13]. Because k is much smaller than the data size N, the additional cost of NEML is negligible.

The right panel of Fig. 3 shows a perfect improvement by NEML. In the last section, we will give numerical results of the algorithm NEML and compare NEML with LLE in applications.

5. A simple analysis of NEML for isometric manifolds

Consider the application of NEML on an isometric manifold $\mathcal{M} = f(\Omega)$ with open set $\Omega \subset \mathcal{R}^d$ and smooth function f. Assume that $\{x_i\}$ are sampled from $\mathcal{M}$, $x_i = f(\tau_i), i = 1, \ldots, N$.

Using first-order Taylor expansion of f at x_i, a neighbor x_j, $j \in J_i$ of x_i can be represented by

$$x_j = x_i + J_{\tau_i} \cdot (\tau_j - \tau_i) + O(\|\tau_j - \tau_i\|^2), \tag{5.1}$$

where $J_{\tau_i} \in \mathcal{R}^{m \times d}$ is the Jacobian matrix of f at τ. Let $\varepsilon_i = \max_j \|\tau_j - \tau_i\|$. Then we have

$$\left\| x_i - \sum_{j \in J_i} w_{ji} x_j \right\| = \left\| \tau_i - \sum_{j \in J_i} w_{ji} \tau_j \right\| + O(\varepsilon_i^2) \tag{5.2}$$

due to the isometry of f. Since $k_i > d$, the first term on the right of the equality above should be zero for suitable weights. So the optimality implies $\|x_i - \sum_{j \in J_i} w_{ji}^* x_j\| = O(\varepsilon_i^2)$.

On the other hand, assume that $\|G_i w(\gamma)\| \approx O(\varepsilon_i^2)$. By (3.18), we have that for the approximately optimal weight vectors $w_i^{(\ell)}$ given by (4.1),

$$\left\| x_i - \sum_{j \in J_i} w_{ji}^{(\ell)} x_j \right\| = \|G_i w_i^{(\ell)}\| \leq |2 - \alpha_i| \sigma_{k_i - s_i + 1}(G_i) + O(\varepsilon_i^2).$$

Thus, by (5.2) we also have the estimation

$$\left\| \tau_i - \sum_{j \in J_i} w_{ji}^{(\ell)} \tau_j \right\| \approx |2 - \alpha_i| \sigma_{k_i - s_i + 1}(G_i) + O(\varepsilon_i^2).$$

Therefore, denoting $T^* = [\tau_1, \ldots, \tau_N]$, we have

$$\begin{aligned} E(T^*) &= \sum_{i=1}^{N} \sum_{\ell=1}^{s_i} \left\| \sum_{j \in J_i} w_{ji}^{(\ell)} \tau_j - \tau_i \right\|^2 \\ &\leq \sum_{i=1}^{N} s_i (2 - \alpha_i)^2 \sigma_{k_i - s_i + 1}^2(G_i) + O\left(\max_i \varepsilon_i^2 \right). \end{aligned} \tag{5.3}$$

Now writing $T^* = RU$ with a row-orthonormal matrix U, i.e., $UU^T = I$, we have that $\sigma_d(R) = \sigma_d(T^*)$ and $E(U) \leq E(T^*)/\sigma_d^2(T^*)$. By this inequality, an upper bound of $E(U)$ follows from (5.3) directly. Note that $\sigma_{r_i+1}^2(G_i)$ is usually very small. $E(U)$ is always small and approximately achieves the minimum of E over the set of row-orthonormal matrices with d rows. So roughly speaking, we can expect that NEML can retrieve the isometric embedding.

6. Comparison with LTSA

NEML has properties similar to those of the local tangent space alignment (LTSA). Firstly, both NEML and LTSA learn the local linear dependence of each neighborhood using multiple constraints. Secondly, they construct low dimensional embeddings by solving eigenvectors of matrices that have structures similar to each other. In this section, we compare NEML and LTSA in these two issues. For simplicity, we assume that $k_i - d$ weight vectors are used in NEML for each neighbor set.

6.1. Linear dependence of neighbors

NEML builds the linear dependence of neighbors using $k_i - d$ linearly independent weight vectors $w_i^{(1)}, \ldots, w_i^{(k_i-d)}$ that approximately minimize the reconstruction error $\|x_i - \sum_{j\in J_i} w_{ji}x_j\|$. Let us consider the total reconstruction errors

$$\varepsilon^{NEML}(\mathcal{N}_i) = \sum_{\ell=1}^{k_i-d} \left\| x_i - \sum_{j\in J_i} w_{ji}^{(\ell)} x_j \right\|^2 = \|G_i W_i\|_F^2.$$

It measures the linear dependence of the neighbor set. Let $I_i = \{i\} \cup J_i$ and let $\overline{x}_i = (1/|I_i|)\sum_{j\in I_i} x_j$ be the mean of the members in the neighbor set. Let $\overline{X}_i = [\ldots, x_j - \overline{x}_i, \ldots]_{j\in I_i}$. It can be verified that $G_i W_i = \overline{X}_i \tilde{W}_i$ with $\hat{W}_i(J_i,:) = W_i$ and $\tilde{W}_i = \hat{W}_i(I_i,:)$. So

$$\varepsilon^{NEML}(\mathcal{N}_i) = \|\overline{X}_i \tilde{W}_i\|_F^2.$$

Approximately, we have $\|\overline{X}_i \tilde{W}_i\|_F = \min_{\mathbf{1}^T\tilde{W}=0} \|\overline{X}_i \tilde{W}\|_F$.

