

Performance Analysis of a Manifold Learning Algorithm in Dimension Reduction

Xiaoming Huo^{a,b*} and Andrew K. Smith^a

^a*School of Industrial & Systems Engineering, Georgia Institute of Technology,
Atlanta, GA 30332-0205, USA;*

^b*Department of Statistics, University of California, Riverside, CA 92521, USA.*

Abstract

We consider the performance of local tangent space alignment (Zhang and Zha, 2004), one of several manifold learning algorithms which has been proposed as a dimension reduction method, when errors are present in the observations. Matrix perturbation theory is applied to obtain a worst-case upper bound on the deviation of the solution, which is an invariant subspace. Although we only prove this result for one algorithm, we anticipate that analogous results are derivable for several others due to their strong similarities. Our result clears a conceptual barrier in applying manifold learning algorithms to *noisy* data. It characterizes the situations under which these manifold learning algorithms are effective tools for dimension reduction.

Key words: manifold learning, dimension reduction, local tangent space alignment (LTSA), matrix perturbation analysis

1 Introduction

Manifold learning (ML) algorithms are novel and model-free dimension reduction (DR) approaches. Researchers in *manifold learning* have invented many efficient algorithms. Readers can easily find numerous articles in the machine learning literature. These algorithms differ from traditional statistical DR methods in two ways:

- (1) There is *no* parametric model assumed for the observed data; although we assume a mapping between the observations and a set of intrinsic (low-dimensional) vectors.
- (2) The sampling density is high enough to ensure the recovery of the underlying structure on its support.

* Corresponding author. Email: xiaoming@isye.gatech.edu.

We adopt a typical statistical model: $y_i = f(x_i) + \varepsilon_i, i = 1, 2, \dots, n$, where y_i denotes a high-dimensional noisy observation, $f(\cdot)$ is a mapping satisfying some local regularity conditions, x_i is a low-dimensional intrinsic parameter, ε_i is a random error, and n is the sample size. The objective of DR is to find the set $\{x_i\}$, without any parametric model assumption on f except local smoothness. Our performance analysis will be different from manifold learning, which in general considers *noiseless* observations: $y_i = f(x_i), i = 1, 2, \dots, n$; i.e., ignoring additive random errors. In numerical simulations, all of these algorithms are observed to be robust against errors.

Many efficient ML algorithms have been developed. A partial list of them is: locally linear embedding (LLE) (Roweis and Saul, 2000), ISOMAP (Tenenbaum, de Silva, and Langford, 2000), charting (Brand, 2003), local tangent space alignment (LTSA) (Zhang and Zha, 2004), Laplacian eigenmaps (Belkin and Niyogi, 2003), and Hessian eigenmaps (Donoho and Grimes, 2003), etc.

The main contribution of this paper is to establish some performance properties of a manifold learning algorithm under the presence of errors. The key idea in our analysis is to treat the solutions of manifold learning algorithms as invariant subspaces, and then carry out a matrix perturbation analysis. It has been reported by many (e.g., Roweis and Saul (2000); Belkin and Niyogi (2003); Donoho and Grimes (2003); Brand (2003); Zhang and Zha (2004)) that solutions of their manifold learning algorithms correspond to invariant subspaces which are spanned by the eigenvectors associated with the 2nd through $(d + 1)$ st smallest eigenvalues of certain matrices. The form of such a matrix depends on the details of the algorithm. These subspaces are clearly invariant, because they are spanned by eigenvectors (Stewart and Sun, 1990, Section I.3.4).

LTSA is chosen because it is representative. First of all, in numerical simulation (e.g., using the tools offered by Wittman (2005)), 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 allows us to analyze its performance by applying matrix perturbation theory. Third, the similarity between many manifold learning algorithms (e.g., LLE, Laplacian eigenmaps and Hessian eigenmaps) that their solutions can all be interpreted as invariant subspaces indicates that results for LTSA can be generalized to other algorithms.

Our main theoretical result (Theorem 3.9) gives a worst case bound on the performance of LTSA. To be more specific, let $x_i, i = 1, 2, \dots, n$, denote a set of low-dimensional vectors. For reasons which will become evident later, we call this set the *true parametrization*. Let $y_i, i = 1, 2, \dots, n$, denote the observed high-dimensional vectors that are generated according to $y_i = f(x_i) + \varepsilon_i$, and assume that f is locally regular. Let $\{\tilde{x}_i, 1 \leq i \leq n\}$ denote the estimated parameter set. Let $\mathcal{R}(\tilde{X})$ (resp., $\mathcal{R}(X)$) denote the invariant subspace that is associated with the set $\{\tilde{x}_i, 1 \leq i \leq n\}$ (resp., $\{x_i, 1 \leq i \leq n\}$). (Details regarding invariant subspaces will be articulated in Section 2.6.) We prove the following regarding the distance between the two invariant subspaces, which consequently gives

the worst case analysis on the performance of LTSA:

$$\|\tan(\mathcal{R}(\widetilde{X}), \mathcal{R}(X))\|_2 \leq 4 \cdot \frac{C_3(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau) \cdot \|\sum_{i=1}^n S_i\|_\infty}{\ell_{\min}},$$

