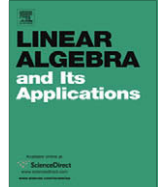




Contents lists available at ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa



Matrix perturbation analysis of local tangent space alignment[☆]

Xiaoming Huo^{*}, Andrew K. Smith

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA

ARTICLE INFO

Article history:

Received 21 August 2007

Accepted 12 September 2008

Available online 22 October 2008

Submitted by R.A. Brualdi

AMS classification:

15A60

65F99

Keywords:

Manifold learning

Dimension reduction

Local tangent space alignment (LTSA)

Matrix perturbation

Perturbation analysis

ABSTRACT

We consider the performance of local tangent space alignment [Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, *SIAM J. Sci. Comput.* 26 (1) (2004) 313–338], one of several manifold learning algorithms, which have been proposed as a dimension reduction method. Matrix perturbation theory is applied to obtain a worst-case upper bound on the angle between the computed linear invariant subspace and the linear invariant subspace that is associated with the embedded intrinsic parametrization. Our result is the first performance bound that has been derived.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

Manifold-based dimensionality reduction methods have attracted substantial attention in both the machine learning and statistics communities, mostly due to their demonstrated potential. Though many methods have been proposed, little work has been done to analyze the performance of these methods. The main contribution of this paper is to establish some asymptotic performance properties of a manifold learning algorithm, as well as a demonstration of some of its limitations. The key idea in our analysis is to treat the solutions of manifold learning algorithms as invariant subspaces, and then carry

[☆] This project is partially supported by NSF Grants 0604736 and 0700152.

^{*} Corresponding author.

E-mail addresses: xiaoming@isye.gatech.edu (X. Huo), asmith@isye.gatech.edu (A.K. Smith).

out a matrix perturbation analysis. A common feature of several manifold learning algorithms (e.g. [2,3,4,5,1]) is that their solutions correspond to invariant subspaces, typically the eigenspace associated with the smallest eigenvalues of a kernel matrix. The exact form of this kernel matrix, of course, depends on the details of the particular algorithm. These subspaces, however, are clearly invariant regardless of the exact form of the matrix involved, because they are spanned by eigenvectors [6, Section I.3.4].

Many efficient manifold-learning algorithms have been developed. A partial list of them is: locally linear embedding (LLE) [2], ISOMAP [7], charting [5], local tangent space alignment (LTSA) [1], Laplacian eigenmaps [3], and Hessian eigenmaps [4], etc. LTSA, in particular, enjoys several advantages. First of all, in numerical simulation (e.g., using the tools offered by [8]), we find empirically that LTSA performs among the best of the available algorithms. Second, the solution to each step of the LTSA algorithm is an invariant subspace, which makes analysis of its performance more tractable. Third, the similarity between LTSA and several other manifold learning algorithms (e.g., LLE, Laplacian eigenmaps and Hessian eigenmaps) suggests that our results may generalize. Thus, it is our hope that this performance analysis will provide a theoretical foundation for the application of manifold learning algorithms. Our main theoretical result is Theorem 3.8, which is a worst-case upper bound on the angle between the subspaces spanned by the computed coordinates and by the intrinsic parameters.

The rest of the paper is organized as follows. The problem formulation and background information are presented in Section 2. In Section 3, perturbation analysis is carried out, and the main theorem is proved. In Section 4, more simulation results are presented to illustrate the analytical properties. Some discussion related to existing work in this area is included in Section 5. Some concluding remarks are in Section 6. Technical proofs are relegated to Appendix wherever convenient.

2. Problem statement and illustration

2.1. Model

To be more specific, we formulate our dimension reduction problem as follows. For a positive integer n , let $y_i \in \mathbb{R}^D, i = 1, 2, \dots, n$, denote n observations. We assume that there is a mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}^D$ which satisfies a set of regularity conditions. In addition, we require another set of (possibly multivariate) values $x_i \in \mathbb{R}^d, d < D, i = 1, 2, \dots, n$, such that

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \dots, n, \tag{1}$$

where $\varepsilon_i \in \mathbb{R}^D$ denotes a random error. For example, we may assume $\varepsilon_i \sim N(\vec{0}, \sigma^2 I_D)$; i.e., a multivariate normal distribution with mean zero and variance-covariance proportional to the identity matrix. The central questions of dimension reduction are: (i) Can we find a set of low-dimensional vectors such that (1) holds? (ii) What kind of regularity conditions should be imposed on f ? (iii) Is the model well defined? These questions will be answered in the following.

2.2. A pedagogical example

An illustrative example of dimension reduction that makes our formulation more concrete is given in Fig. 1. Fig. 1a shows the true underlying structure of a toy example, a 1-D spiral. The *noiseless* observations are equally spaced points on this spiral. In Fig. 1b, 1024 *noisy* observations are generated with multivariate noise satisfying $\varepsilon_i \sim N(\vec{0}, \frac{1}{100} I_3)$. We then apply LTSA to the noisy observations, using $k = 10$ nearest neighbors. In Fig. 1c, the result from LTSA is compared with the true parametrization. When the underlying parameter is faithfully recovered, one should see a straight line, which is observed in Fig. 1c.

2.3. Regularity and uniqueness of the mapping f

If the conditions on the mapping f are too general, the model (1) is not well defined. For example, if the mapping $f(\cdot)$ and point set $\{x_i\}$ satisfy (1), so do $f(A^{-1}(\cdot - b))$ and $\{Ax_i + b\}$, where A is an invertible d by d matrix and b is a d -dimensional vector. As being common in the manifold-learning literature, we adopt the following condition on f .