In LTSA, the local linear structure is established by the optimal linear fitting

$$\min_{c_i,\theta_j,U_i^TU_i=I} \sum_{j\in I_i} \|x_j - (c_i + U_i\theta_j)\|^2.$$

It is solved by $c_i = \overline{x}_i$, $U_i = Q_i$, the matrix of left singular vectors of $\overline{X}_i$ corresponding to the d largest singular values, and $\theta_j = \theta_j^{(i)} = Q_i^T(x_j - \overline{x}_i)$. Q_i can be obtained by the SVD of $\overline{X}_i$,

$$\overline{X}_i = [Q_i, \hat{Q}_i]\,\mathrm{diag}(\Sigma_i, \hat{\Sigma}_i)[V_i, \hat{V}_i]^T, \tag{6.1}$$

where both Q_i and V_i are of d columns. The linear dependence of $\mathcal{N}_i$ can also be measured by the total error

$$\varepsilon^{LTSA}(\mathcal{N}_i) = \sum_{j\in I_i} \|x_j - \overline{x}_i - Q_i\theta_j^{(i)}\|^2 = \|\overline{X}_i - Q_i\Theta_i\|_F^2 = \|\overline{X}_i\hat{V}_i\|_F^2.$$

The measure functions ε^{NEML} and ε^{LTSA} of neighborhood linear dependence used in NEML and LTSA are similar,

$$\varepsilon^{NEML}(\mathcal{N}_i) = \|\overline{X}_i\tilde{W}_i\|_F^2 \approx \min_{\mathbf{1}^T\tilde{W}=0} \|\overline{X}_i\tilde{W}\|_F^2, \varepsilon^{LTSA}(\mathcal{N}_i) = \|\overline{X}_i\hat{V}_i\|_F^2 = \min_{Z^TZ=I} \|\overline{X}_i Z\|_F^2.$$

Note that $\hat{V}_i$ is an orthogonal bases of the null space of $\mathrm{span}([\mathbf{1}_{k_i+1}, \Theta_i^T]^T)$, i.e., $\hat{V}_i[\mathbf{1}_{k_i+1}\Theta_i^T] = 0$. We also have $\tilde{W}_i\mathbf{1}_{k_i+1} = 0$.

6.2. Alignment matrices

Both NEML and LTSA minimize a trace function of an *alignment* matrix Φ to obtain an embedding,

$$\min_{TT^T=I} \mathrm{trace}(T\Phi T^T).$$

The alignment matrix can be written in the same form $\Phi = \sum_{i=1}^N S_i\Phi_i S_i^T$, where Φ_i is a positive semidefinite matrix determined by the local linear structures learned, and S_i is a selection matrix consisting of the columns $j \in I_i$ of the large identity matrix of order N. $S_i y$ is the embedding of a (k_i+1)-dimensional vector y to an N-dimensional space by setting zero components in the complement set of I_i. In LTSA, the local matrix Φ_i is given by the orthogonal projection

$$\Phi_i^{LTSA} = \hat{V}_i\hat{V}_i^T,$$

whose null space is $\mathrm{span}([\mathbf{1}_{k_i+1}, \Theta_i^T])$, see [20]. We denote by Φ^{LTSA} the global alignment matrix of LTSA, $\Phi^{LTSA} = \sum_{i=1}^N S_i\Phi_i^{LTSA}S_i^T$.

For NEML, since $\hat{W}_i = S_i\tilde{W}_i$,

$$\Phi^{NEML} = \sum_{i=1}^N S_i\tilde{W}_i\tilde{W}_i^T S_i^T = \sum_{i=1}^N S_i\Phi_i^{NEML}S_i^T,$$

with

$$\Phi_i^{NEML} = \tilde{W}_i\tilde{W}_i^T.$$

It is interesting that the range space $\mathrm{span}(\tilde{W}_i)$ of $\tilde{W}_i$ and the range space $\mathrm{span}(\hat{V}_i)$ of $\hat{V}_i$ are very close to each other if the reconstruction error of x_i is small. The following theorem gives an upper bound of the closeness using the distance $\mathrm{dist}(\tilde{W}_i, \hat{V}_i)$ between the subspaces $\mathrm{span}(\tilde{W}_i)$ and $\mathrm{span}(\hat{V}_i)$ that denotes the largest angle between the two subspaces. (See [5] for discussion about distance of subspaces.)

Theorem 6.1. *Let* $G_i = [\ldots, x_j - x_i, \ldots]_{j\in J_i}$. *Then*

$$\mathrm{dist}(\tilde{W}_i, \hat{V}_i) \le \frac{\|G_iW_i\|}{\sigma_d(\tilde{W}_i)\sigma_d(\overline{X}_i)}.$$

The proof is given in the Appendix. As we have shown before, if x_i and its neighbors are well selected from a d dimensional manifold, $\|G_iW_i\|$ is small and hence by Theorem 6.1, the range spaces of $\tilde{W}_i$ and $\hat{V}_i$ are approximately equal. The difference is that Φ_i^{LTSA} is an orthogonal projection while the projection Φ_i^{NEML} is an affine transformation.