where C_3 is a constant that depends on the dimension of the observations, the regularity of the function f and the value of an algorithmic parameter that is used in LTSA, σ is an upper bound on the absolute values of the random errors, τ denotes the size of the neighborhoods, within which f is assumed to be well-behaved, $\|\sum_{i=1}^n S_i\|_\infty$ is equal to the maximum number of times that a single observation appears in nearest-neighbor sets, and ℓ_{\min} is a constant determined by the global structure of the mapping f . The above inequality is established under the conditions that $\tau \rightarrow 0$, $\frac{\sigma}{\tau} \rightarrow 0$, and $\ell_{\min} \rightarrow 0$ at a rate slower than the first two so that the right hand side of the inequality goes to 0; more specifically,

$$\frac{C_3(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau) \cdot \|\sum_{i=1}^n S_i\|_\infty}{\ell_{\min}} \rightarrow 0.$$

Performance analysis will provide theoretical foundation for the application of ML algorithms. To the best of our knowledge, this paper is the first attempt of this kind.

Our result differs from most existing classical DR methodologies. Due to the amount of literature in DR and the space of a single paper, we discuss two of the most popular branches in DR, while emphasizing that there are many more.

- Principal component analysis (PCA) and multidimensional scaling (MDS), together with many extensions, are widely known in statistics. In contrast to PCA, a manifold learning algorithm does *not* require the underlying structure to be a linear subspace. Unlike MDS, a manifold learning algorithm does not impose the same pairwise distances in the data (or observation) space. For example, MDS will fail in the pedagogical numerical example that we will provide later. That example gives a case in which it is not necessary to keep the pairwise distances between observations.
- A recent branch of research in DR involves the idea of a *central subspace*. The most recent work that we are aware of in this area is Cook and Ni (2005). A central subspace is defined only when a response is present - hence it is a supervised learning problem, compared to the unsupervised learning problem that is discussed here. A central subspace is still a globally linear subspace, however, while manifold learning makes no such assumption.

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. Finally, we present concluding remarks in Section 6. Technical proofs are relegated to an Appendix when convenient.

2 Problem Statement and Illustration

This section is organized as follows. A model is introduced in Section 2.1. We give an illustrative example in Section 2.2. The regularity and uniqueness of the mapping f is discussed in Section 2.3. A condition of a *covering* of the observations, which is related to sampling density, is discussed in Section 2.4. The solution that we pursue is specified in Section 2.5. The key technical point of this paper is to interpret the solution to LTSA algorithm as an invariant subspace, which is described in Section 2.6. Some other necessary notation and results are presented in Section 2.7. Finally in Section 2.8, we review LTSA and its modification, on which our analysis will be based.

2.1 Model

To be more specific, we formulate our DR 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}^D$, such that f satisfies a set of regularity conditions. In addition, we require another set of multivariate values $x_i \in \mathbb{R}^d, d < D, i = 1, 2, \dots, n$, such that

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (2.1)$$

where $\varepsilon_i \in \mathbb{R}^D$ denotes a random error. (Recall that the above has been introduced at the beginning of the Introduction.) 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 DR are: (1) Can we find a set of low-dimensional vectors x_i 's and a mapping f , such that (2.1) holds? (2) Can we achieve the smallest for d ? (3) Which kind of regularity conditions should be imposed on f ? (4) Is the model well defined? These questions will be answered in the following.

2.2 A Pedagogical Example

An illustrative example of DR that concretizes our formulation is given in Figure 1. Subfigure (a) shows the true underlying structure of a toy example, a 1-D spiral. The *noiseless* observations are equally spaced points on this spiral. In subfigure (b), 1024 *noisy* observations are generated with multivariate noise satisfying $\varepsilon_i \sim N(\vec{0}, \frac{1}{100} \mathbf{I}_3)$. We then apply LTSA to the noisy observations, using $k = 10$ nearest neighbors. In subfigure (c), 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 subfigure (c).