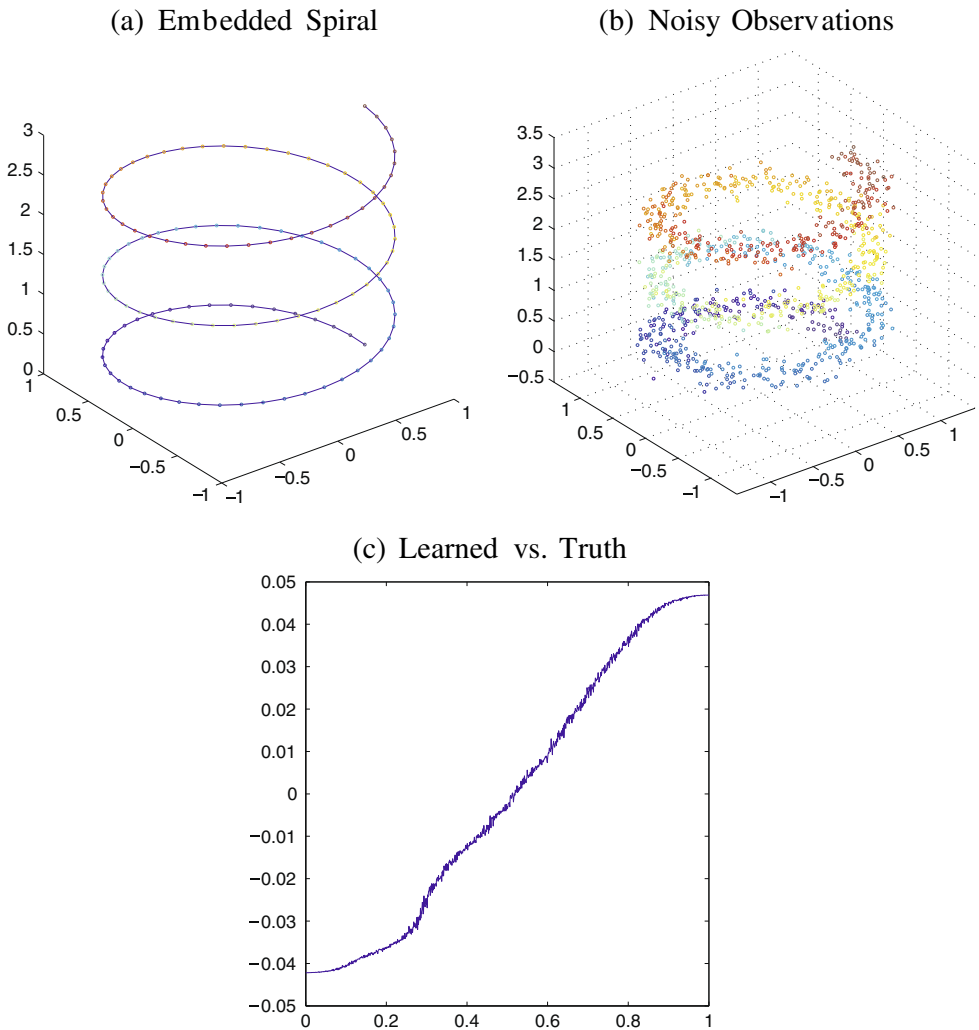


Fig. 1. An illustrative example of LTSA in nonparametric dimension reduction. The straight line pattern in (c) indicates that the underlying parametrization has been approximately recovered.

Condition 2.1 (Local isometry). *The mapping f is locally isometric: For any $\varepsilon > 0$ and x in the domain of f , let $N_\varepsilon(x) = \{z : \|z - x\|_2 < \varepsilon\}$ denote an ε -neighborhood of x using Euclidean distance. We have*

$$\|f(x) - f(x_0)\|_2 = \|x - x_0\|_2 + o(\|x - x_0\|_2).$$

The above condition indicates that in a local sense, f preserves Euclidean distance. Let $J(f; x_0)$ denote the Jacobian of f at x_0 . We have $J(f; x_0) \in \mathbb{R}^{D \times d}$, where each column (resp., row) of $J(f; x_0)$ corresponds to a coordinate in the feature (resp., data) space. The above in fact implies the following lemma.

Lemma 2.2. *The matrix $J(f; x_0)$ is orthonormal for any x_0 , i.e.,*

$$J^T(f; x_0)J(f; x_0) = I_d.$$

A reference for this result is Zhang and Zha [1].

Given the previous condition, model (1) is still not uniquely defined. For example, for any d by d orthogonal matrix O and any d -dimensional vector b , if $f(\cdot)$ and $\{x_i\}$ satisfy (1) and Condition 2.1, so do $f(O^T(\cdot - b))$ and $\{Ox_i + b\}$. We can force b to be $\bar{0}$ by imposing the condition that $\sum_i x_i = 0$. In dimension reduction, we can consider the sets $\{x_i\}$ and $\{Ox_i\}$ “invariant,” because one is just a rotation of the other. In fact, the invariance coincides with the concept of “invariant subspace” that will be discussed later.

Condition 2.3 (Local linear independence condition). Let $Y_i \in \mathbb{R}^{D \times k}$, $1 \leq i \leq n$, denote a matrix whose columns are made by the i th observation y_i and its $k - 1$ nearest neighbors. We choose $k - 1$ neighbors so that the matrix Y_i has k columns. It is generally assumed that $d < k$. For any $1 \leq i \leq n$, the rank of $Y_i \bar{P}_k$ is at least d ; in other words, the d th largest singular value of matrix $Y_i \bar{P}_k$ is greater than 0.

The regularity of the manifold can be determined by the Hessians of the mapping. Rewrite $f(x)$ for $x \in \mathbb{R}^d$ as

$$f(x) = (f_1(x), f_2(x), \dots, f_D(x))^T.$$

Furthermore, let $x = (x_1, \dots, x_d)^T$. A Hessian is

$$[H_i(f; x)]_{jk} = \frac{\partial^2 f_i(x)}{\partial x_j \partial x_k}$$

for $1 \leq i \leq D, 1 \leq j, k \leq d$.

The following condition ensures that f is locally smooth. We impose a bound on all the components of the Hessians.

Condition 2.4 (Regularity of the manifold). $|[H_i(f; x)]_{jk}| \leq C_1$ for all i, j , and k , where $C_1 > 0$ is a prescribed constant.

2.4. Solutions as invariant subspaces and a related metric

We now give a more detailed discussion of invariant subspaces. Let $\mathcal{R}(X)$ denote the subspace spanned by the columns of X . Recall that $x_i, i = 1, 2, \dots, n$, are the true low-dimensional representations of the observations. We treat the x_i 's as column vectors. Let

$$X = (x_1, x_2, \dots, x_n)^T;$$

i.e., the i th row of X corresponds to $x_i, 1 \leq i \leq n$. If the set $\{Ox_i\}$, where O is a d by d orthogonal square matrix, forms another solution to the dimension reduction problem, we have

$$(Ox_1, Ox_2, \dots, Ox_n)^T = XO^T.$$

It is evident that $\mathcal{R}(XO^T) = \mathcal{R}(X)$. This justifies the *invariance* that was mentioned earlier.