7. Experimental results

In this section, we give several numerical results that illustrate the performance of the algorithm NEML. We first consider two synthetic data sets which are generated in a non-convex domain or have variant curvatures, respectively, to highlight the essential improvement without interference from noise. To show the efficiency of NEML on real data which usually contain large noises, five real world examples from recognition, classification, and data visualization are tested: a handwritten digit set, a news data set, a cancer data set, an image set which basically depends on a parameter, and an image set depending on three generating parameters. We will compare LLE and NEML, together with Isomap, LTSA and LE, applied on the real data for classification, recognition, and retrieving the hidden parameters.

7.1. Synthetic data

The first example is the data of swiss roll with a rectangle hole in its generating domain. The manifold is parameterized by

$$f(s,t) = (t\cos(t), s, t\sin(t))^T.$$

We set $N = 1500$ points $x_i = f(s_i, t_i)$ with generating parameters (s_i, t_i) sampled from the domain $(0,21) \times (3\pi/2, 9\pi/2)$ with a small rectangle in normal distribution. We applied Isomap, LLE, LTSA, and NEML on the data set with neighborhood size $k = 15$. Since Isomap cannot retrieve the intrinsic low dimension structure if the domain is not convex [13], there is an obvious dilation in the computed two-dimensional domain by Isomap near the missing region. A strong distortion also occurs on the coordinates computed by LLE. NEML, as well as LTSA, can greatly reduce the distortion and retrieve the generating coordinates well for this example. The computed coordinates are plotted in Fig. 4.

To illustrate the influence of the regularizer γ on the computed embeddings of LLE and NEML, we tested γ ranging from 10^{-10} to 10^{-1} and checked the deviation for the computed coordinates T from the generating coordinates T_{true}, which is defined by the relative affine error

$$\mathrm{dev}(T) = \min_{c,L} \frac{\|T_{true} - (c\mathbf{1}^T + LT)\|_2}{\|T_{true}\|_2}. \tag{7.1}$$

Of course, smaller deviation yields better computed embedding for recovering the hidden manifold structures. In Fig. 5, we plot the deviation functions $\mathrm{dev}(T)$ of the two-dimensional embedding T computed by NEML and LLE using regularized weights with respect to the regularization parameter γ. It is clear that NEML

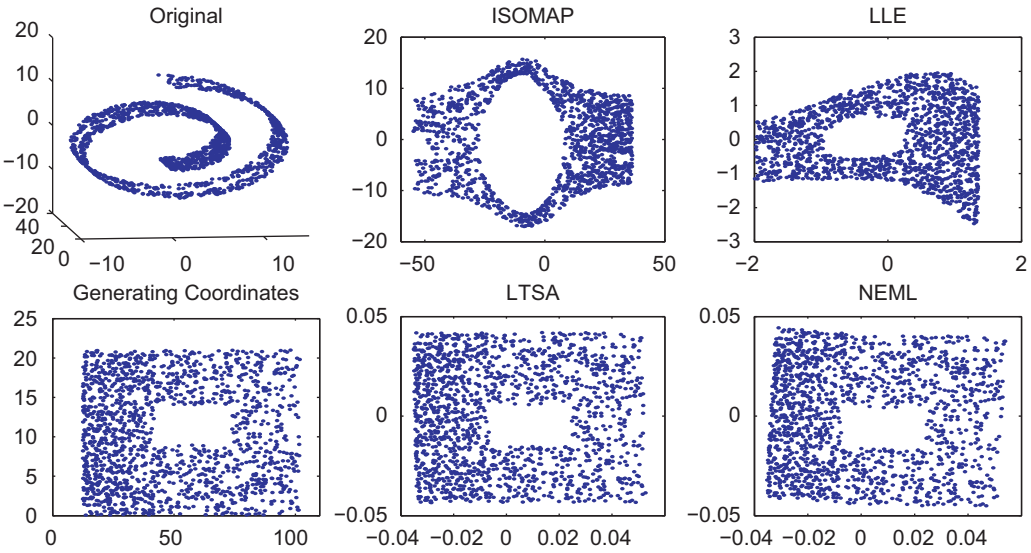

Fig. 4. Comparison of Isomap, LLE, LTSA, and NEML applied on the swiss roll with a hole. The computed coordinates by Isomap and LLE have strong distortions. NEML, as well as LTSA, performs well.