The goal of our performance analysis is to answer the following question: Letting $\|\tan(\cdot, \cdot)\|_2$ denote the Euclidean norm of the vector of canonical angles between two invariant subspaces [6, Section I.5], and letting X and \tilde{X} denote the true and estimated parameters, respectively, how do we evaluate $\|\tan(\mathcal{R}(X), \mathcal{R}(\tilde{X}))\|_2$?

2.5. LTSA: local tangent space alignment

We now review LTSA. There are two main steps in the LTSA algorithm [1].

- (1) The first step is to compute the local representation on the manifold. Consider a projection matrix $\bar{P}_k = I_k - \frac{1}{k} \cdot \mathbf{1}_k \mathbf{1}_k^T$, where I_k is the k by k identity matrix and $\mathbf{1}_k$ is a k -dimensional column vector of ones. It is easy to verify that $\bar{P}_k = \bar{P}_k \cdot \bar{P}_k$, which is a characteristic of projection matrices.

We solve the minimization problem:

$$\min_{A,V} \|Y_i \bar{P}_k - AV\|_F,$$

where $A \in \mathbb{R}^{D \times d}, V \in \mathbb{R}^{d \times k}$, and $VV^T = I_d$. Let V_i denote optimal V . Then the row vectors of V_i are the d right singular vectors of $Y_i \bar{P}_k$.

(2) The solution to LTSA corresponds to the invariant subspace which is spanned and determined by the eigenvectors associated with the 2nd to the $(d + 1)$ st smallest eigenvalues of the matrix

$$(S_1, \dots, S_n) \begin{pmatrix} \bar{P}_k - V_1^T V_1 & & & \\ & \bar{P}_k - V_2^T V_2 & & \\ & & \ddots & \\ & & & \bar{P}_k - V_n^T V_n \end{pmatrix} (S_1, \dots, S_n)^T.$$

where $S_i \in \mathbb{R}^{n \times k}$ is a selection matrix such that $Y^T S_i = Y_i$, where $Y = (y_1, y_2, \dots, y_n)^T$.

We have slightly reformulated the original algorithm as presented in [1], in order to simplify the theoretical analysis. The verification of the equivalence is a standard exercise in linear algebra, and it is given in the Appendix of [9].

3. Perturbation analysis

We now carry out a perturbation analysis on the reformulated version of LTSA. There are two steps in our analysis: in the *local* step (Section 3.1), we characterize the deviation of the null spaces of the matrices $\bar{P}_k - V_i^T V_i, i = 1, 2, \dots, n$. In the *global* step (Section 3.2), we derive the variation of the null space under global alignment. The detailed calculations are again relegated to the Appendix.

3.1. Local coordinates

Let X be the matrix of true parameters. We define

$$X_i = X^T S_i = (x_1, x_2, \dots, x_n) S_i;$$

i.e., the columns of X_i are made by x_i and those x_j 's that correspond to the $k - 1$ nearest neighbors of y_i . We require a bound on the size of the local neighborhoods defined by the X_i 's.

Condition 3.1 (Universal bound on the sizes of neighborhoods). For all $i, 1 \leq i \leq n$, we have $\tau_i < \tau$, where τ is a prescribed constant and τ_i is an upper bound on the distance between two columns of X_i : $\tau_i = \max_{x_j, x_k} \|x_j - x_k\|$, where the maximum is taken over all columns of X_i .

In this paper, we are interested in the case when $\tau \rightarrow 0$.

We will need conditions on the local tangent spaces. Let $d_{\min,i}$ (respectively, $d_{\max,i}$) denote the minimum (respectively, maximum) singular values of $X_i \bar{P}_k$. Let

$$d_{\min} = \min_{1 \leq i \leq n} d_{\min,i}$$

and

$$d_{\max} = \max_{1 \leq i \leq n} d_{\max,i}.$$

We have the following result regarding d_{\max} :

Lemma 3.2

$$d_{\min} \leq d_{\max} \leq \tau \sqrt{k}.$$

For the proof, see Appendix A.1.

Condition 3.3 (Local tangent space). *There exists a constant $C_2 > 0$, such that*

$$C_2 \cdot \tau \leq d_{\min}.$$

The above can roughly be thought of as requiring that the local dimension of the manifold remain constant (i.e., the manifold has no singularities.)

The following condition defines a global bound on the errors (ε_i).

Condition 3.4 (Universal error bound). *There exists $\sigma > 0$, such that $\forall i, 1 \leq i \leq n$, we have $\|y_i - f(x_i)\|_\infty < \sigma$. Moreover, we assume $\sigma = o(\tau)$; i.e., we have $\frac{\sigma}{\tau} \rightarrow 0$, as $\tau \rightarrow 0$.*

It is reasonable to require that the error bound (σ) be smaller than the size of the neighborhood (τ), which is reflected in the above condition. We discuss the necessity of this condition in Section 3.3.

Within each neighborhood, we give a perturbation bound between an invariant subspace spanned by the true parametrization and the invariant subspace spanned by the singular vectors of the matrix of noisy observations. Let

$$X_i \bar{P}_k = A_i D_i B_i$$

be the singular value decomposition of the matrix $X_i \bar{P}_k$; here $A_i \in \mathbb{R}^{d \times d}$ is orthogonal ($A_i A_i^T = I_d$), $D_i \in \mathbb{R}^{d \times d}$ is diagonal, and the rows of $B_i \in \mathbb{R}^{d \times k}$ are the right singular vectors corresponding to the largest singular values ($B_i B_i^T = I_d$). It is not hard to verify that

$$B_i = B_i \bar{P}_k. \tag{2}$$

Let $Y_i \bar{P}_k = \tilde{A}_i \tilde{D}_i \tilde{B}_i$ be the singular value decomposition of $Y_i \bar{P}_k$, and assume that this is the “thin” decomposition of rank d . We may think of this as the perturbed version of $J(f; x_i^{(0)}) X_i \bar{P}_k$. The rows of \tilde{B}_i are the eigenvectors of $(Y_i \bar{P}_k)^T (Y_i \bar{P}_k)$ corresponding to the d largest eigenvalues. Let $\mathcal{R}(B_i^T)$ (respectively, $\mathcal{R}(\tilde{B}_i^T)$) denote the invariant subspace that is spanned by the columns of matrix B_i^T (respectively, \tilde{B}_i^T).

Theorem 3.5. *Given invariant subspaces $\mathcal{R}(B_i^T)$ and $\mathcal{R}(\tilde{B}_i^T)$ as defined above, we have*

$$\lim_{\tau \rightarrow 0} \|\sin(\mathcal{R}(B_i^T), \mathcal{R}(\tilde{B}_i^T))\|_2 \leq C_3 \left(\frac{\sigma}{\tau} + C_1 \tau \right),$$

where C_3 is a constant that depends on k, D and C_2 .

The proof is presented in Appendix A.2. The above gives an upper bound on the deviation of the local invariant subspace in step (1) of the modified LTSA. It will be used later to prove a global result.

3.2. Global alignment

Condition 3.6 (No overuse of one observation). *There exists a constant C_4 , such that*

$$\left\| \sum_{i=1}^n S_i \right\|_\infty \leq C_4.$$

Note that we must have $C_4 \geq k$. The next condition (Condition 3.7) will implicitly give an upper bound on C_4 .

Recall that the quantity $\|\sum_{i=1}^n S_i\|_\infty$ is the maximum row sum of the absolute values of the entries in $\sum_{i=1}^n S_i$. The value of $\|\sum_{i=1}^n S_i\|_\infty$ is equal to the maximum number of nearest neighbor subsets to which a single observation belongs.

We will derive an upper bound on the angle between the invariant subspace spanned by the result of LTSA and the space spanned by the true parameters.

Given (2), it can be shown that

$$X_i \bar{P}_k (\bar{P}_k - B_i^T B_i) (X_i \bar{P}_k)^T = 0.$$

Recall $X = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^{n \times d}$. It is not hard to verify that the row vectors of

$$(\mathbf{1}_n, X)^T \tag{3}$$

span the $(d + 1)$ -dimensional null space of the matrix:

$$(S_1, \dots, S_n) \bar{P}_k \begin{pmatrix} I - B_1^T B_1 & & & \\ & I - B_2^T B_2 & & \\ & & \ddots & \\ & & & I - B_n^T B_n \end{pmatrix} \bar{P}_k (S_1, \dots, S_n)^T. \tag{4}$$

Assume that

$$\begin{pmatrix} \mathbf{1}_n^T \\ \sqrt{n} \\ X^T \\ (X^c)^T \end{pmatrix}$$

is orthogonal, where $X^c \in \mathbb{R}^{n \times (n-1-d)}$. Although in our original problem formulation, we made no assumptions about the x_i 's, we can still assume that the columns of X are orthonormal because we can transform any set of x_i 's into an orthonormal set by rescaling the columns and multiplying by an orthogonal matrix. Based on the previous paragraph, we have

$$\begin{pmatrix} \mathbf{1}_n^T \\ \sqrt{n} \\ X^T \\ (X^c)^T \end{pmatrix} M_n \begin{pmatrix} \mathbf{1}_n \\ \sqrt{n} \\ X \\ X^c \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{(d+1) \times (d+1)} & \mathbf{0}_{(d+1) \times (n-d-1)} \\ \mathbf{0}_{(n-d-1) \times (d+1)} & L_2 \end{pmatrix}$$

where

$$M_n = (S_1, \dots, S_n) \bar{P}_k \begin{pmatrix} I_k - B_1^T B_1 & & \\ & \ddots & \\ & & I_k - B_n^T B_n \end{pmatrix} \bar{P}_k (S_1, \dots, S_n)^T$$

and

$$L_2 = (X^c)^T M_n X^c.$$

Let ℓ_{\min} denote the minimum singular value (i.e., eigenvalue) of L_2 . We will need the following condition on ℓ_{\min} .

Condition 3.7 (Appropriateness of global dimension). $\ell_{\min} > 0$ and ℓ_{\min} goes to 0 at a slower rate than $\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau$; i.e., as $\tau \rightarrow 0$, we have

$$\frac{\left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau\right) \cdot \|\sum_{i=1}^n S_i\|_\infty}{\ell_{\min}} \rightarrow 0.$$

As discussed in [10], this condition is actually related to the amount of overlap between the nearest neighbor sets.

Theorem 3.8 (Main theorem)

$$\lim_{\tau \rightarrow 0} \|\tan(\mathcal{R}(\tilde{X}), \mathcal{R}(X))\|_2 \leq \frac{C_3 \left(\frac{\sigma}{\tau} + C_1 \tau\right) \cdot \|\sum_{i=1}^n S_i\|_\infty}{\ell_{\min}}. \tag{5}$$

As mentioned in the Introduction, the above theorem gives a worst-case bound on the performance of LTSA. A discussion on when Condition 3.7 is satisfied will be long and beyond the scope of this paper. We leave it to future investigation.

3.3. The requirement that $\sigma \rightarrow 0$

A natural question to ask, in light of the above analysis, is whether LTSA is still consistent without the restrictive assumption that $\sigma \rightarrow 0$. In this section, we discuss a simple example which demonstrates that the answer is, surprisingly, no.

Consider the following model:

$$y_i = \begin{pmatrix} x_i \\ 0 \end{pmatrix} + \epsilon_i,$$

where

$$x_i \sim N(0, \sigma_x^2), \quad \epsilon_i \sim MVN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_\epsilon^2 I_2 \right], \quad x_i \perp \epsilon_i.$$

It is then easy to see that

$$y_i \sim MVN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 + \sigma_x^2 & 0 \\ 0 & \sigma_\epsilon^2 \end{pmatrix} \right].$$

Suppose, as usual, that we wish to reconstruct the x_i 's from the given y_i 's. This is a particularly simple case of dimension reduction, where $D = 2, d = 1$, and the data lie near a linear manifold. Thus, the entire manifold may be thought of as a single linear patch. In applying LTSA to this model, we may therefore assume that $k = n$, that is, that all points in the data set are neighbors of one another. This implies that $S_i = I_n$ for each i .