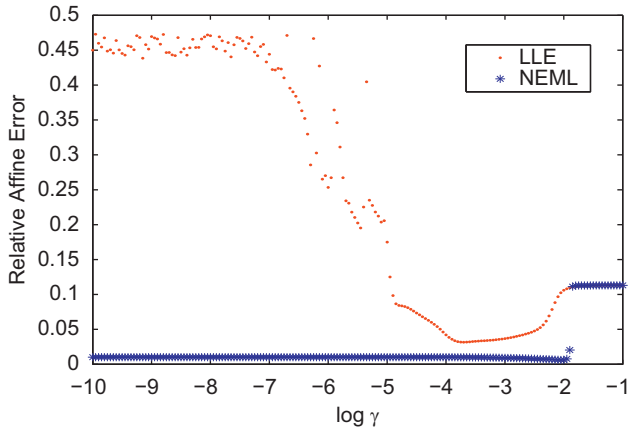

Fig. 5. Function dev(T) with respect to γ variant from 10^{-10} to 10^{-1}, computed by LLE and NEML applied on the data set of the swiss roll with a hole.

embedding has much smaller deviation than LLE embedding. The effect of γ on NEML is very slight if γ is not large, while LLE is very sensitive to γ, especially to small one.

The second synthetic data are a set of points on the triple-peak manifold $f(t,s)=(t,s,h(t,s))^T$ with

$$h(t,s)=e^{-10((t-0.5)^2+(s-0.5)^2)}-e^{-10(t^2+(s+1)^2)}-e^{-10((1+t)^2+s^2)}.$$

We generated $N=1225$ samples with the generating parameters t and s equally spaced in the interval $[-1.5,1.5]$, containing the triple-peak. The data points and the generating parameters are plotted on the left of Fig. 6. We applied Isomap, LLE, LTSA, and NEML on this data set with $k=15$, and compare the computed coordinates with the generating parameters.[6] Both Isomap and LLE cannot recover the generating parameters—there are large deformations nearby the parameters of the peaks on the computed two-dimensional coordinates by Isomap and LLE, due to the large variation in local curvatures. This bias is much reduced by the modified curvature model of LTSA proposed in [18]. It is a little bit surprising that the coordinates obtained by NEML are quite perfect—NEML recovers the generating parameter up to an affine transformation. The distortions are almost invisible. See the right two columns in Fig. 6 for the computed two-dimensional coordinates by Isomap, LLE, LTSA, and NEML.

[6] Since the manifold is not isometric nearby the peaks, it is difficult to give a criterium to evaluate which embedding is 'ideal'. Intuitively, it is reasonable to compare the computed embedding with the generating parameters.

7.2. Classification

In this subsection, we consider the application of NEML, together with LLE and Laplacian Eigenmaps (LE), on classification. Three real data sets from handwritten recognition, texture recognition, and breast cancer data classification are tested. The data points—training points and testing points—are nonlinearly projected into a lower-dimensional space by these manifold learning algorithms at first, and then the projected test points are classified according to the projection and the classification labels of the training points. Two simple classifiers, nearest-neighbor (NN) and nearest feature line (NFL) classifiers [10], are used for recognizing the testing points in our experiments. Other kinds of classification methods such as LDA [1,15] can be applied on the projected points. We do not want to say much on this topic in this paper.

There are two strategies for projecting the training set and testing set: separate projection or holistic projection. By separate projection we mean that training points are first projected, using the existing dimensionality reduction algorithms, and then testing points are mapped by linear interpolation in terms of the projected coordinates of their nearest training points. This strategy is convenient for classifying testing points that arrive successively, but the projection ways for testing points and for training points are not completely conformed to each other. The holistic projection maps all training points and testing points simultaneously. It provides conformed low-dimension coordinates for both training set and testing set, and results in a higher accuracy than the separate projection. However, this strategy is computationally demanding for newly arrived points since it requires to repeat the algorithms for newly arrived data points.

Handwritten digits: We first consider the applications of LLE, LE, and NEML in recognition of the handwritten digits in 10 classes ('0'–'9') from the USPS database [7]. The database has 11 000 samples of the 10 digits (1100 per digit) in 16×16 grayscale

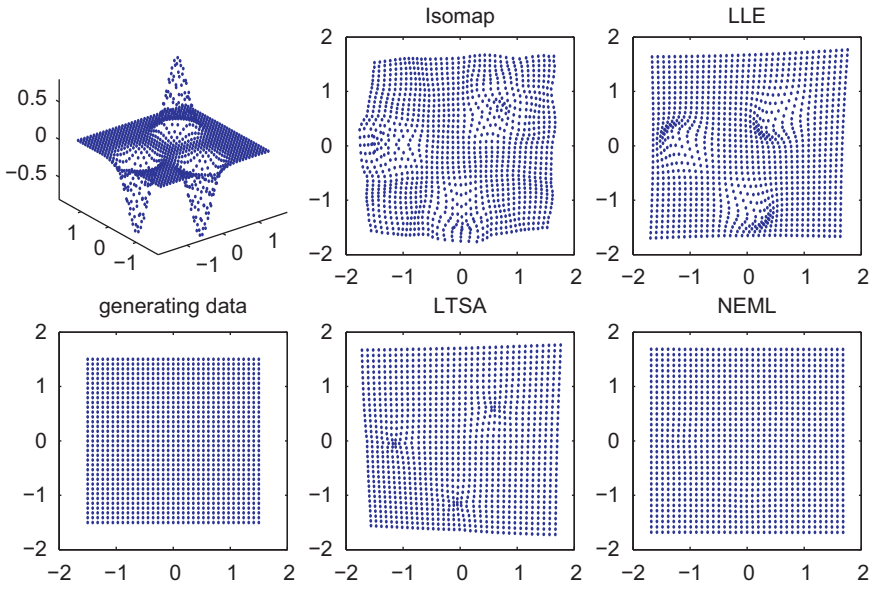

Fig. 6. Left column: the data set and its generating points. Right two columns: the computed two-dimensional coordinates (with affine transformation) by Isomap, LLE, LTSA (curvature version), and NEML. NEML recovers the generating coordinates perfectly up to an affine transformation.