Now, in the first step, LTSA will find the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix. It is a standard result [11, Section 13.5] that the space spanned by the leading eigenvector converges to the space spanned by $(1, 0)^T$. Without loss of generality, we may suppose that the eigenvector is chosen so that the first component is positive, and therefore the leading eigenvector converges to $(1, 0)^T$. The estimated local coordinates will then be

$$\theta_i = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} y_i = x_i + \epsilon_i^{(1)},$$

where we have denoted the first component of ϵ_i by $\epsilon_i^{(1)}$. Let $\Theta = (\theta_1, \theta_2, \dots, \theta_n)$ denote the row vector formed by the n estimated local coordinates. We assume that $\|\Theta\|_2 = 1$, that is, that the eigenvector associated with the largest eigenvalue of the sample covariance matrix is normalized. Also note that we have $\Theta \perp \mathbf{1}_n$.

The alignment step is especially simple due to the structure of our artificial example. The computed \hat{x}_i 's are given by the eigenvector corresponding to the smallest eigenvalue of

$$(S_1, S_2, \dots, S_n) \bar{P}_{k \times n} \begin{pmatrix} I_k - \Theta_1^+ \Theta_1 & & & \\ & I_k - \Theta_2^+ \Theta_2 & & \\ & & \ddots & \\ & & & I_k - \Theta_n^+ \Theta_n \end{pmatrix} \cdot \bar{P}_{k \times n} (S_1, S_2, \dots, S_n)^T,$$

The computation is easily simplified, however. As noted above, each S_i is the identity, and the diagonal blocks in the center matrix are all the same. Therefore, the \hat{x} 's can be expressed as the eigenvector corresponding to the second smallest eigenvalue of

$$\bar{P}_n (I_n - \Theta^+ \Theta) \bar{P}_n.$$

It is easy to see that the correct eigenvector is proportional to $\Theta^+ = \Theta^T$ by noting that

$$(I_n - \Theta^+ \Theta) \Theta^+ = \Theta^+ - \Theta^+ \Theta \Theta^+ = 0.$$

Therefore, the vector Θ^+ corresponds to the eigenvalue 0 of $\bar{P}_n(I_n - \Theta^+ \Theta) \bar{P}_n$. Further, we know that the dimension of the nullspace of $\bar{P}_n(I_n - \Theta^+ \Theta) \bar{P}_n$ is exactly $d + 1 = 2$, so there can be no other vector in the nullspace except, of course, for $\mathbf{1}_n$. If σ_ϵ^2 is constant, then in general we will have

$$X = (x_1, x_2, \dots, x_n)^T$$

but

$$\tilde{X} = (x_1 + \epsilon_1^{(1)}, x_2 + \epsilon_2^{(1)}, \dots, x_n + \epsilon_n^{(1)})^T.$$

Now, we consider the angle formed between the two subspaces $\mathcal{R}(X)$ and $\mathcal{R}(\tilde{X})$. In this special one-dimensional case, this has a particularly simple form:

$$\angle(\mathcal{R}(X), \mathcal{R}(\tilde{X})) = \cos^{-1} \left(\frac{X^T \tilde{X}}{\|X\| \cdot \|\tilde{X}\|} \right).$$

Supposing that n is sufficiently large, we may use the strong law of large numbers to evaluate the limits of the quantities on the right-hand side. For the numerator of the fraction, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{X^T \tilde{X}}{n} &= \frac{1}{n} \lim_{n \rightarrow \infty} \sum_{i=1}^n x_i \cdot (x_i + \epsilon_i^{(1)}) \\ &\stackrel{\text{SLLN}}{=} E(x_i \cdot (x_i + \epsilon_i^{(1)})) \\ &= E(x_i^2) + E(x_i) \cdot E(\epsilon_i^{(1)}) \\ &= \sigma_x^2. \end{aligned}$$

For the denominator, a similar argument shows that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \|X\|^2 = \sigma_x^2 \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{1}{n} \|\tilde{X}\|^2 = \sigma_x^2 + \sigma_\epsilon^2.$$

Putting these limits together, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \angle(\mathcal{R}(X), \mathcal{R}(\tilde{X})) &= \cos^{-1} \left(\frac{n\sigma_x^2}{\sqrt{n\sigma_x^2} \cdot \sqrt{n(\sigma_x^2 + \sigma_\epsilon^2)}} \right) \\ &= \cos^{-1} \left(\frac{\sigma_x^2}{\sqrt{\sigma_x^4 + \sigma_x^2 \sigma_\epsilon^2}} \right). \end{aligned}$$

If σ_ϵ^2 is constant (i.e., does not have limit 0,) then the argument of \cos^{-1} will not have limit 1, and $\lim_{n \rightarrow \infty} \angle(\mathcal{R}(X), \mathcal{R}(\tilde{X})) \neq 0$. Note that this inconsistency would still apply even if we add the stronger assumption that the distribution of the error is bounded.

Thus, we see that our assumption that $\sigma \rightarrow 0$ is, in fact, necessary to ensure the consistency of LTSA, even in what might be considered the simplest possible case of the dimension reduction problem. While this result may at first seem somewhat counterintuitive, it is less surprising when one considers the fact that the number of unknown parameters (in this case, the x_i 's) grows as n increases, so our dimension reduction problem is not analogous to traditional parameter estimation problems such as the classical model

$$y_i = \mu + \epsilon_i$$

with only μ (1 parameter) unknown.

An additional difficulty which arises in the absence of the assumption that $\sigma \rightarrow 0$ is the fact that the estimated selection matrices (the \hat{S}_i 's) may not converge to the correct population counterparts (the S_i 's). An implicit assumption throughout our analysis is that, at least asymptotically $\hat{S}_i = S_i$, which is crucial in our derivation of bounds on the deviation of the estimated alignment matrix from the true alignment matrix. In the asymptotically noiseless case, this convergence is automatic, provided

that the underlying manifold is not self-intersecting. However, in the asymptotic case with noise, such convergence is not guaranteed and in fact, will not hold in general. This is compounded with the difficulties discussed above related to our toy example.

Considering the problem from a geometric perspective is also illuminating. While it is well-known that we can asymptotically recover the correct local tangent space at each point (at least in our simplified example), the problem occurs in the alignment step. The simple structure of the example makes it easy to see what is going on – we extract the *projection* of y_i onto the local tangent space. However, this of course does not correspond to the generating coordinate x_i in the general case, though it could be construed as a maximum likelihood estimate of the generating coordinate, being the closest point in the transformed parameter space to the actual observation in terms of Euclidean distance. What we can recover, then, is the projection of $f(x_i) + \epsilon_i$ onto the (asymptotically correct) tangent space in each neighborhood, but the original generating coordinate itself is unrecoverable.

It would certainly be interesting to know whether f can still be recovered asymptotically if σ is constant, but this question remains open. The analysis of the reconstruction of f is more complicated because LTSA does not compute any function explicitly – an estimated f can only be computed implicitly, for example by polynomial regression of Y on \tilde{X} as discussed in Section 5 of [1]. An analysis of this situation would involve consideration of the interplay of the errors in \tilde{X} with the errors in reconstructing the function f via indirect methods based on \tilde{X} . We leave this to future investigation.

A further consequence of this result is that while plots such as those shown in Fig. 2 can be useful as rough indicators of LTSA’s performance, they are not reliable in a strict sense for determining consistency. Although the relationship between the true and estimated coordinates may appear to be roughly linear, this alone does not imply that the algorithm will asymptotically recover the correct coordinates – the trouble is the “bandwidth” of the graph. If the underlying parameters are truly recovered, the graph must eventually converge to *exactly* a straight line with no dispersion. Such information is difficult to discern from plots of this type.

4. Simulations

In the same setting as in Section 2.2, if we change the value of σ from $\sigma = 0.1$ to $\sigma = 0.025$ and 0.2 , we have Fig. 2. Based on our theorem, the smaller the error standard deviation is, the closer the result of LTSA is to the true parametrization. In the case of $\sigma = 0.2$, the result of LTSA breaks down.

When X and \tilde{X} are one-dimensional, we have

$$\sqrt{1 - [\text{corr}(X, \tilde{X})]^2} = \|\sin(\mathcal{R}(X), \mathcal{R}(\tilde{X}))\|_2 \leq \|\tan(\mathcal{R}(X), \mathcal{R}(\tilde{X}))\|_2,$$

where $\text{corr}(X, \tilde{X})$ is the correlation coefficient between two vectors. If

$$\|\tan(\mathcal{R}(X), \mathcal{R}(\tilde{X}))\|_2 \rightarrow 0,$$

we have $\text{corr}(X, \tilde{X}) \rightarrow 1$, which corresponds to the consistency.

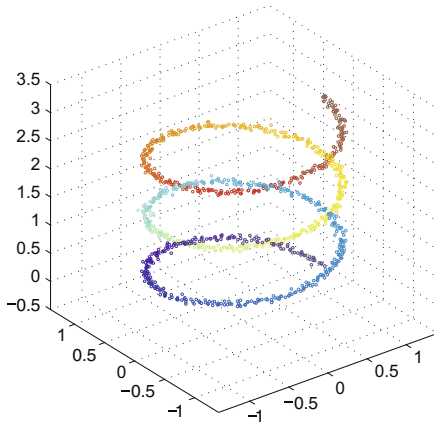
In Fig. 2b, when σ is small, we observe a nearly straight line; while in Fig. 2d, where σ is large, the estimates are drastically different from what they are supposed to be. This phenomenon is consistent with our theory.

5. Discussion

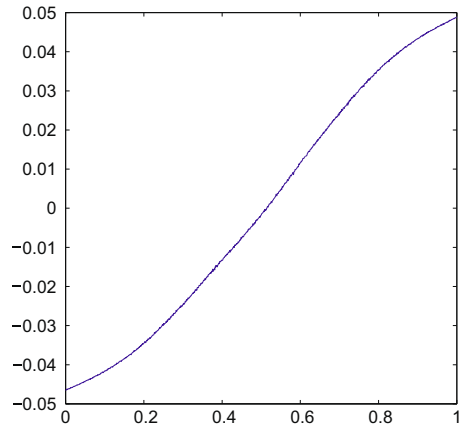
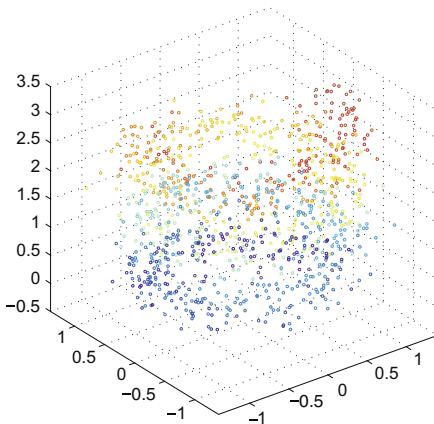
To the best of our knowledge, the performance analysis that is based on invariant subspaces is new. Consequently the worst-case upper bound is the first of its kind. There are still open questions to be addressed (Section 5.1). In addition to a discussion on the relation of LTSA to existing dimension reduction methodologies, we will also address relation with known results as well (Section 5.2).

5.1. Open questions

The rate of convergence of ℓ_{\min} is determined by the topological structure of f . It is important to estimate this rate of convergence, but this issue has not been addressed here.

(a) Noisy Observations when $\sigma = 0.025$ 

(b) Result of LTSA

(c) Noisy Observations when $\sigma = 0.2$ 

(d) Result of LTSA

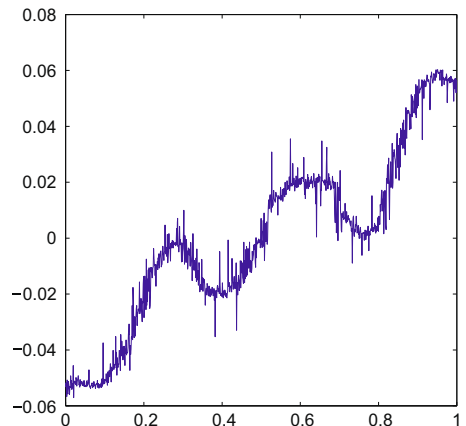


Fig. 2. Reruns of the illustrative example in Section 2.2, with different noise standard deviations.

We assume that $\tau \rightarrow 0$. One can imagine that it is true when the error bound (σ) goes to 0 and when the x_i 's are sampled with a sufficient density in the support of f . An open problem is how to derive the rate of convergence of $\tau \rightarrow 0$ as a function of the topology of f and the sampling scheme. After doing so, we may be able to decide where our theorem is applicable.

Given a covering scheme, such as choosing the k -nearest neighbors, a verification of $\tau \rightarrow 0$ and a derivation of its corresponding rate is an open question, too. The answer to this will depend on the topology of f , which is not covered in this paper, and the sampling scheme.