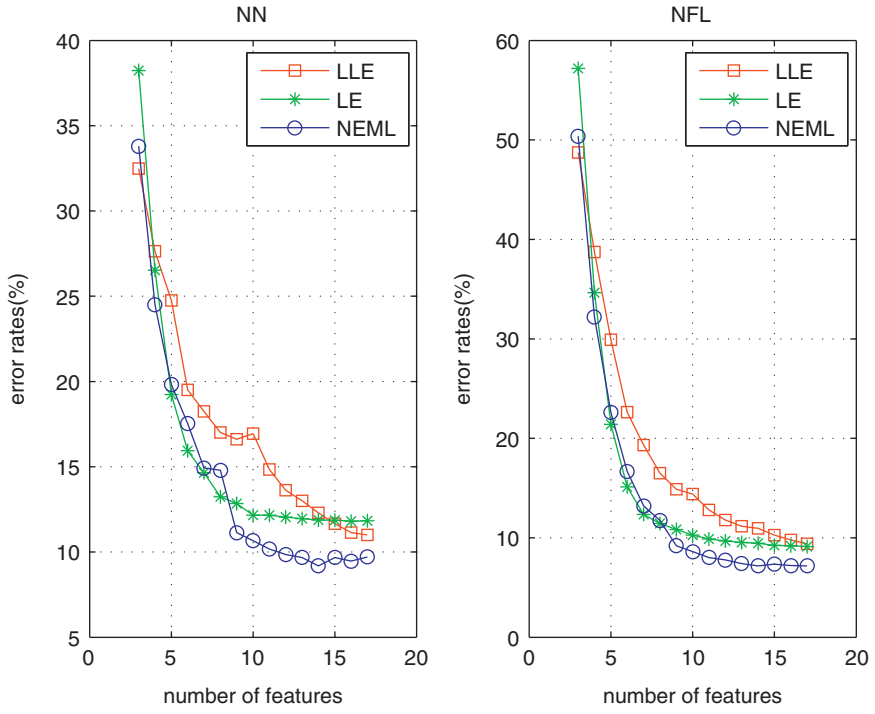

Fig. 7. The average NN- and NFL-recognition errors (in %) of the 6000-point set of handwritten digits USPS embedded by LLE and NEML with neighborhood size $k=20$.

Table 1
The average classification errors and standard deviation (in %) for 20newsgroups.

Classi.	Sp. proj.	$d=3$	6	9	12	15
NN	LLE	65.8 ± 1.5	53.0 ± 1.1	45.7 ± 1.0	42.2 ± 1.1	39.1 ± 0.9
	LE	62.6 ± 1.3	46.5 ± 0.1	39.4 ± 0.1	35.2 ± 0.4	30.5 ± 0.2
	NEML	42.1 ± 3.9	33.9 ± 4.8	29.8 ± 1.7	28.0 ± 0.9	27.0 ± 0.7
NFL	LLE	68.2 ± 0.9	55.8 ± 1.1	44.0 ± 1.6	36.7 ± 1.3	32.4 ± 1.1
	LE	64.5 ± 1.5	45.4 ± 0.4	33.6 ± 0.8	28.2 ± 0.4	25.3 ± 0.5
	NEML	48.4 ± 3.6	32.3 ± 2.7	27.6 ± 1.5	25.0 ± 0.9	22.9 ± 0.4

images. Each digit image is represented by a 256-dimension column vector in our experiment. Due to the failure of MATLAB's eigs function for computing eigenvectors of the resulting large sparse matrix for the whole data set of 11 000 samples, we just used 6000 data points (600 each digit class). Half of each class is randomly selected to form the training set and other half is the testing set. The holistic strategy is used for projection. We set the neighborhood size $k=20$ and variant d from 3 to 17. The random training–testing partition is repeated for 25 times for each d. In Fig. 7, we plot the percentages of recognition errors versus the

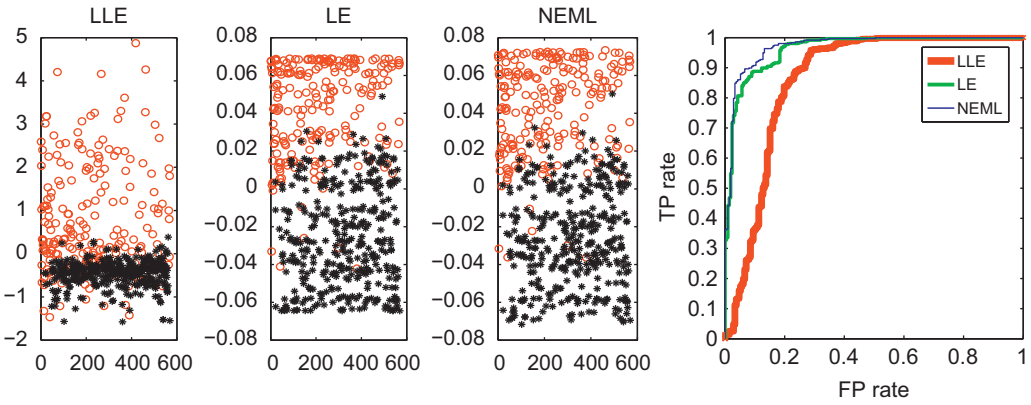

Fig. 8. Left: the computed one-dimensional results by LLE, LE and NEML with $k=10$ on the WDBC data. The red 'o' points represent the benign samples and the black '*' points represent the malignant ones. Right: ROC curves comparing NEML with LLE and LE. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

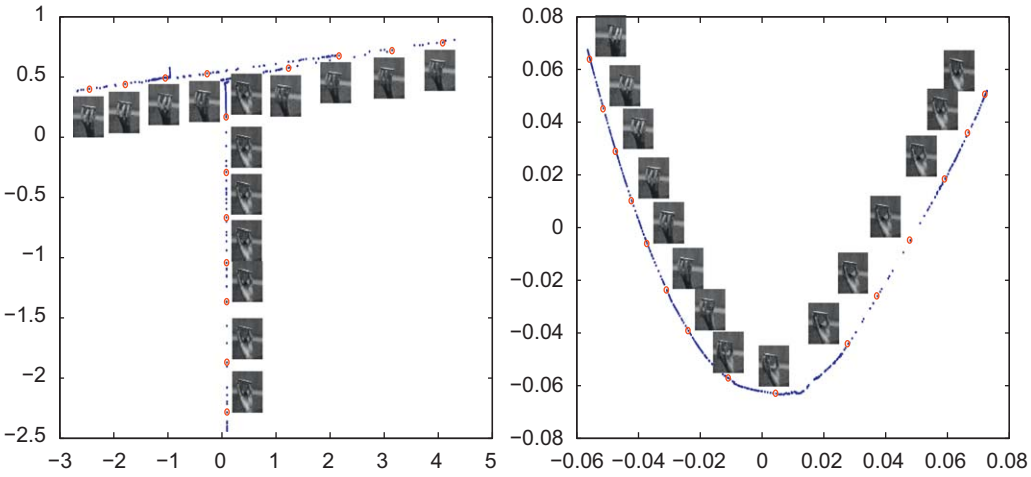

Fig. 9. Two-dimensional embeddings of the hand image set computed by LLE (left) and NEML (right) with $k=10$ and some corresponding images.

feature numbers. As can be seen, the NEML projection performs better than LLE embedding in both the NN or NFL classification for almost every feature number. Compared with the LE embedding, NEML leads to similar error rates for small number of features. However, as the number of features approaches to the neighborhood size, NEML outperforms LE in decreasing the recognition errors.

The 20newsgroups data.[7] The testing data are from the 20newsgroups data with binary occurrence and consists of 9060 postings each of which has 100 words. It is obtained by deleting repeated postings in the original data set. The data points can be divided into four groups with group sizes 2804, 1790, 1684, and 2782. Half of each group is randomly selected as training set and the remainder forms the testing set. The separate projection strategy is used in this experiment. We set $k=20$ and $d=3,6,9,12,15$. The experiment is repeated for 10 times, each has different training/testing partitioning. Table 1 lists the average and standard deviation of percentages of classification errors among the repeated testings for each d. Obviously, NEML improves LLE and LE significantly in both NN and NFL, whatever the value of d is set. We remark that the classification errors of NN and NFL applied to the original high dimensional data are $35.21\pm0.54\%$ and $23.96\pm0.31\%$, respectively. LLE projection cannot improve the NN or NFL-classification for any d we tested in this example. The improvement of NEML-projection on both NN or NFL-classification is significant, especially on NN.

WDBC data.[8] It is to detect malignant breast tumors from a set mixing benign and malignant samples. The data set, called Wisconsin Diagnostic Breast Cancer (WDBC) data, contains 569 samples—357 tumors are benign and the other 212 tumors are malignant. Each sample is represented by a 32-dimensional vector whose first two attributes (ID and diagnosis) are deleted since they are inessential to the detection. So the data points in this experiment are 569 30-dimensional vectors. We consider one-dimensional projections by LLE, LE, and NEML with $k=10$. The projected one-dimensional points are plotted in the left three sub-figures of Fig. 8. The points of benign tumors are marked by small red circles and small black stars for malignant tumors. NEML and LE outperform LLE to a large extent

[7] http://www.cs.toronto.edu/roweis/data.html.

[8] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

in this example: NEML and LE separate most of the benign cases and the malignant ones, while the LLE projection mixes the two kinds of tumors almost together. We also plot the receiver operating characteristic (ROC) curves [14] of the recognition by LLE, LE, and NEML in the right of Fig. 8. In the ROC curves, the vertical axis is the rate of true positives (TP) and the horizontal axis is the rate of false positives (FP). Here, TP means that a malignant tumor is detected correctly, and FP is that a benign tumor is incorrectly detected as a malignant tumor. The larger the area under the curve, the better the performance. It is clear that our algorithm has the best performance in this experiment.