5.2. Relation to existing work

The error analysis in the original paper about LTSA is the closest to our result. However, Zhang and Zha [1] do not interpret their solutions as invariant subspaces, and hence their analysis does not yield a worst case bound as we have derived here.

Reviewing the original papers on LLE [2], Laplacian eigenmaps [3], and Hessian eigenmaps [4] reveals that their solutions are subspaces spanned by a specific set of eigenvectors. This naturally suggests that results analogous to ours may be derivable as well for these algorithms. A recent book

chapter [12] stresses this point. After deriving corresponding upper bounds, we can establish different proofs of consistency than those presented in these papers.

ISOMAP, another popular manifold learning algorithm, is an exception. Its solution cannot immediately be rendered as an invariant subspace. However, ISOMAP calls for MDS, which can be associated with an invariant subspace; one may derive an analytical result through this route.

6. Conclusion

We derive an upper bound of the distance between two invariant subspaces that are associated with the numerical output of LTSA and an assumed intrinsic parametrization. Such a bound describes the performance of LTSA with errors in the observations, and thus creates a theoretical foundation for its use in real-world applications, in which we would naturally expect such errors to be present. Our results can also be used to show other desirable properties, including consistency. Similar bounds may be derivable for other machine learning algorithms.

Appendix A. Proofs

A.1. Proof of Lemma 3.2

The first inequality in the lemma is obvious. For the second inequality, we have

$$d_{\max,i} = \|X_i \bar{P}_k\|_2 = \|(X_i - x_0 \cdot \mathbf{1}_k^T) \bar{P}_k\|_2 \leq \|X_i - x_0 \cdot \mathbf{1}_k^T\|_2 \tag{A.1}$$

$$\leq \sqrt{k} \cdot \max_{j \in P_i} \|x_j - x_0\|_2 \tag{A.2}$$

$$\leq \sqrt{k} \cdot \tau.$$

Taking the maximum over i on both sides, we obtain the second inequality.

In the above, inequality (A.1) is true because in general, for two matrices A and B , we have $\|AB\|_2 \leq \|A\|_2 \cdot \|B\|_2$ [6, p. 69]. The inequality (A.2) is also standard linear algebra [6, p. 71].

A.2. Proof of Theorem 3.5

The following two equations will be used:

$$Y_i \bar{P}_k = (Y_i - f(x_i^{(0)}) \cdot \mathbf{1}_k^T) \bar{P}_k \tag{A.3}$$

and

$$X_i \bar{P}_k = (X_i - x_i^{(0)} \cdot \mathbf{1}_k^T) \bar{P}_k, \tag{A.4}$$

where $x_i^{(0)}$ is the coordinates of the i th point in the true parametrization of the underlying manifold; cf. formulation in (1). The above equations can easily be verified by recalling the definition of \bar{P}_k .

To exploit the local isometry, we consider the Taylor expansion at $x_i^{(0)}$. It is not hard to verify the following: for $j \in P_i$, $1 \leq i \leq n$,

$$\begin{aligned} & \|y_j - f(x_i^{(0)}) - J(f; x_i^{(0)})(x_j - x_i^{(0)})\|_\infty \\ & \leq \|y_j - f(x_j)\|_\infty + \|f(x_j) - f(x_i^{(0)}) - J(f; x_i^{(0)})(x_j - x_i^{(0)})\|_\infty \\ & \leq \sigma + \frac{1}{2} C_1 \|x_j - x_i^{(0)}\|_2^2 + O(\|x_j - x_i^{(0)}\|_2^3) \\ & \leq \sigma + \frac{1}{2} C_1 \tau^2. \end{aligned}$$

Note that in the last step, we dropped an $O(\tau^3)$ term because we are only interested in the case when $\tau \rightarrow 0$, in which case the quadratic term dominates.

Let $E_i = Y_i \bar{P}_k - J(f; x_i^{(0)}) X_i \bar{P}_k$. Note that $E_i \in \mathbb{R}^{D \times k}$. We have the following upper bound for $\|E_i\|_2$:

$$\begin{aligned} \|E_i\|_2 &= \|Y_i \bar{P}_k - J(f; x_i^{(0)}) X_i \bar{P}_k\|_2 \\ &\leq \sqrt{k} \cdot \sup_{j \in P_i} \|(y_j - f(x_i^{(0)})) \cdot \bar{P}_k - J(f; x_i^{(0)})(x_j - x_i^{(0)}) \cdot \bar{P}_k\|_2 \\ &\leq \sqrt{k} \cdot \sup_j \|(y_j - f(x_i^{(0)})) - J(f; x_i^{(0)})(x_j - x_i^{(0)})\|_2 \\ &\leq \sqrt{kD} \cdot \sup_j \|(y_j - f(x_i^{(0)})) - J(f; x_i^{(0)})(x_j - x_i^{(0)})\|_\infty \\ &\leq \sqrt{kD} \cdot \left[\sigma + \frac{1}{2} C_1 \tau^2 \right]. \end{aligned} \tag{A.5}$$

In the above, the first and third inequalities are standard linear algebra, the second inequality is due to the fact that \bar{P}_k is a projection matrix.

We now wish to derive a bound on the angle between the subspaces spanned by the right singular vectors associated with the d largest singular values of $J(f; x_i^{(0)}) X_i \bar{P}_k$ and by those of $Y_i \bar{P}_k$. To this end, define the following two quantities:

$$\begin{aligned} R &= J(f; x_i^{(0)}) X_i \bar{P}_k \tilde{B}_i - \tilde{A}_i \tilde{D}_i, \\ S &= \bar{P}_k X_i^T J^T(f; x_i^{(0)}) \tilde{A}_i - \tilde{B}_i \tilde{D}_i. \end{aligned}$$

By substituting the identity $E_i = Y_i \bar{P}_k - J(f; x_i^{(0)}) X_i \bar{P}_k$, it is easy to see that $\|R\|_2 \leq \|E_i\|_2$ and $\|S\|_2 \leq \|E_i\|_2$. Finally, consider the smallest singular value of $Y_i \bar{P}_k$. We have

$$\begin{aligned} \sigma_{\min}(Y_i \bar{P}_k) &= \sigma_{\min}(J(f; x_i^{(0)}) X_i \bar{P}_k + E_i) \\ &\geq \sigma_{\min}(J(f; x_i^{(0)}) X_i \bar{P}_k) - \sigma_{\max}(E_i) \\ &\stackrel{(A.5)}{\geq} C_2 \cdot \tau - \sqrt{kD} \left[\sigma + \frac{1}{2} C_1 \tau^2 \right]. \end{aligned}$$