7.3. Visualization

Two image data sets are considered. One is a collection of 481 images that show a rolling hand that holds a rice bowl.[9] These images depend on only one parameter, namely, horizontal rotation. The original image is of size 512×480. We resize each image to 51×48 using the bilinear interpolation method, and each image is represented in a 2448-dimensional column vector. So the data matrix X is of size 2448. For this data set, LLE cannot generate a reasonable two-dimensional embedding, see the left of Fig. 9 for the two-dimensional coordinates computed by LLE and some corresponding images to the circled points. This result can be improved by using multiple local weight vectors significantly. The right of Fig. 9 shows the NEML embedding results. The smooth trace of the two-dimensional points reflects the true horizontal rotation of the hand.

The other image set is of 698 face images from [16]. These face images depend on three parameters: two for poses (left-right and up-down) and one for lighting. The image size is 64-by-64 pixel and hence each sample is converted to an $m = 4096$- dimensional column vector. We project the data into a three-dimensional space nonlinearly by LLE and NEML with $k = 14$ in order to recover the three hidden parameters. The efficiency of the computed embedding is measured by the deviation function (relative affine error) $\mathrm{dev}(T)$ defined in (7.1) with T_{true} the matrix of the true light-parameter and the two pose-parameters. NEML produces a better nonlinear projection than LLE since the deviation of NEML is smaller than that of LLE,

$$\mathrm{dev}(T_{LLE}) = 0.2286, \quad \mathrm{dev}(T_{NEML}) = 0.0930.$$

Appendix

Proof of Theorem 2.1. If $\mathbf{1}_k$ is not orthogonal to $\mathcal{N}(F)$, there is a vector $y \in \mathcal{N}(F)$ such that $y^T\mathbf{1}_k \neq 0$. Thus $w = y/y^T\mathbf{1}_k$ satisfies $Fw = 0$ and $w^T\mathbf{1}_k = 1$. That is, w is optimal and any optimal weight vector should be also a null vector.

On the other hand, applying Lagrange multiplier method on the constrained minimization problem $\min_{w^T\mathbf{1}_k = 1}\|Fw\|$, an optimal solution w must satisfy $F^TFw = \lambda\mathbf{1}_k$ with a certain scalar λ. So we can write w as

$$w = z + \lambda(F^TF)^{\dagger}\mathbf{1}_k = z + \lambda y_1,$$

with a null vector z of G. If $\mathbf{1}_k$ is orthogonal to the null space $\mathcal{N}(F)$, then $\mathbf{1}_k^Tz = 0$. Since $\mathbf{1}_k^Tw = 1$, we have $1 = \lambda\mathbf{1}_k^Ty_1$. It follows that $\mathbf{1}_k^Ty_1 \neq 0$ and $\lambda = 1/\mathbf{1}_k^Ty_1$. Thus,

$$w = z + \frac{y_1}{\mathbf{1}_k^Ty_1} = z + w^*.$$

So by $Fw = Fw^*$ we conclude that w is an optimal weight vector if and only if $w = z + w^*$ for some null vector z of G. □

Proof of Theorem 3.1. Obviously,

$$\sigma_{\max}(W) = \|W\| \leq |(1-\lambda)(1-\alpha)|\sqrt{s}\|w(\gamma)\| + \left|\frac{\lambda}{\alpha} + 1 - \lambda\right| = \sqrt{s}\varsigma + \left|\frac{\lambda}{\alpha} + 1 - \lambda\right|,$$

where $\varsigma = |(1-\lambda)(1-\alpha)|\|w(\gamma)\|$. On the other hand,

$$\|Wy\| = \left\|(1-\lambda)(1-\alpha)w(\gamma)\mathbf{1}_s^Ty + \left(\frac{\lambda}{\alpha} + 1 - \lambda\right)V_0Hy\right\| \geq \left|\frac{\lambda}{\alpha} + 1 - \lambda\right|\|y\| - \varsigma\|\mathbf{1}_s^Ty\|.$$

Notice that $\|\mathbf{1}_s^Ty\| = \|\mathbf{1}_k^TWy\| \leq \sqrt{k}\|Wy\|$. Substituting it into the inequality above, we have

$$\|Wy\| \geq \left|\frac{\lambda}{\alpha} + 1 - \lambda\right|\|y\| - \sqrt{k}\varsigma\|Wy\|,$$

giving $\|Wy\| \geq |\lambda/\alpha + 1 - \lambda|\|y\|/(1 + \sqrt{k}\varsigma)$. Hence,

$$\sigma_{\min}(W) = \min_{\|y\|=1}\|Wy\| \geq \frac{\left|\frac{\lambda}{\alpha} + 1 - \lambda\right|}{1 + \sqrt{k}}\varsigma.$$

Therefore we have that

$$\mathrm{cond}(W) = \frac{\sigma_{\max}(W)}{\sigma_{\min}(W)} \leq \frac{\left(\sqrt{s}\varsigma + \left|\frac{\lambda}{\alpha} + 1 - \lambda\right|\right)(1 + \sqrt{k}\varsigma)}{\left|\frac{\lambda}{\alpha} + 1 - \lambda\right|} \leq \frac{\left(\max\left(\left|\frac{\lambda}{\alpha} + 1 - \lambda\right|, 1\right) + \sqrt{k}\varsigma\right)^2}{\left|\frac{\lambda}{\alpha} + 1 - \lambda\right|}.$$

The bound of $\mathrm{cond}(W)$ follows immediately by setting $b = \max(|\lambda/\alpha + 1 - \lambda|, 1)$. □

Proof of Theorem 6.1. Let Q_i and V_i be the matrices of the left and right singular vectors corresponding to the d largest singular values of $\overline{X}_i$ and let $\tilde{W}_i = \tilde{Q}_iR_i$ be a QR decomposition of $\tilde{W}_i$. The distance $\mathrm{dist}(\tilde{W}_i, \hat{V}_i)$ is defined by

$$\mathrm{dist}(\tilde{W}_i, \hat{V}_i) = \|\tilde{Q}_i^TV_i\|.$$

Notice that $G_iW_i = \overline{X}_i\tilde{W}_i = \overline{X}_i\tilde{Q}_iR_i$ and $Q_i^T\overline{X}_i = \Sigma_iV_i^T$. We have

$$Q_i^TG_iW_i = \Sigma_iV_i^T\tilde{Q}_iR_i.$$

Since $\tilde{W}_i$ is of full column rank, R_i is nonsingular. It follows that $V_i^T\tilde{Q}_i = \Sigma_i^{-1}Q_i^TG_iW_iR_i^{-1}$. Therefore,

$$\|V_i^T\tilde{Q}_i\| \leq \|\Sigma_i^{-1}\|\|G_iW_i\|\|R_i^{-1}\| = \frac{\|G_iW_i\|}{\sigma_d(\tilde{W}_i)\sigma_d(\overline{X}_i)},$$

completing the proof. □

References

[1] J. Anderson, Logistic Discrimination, North-Holland Publishing Company, 1982, pp. 169–191.
[2] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (6) (2003) 1373–1396.
[3] M. Brand, Charting a manifold, in: Advances in Neural Information Processing Systems, vol. 15, MIT Press, Cambridge, 2003.
[4] D. Donoho, C. Grimes, Hessian eigenmaps: new tools for nonlinear dimensionality reduction, Proceedings of National Academy of Science 100 (2003) 5591–5596.
[5] G.H. Golub, C.F. Van Loan, Matrix Computations, third ed., Johns Hopkins University Press, Baltimore, Maryland, 1996.
[6] J. Ham, D. Lee, S. Mika, B. Scholkopf, A kernel view of the dimensionality reduction of manifolds, in: International Conference on Machine Learning, vol. 21, 2004.

[9] http://vasc.ri.cmu.edu//idb/html/motion/hand/index.html.

[7] J. Hull, A database for handwritten text recognition research, IEEE Transaction on Pattern Analysis and Machine Intelligence 16 (5) (1994) 550–554.
[9] D. Mateus, F. Cuzzolin, R. Horaud, E. Boyer, Articulated shape matching using locally linear embedding and orthogonal alignment, in: International Conference on Computer Vision, 2007.
[10] S. Li, K. Chan, C. Wang, Performance evaluation of the nearest feature line method in image classification and retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (11) (2000) 1335–1339.
[11] Y. Pan, S. Ge, A. Mamuna, Weighted locally linear embedding for dimension reduction, Pattern Recognition 42 (5) (2009) 798–811.
[12] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[13] L. Saul, S. Roweis, Think globally, fit locally: unsupervised learning of nonlinear manifolds, Journal of Machine Learning Research 4 (2003) 119–155.
[14] K. Spackman, Signal detection theory: valuable tools for evaluating inductive learning, in: Proceedings of the Sixth International Workshop on Machine Learning, 1989, pp. 160–163.
[15] D. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligences 18 (8) (1996) 831–836.
[16] J. Tenenbaum, V. De Silva, J. Langford, A global geometric framework for nonlinear dimension reduction, Science 290 (2000) 2319–2323.
[17] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, N. Koudas, Non-linear dimensionality reduction techniques for classification and visualization, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.
[18] Z. Zhang, J. Wang, H. Zha, Adaptive manifold learning, in: L.K. Saul, Y. Weiss, L. Bottou, Advances in Neural Information Processing Systems, vol. 17, MIT Press, Cambridge, MA, 2005, pp. 1473–1480.
[19] H. Zha, Z. Zhang, Spectral properties of the alignment matrices in manifold learning, SIAM Review 51 (3) (2009) 454–566.
[20] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimensionality reduction via tangent space alignment, SIAM Journal on Scientific Computing 26 (1) (2005) 313–338.

About the Author—JING WANG received the Ph.D. in Department of Mathematics, ZheJiang University in 2006. He is now the associate professor of School of Computer Science and Technology, Huaqiao University. His research interests include manifold learning, data mining, and numerical linear algebra.

About the Author—ZHENYUE ZHANG received the Ph.D. in Department of Mathematics, Fudan University in 1989. He is now the professor of Department of Mathematics, Zhejiang University. His research interests include manifold learning, numerical linear algebra, and scientific computing.