We can now apply Theorem V.4.4 in [6], and conclude

$$\begin{aligned} \|\sin(\mathcal{R}(B_i^T), \mathcal{R}(\tilde{B}_i^T))\|_2 &\leq \frac{\|E_i\|_2}{\sigma_{\min}(Y_i \bar{P}_k)} \\ &\leq \frac{\sqrt{kD} \cdot \left[\sigma + \frac{1}{2} C_1 \tau^2 \right]}{C_2 \cdot \tau - \sqrt{kD} \cdot \left[\sigma + \frac{1}{2} C_1 \tau^2 \right]} \end{aligned}$$

If we ignore higher-order terms, we can take $C_3 = \frac{\sqrt{kD}}{C_2}$, and the theorem is established.

A.3. Proof of Theorem 3.8

Now we consider the step of global alignment. Recall that the columns of $(\mathbf{1}_n, X)$, where X is defined in (??), are eigenvectors associated with the zero eigenvalue of (??).

First, similar to M_n , define \tilde{M}_n as

$$\tilde{M}_n = (S_1, \dots, S_n) \bar{P}_{k:n} \begin{pmatrix} I_k - \tilde{B}_1^T \tilde{B}_1 & & \\ & \ddots & \\ & & I_k - \tilde{B}_n^T \tilde{B}_n \end{pmatrix} \bar{P}_{k:n} (S_1, \dots, S_n)^T,$$

where \tilde{B}_i is defined right before Theorem 3.5.

We now consider $\|M_n - \tilde{M}_n\|_2$, the norm of the difference between the alignment matrices formed from the true and estimated local coordinates. This is equivalent to

$$\left\| \sum_{i=1}^n S_i \cdot \bar{P}_k \left(\tilde{B}_i^T \tilde{B}_i - B_i^T B_i \right) \bar{P}_k \cdot S_i^T \right\|_2.$$

Theorem 3.5 and [6, Theorem I.5.5] together imply that

$$\|\tilde{B}_i^T \tilde{B}_i - B_i^T B_i\|_2 \leq C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right), \quad i = 1, \dots, n.$$

Now, since $M_n - \tilde{M}_n$ is symmetric, we have $\|M_n - \tilde{M}_n\|_2 \leq \|M_n - \tilde{M}_n\|_1$. By Condition 3.6, each column of $M_n - \tilde{M}_n$ will be the sum of at most C_4 terms, each of which is a column of one of the matrices $B_i^T B_i - \tilde{B}_i^T \tilde{B}_i$. Therefore, we have

$$\|M_n - \tilde{M}_n\|_2 \leq C_4 \cdot C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right).$$

To verify the conditions of Theorem V.2.7 in [6], consider

$$\begin{pmatrix} \frac{\mathbf{1}_n^T}{\sqrt{n}} \\ X^T \\ (X^c)^T \end{pmatrix} (M_n - \tilde{M}_n) \begin{pmatrix} \mathbf{1}_n \\ \sqrt{n} \\ X \\ X^c \end{pmatrix} = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix},$$

where $E_{11} \in \mathbb{R}^{(d+1) \times (d+1)}$, $E_{12} \in \mathbb{R}^{(d+1) \times (n-d-1)}$, $E_{21} \in \mathbb{R}^{(n-d-1) \times (d+1)}$, and $E_{22} \in \mathbb{R}^{(n-d-1) \times (n-d-1)}$. Since we have assumed that

$$\begin{pmatrix} \frac{\mathbf{1}_n^T}{\sqrt{n}} \\ X^T \\ (X^c)^T \end{pmatrix}$$

is unitary, and since an upper bound on the spectral norm of a matrix is also an upper bound on the spectral norm of any submatrix, we have $\|E_{11}\|_2 \leq C_4 \cdot C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right)$, and similarly for all the other blocks.

It now easily follows that we can apply Theorem V.2.7 in [6], and therefore

$$\begin{aligned} \|\tan(\mathcal{R}(\tilde{X}), \mathcal{R}(X))\|_2 &\leq \frac{2\|E_{12}\|_2}{\ell_{\min} - \|E_{11}\|_2 - \|E_{22}\|_2} \\ &\leq \frac{C_4 \cdot C_3 \left(\frac{\sigma}{\tau} + C_1 \tau \right)}{\ell_{\min} - 2C_4 \cdot C_3 \left(\frac{\sigma}{\tau} + C_1 \tau \right)} \end{aligned}$$

Acknowledgements

The authors would like to thank Professors Hongyuan Zha and Zhenyue Zhang for helpful discussion.

References

[1] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, *SIAM J. Sci. Comput.*, 26 (1) (2004) 313–338.
 [2] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* 290 (2000) 2323–2326.
 [3] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Comput.* 15 (6) (2003) 1373–1396.
 [4] D.L. Donoho, C.E. Grimes, Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data, *Proc. Natl. Acad. Arts Sci.* 100 (2003) 5591–5596.
 [5] M. Brand, Charting a manifold, *Neural Information Processing Systems*, Mitsubishi Electric Research Labs, vol. 15, MIT Press, 2003.

- [6] G.W. Stewart, J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, MA, 1990.
- [7] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* 290 (2000) 2319–2323.
- [8] T. Wittman, MANifold learning Matlab demo, April, 2005, <<http://www.math.umn.edu/~wittman/mani/index.html>>.
- [9] X. Huo, A.K. Smith, Performance analysis of a manifold learning algorithm in dimension reduction, Tech. rep., Georgia Institute of Technology, March, 2006, <<http://www2.isye.gatech.edu/statistics/papers/06-06.pdf>>.
- [10] H. Zha, Z. Zhang, Spectral analysis of alignment in manifold learning, in: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.
- [11] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, third ed., John Wiley and Sons, Hoboken, 2003.
- [12] X. Huo, X.S. Ni, A.K. Smith, *Recent Advances in Data Mining of Enterprise Data*, World Scientific, Singapore, 2007 (Chapter A survey of manifold-based learning methods, pp. 691–745).