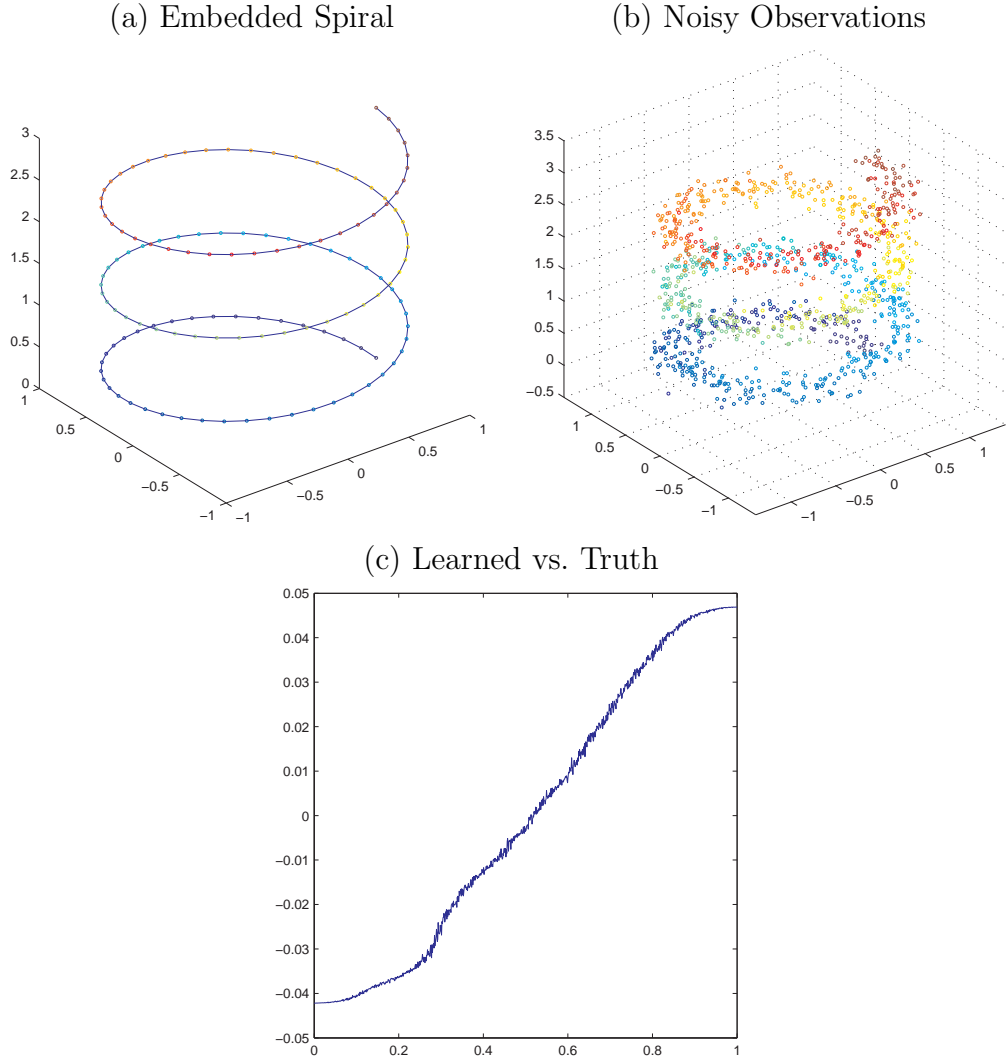


Fig. 1. An illustrative example of LTSA in nonparametric dimension reduction.

2.3 Regularity and Uniqueness of the Mapping f

If the conditions on the mapping f are too general, the model (2.1) is not well defined. For example, if the mapping $f(\cdot)$ and point set $\{x_i\}$ satisfy (2.1), so do $f(A^{-1}(\cdot - b))$ and point set $\{Ax_i + b\}$, where A is an invertible d by d matrix and b is a d -dimensional vector. Borrowing from the manifold learning literature, we adopt the following condition on f .

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

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

for any x_0 in the domain of f and $x \in N_\varepsilon(x_0)$.

The above condition indicates that in a local sense, f preserves the Euclidean distance.

Given the previous condition, model (2.1) 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 (2.1) and Condition 2.1, so do $f(O^T(\cdot - b))$ and $\{Ox_i + b\}$. We can force b to be $\vec{0}$ by imposing the condition that $\sum_i x_i = 0$. In DR, 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.

Our DR problem can be stated as follows. For a set of high dimensional observations $y_i, i = 1, 2, \dots, n$, find low-dimensional $x_i, i = 1, 2, \dots, n$ and a locally isometric mapping f , such that

$$\sum_{i=1}^n \|y_i - f(x_i)\|_2^2 \tag{2.2}$$

is minimized. If the random errors are distributed as $N(\vec{0}, \sigma^2 I_D)$, the above can be justified as a maximum likelihood procedure.

2.4 Covering Conditions and Sampling Density

The above is still too general to determine a solution for f and $\{x_i\}$. The following assumptions are widely adopted in the manifold learning literature. We call them *covering conditions*.

Condition 2.2 (Weak Covering Condition) *There exists a set of index subsets $P_i, i = 1, 2, \dots, m$, which form a (possibly overlapping) covering of the entire index set — $\{1, 2, \dots, n\} = P_1 \cup P_2 \cup \dots \cup P_m$ — and for $j \in P_i$, we have $f(x_j) = f_i(x_j)$, where $f_i(\cdot)$ is locally isometric.*

Recall that if $f_i(\cdot)$ is locally isometric, there exist a constant $\tau_i > 0$ and a vector $x_i^{(0)} \in \mathbb{R}^d$ such that for any $x \in N_{\tau_i}(x_i^{(0)})$, we have

$$\|f_i(x) - f_i(x_i^{(0)})\|_2 = \|x - x_i^{(0)}\|_2 + o(\|x - x_i^{(0)}\|_2).$$

The following condition is slightly stronger.

Condition 2.3 (Strong Covering Condition) *For the same subsets as mentioned in the previous condition, if $j \in P_i$, then $x_j \in N_{\tau_i}(x_i^{(0)})$, where $N_{\tau_i}(x_i^{(0)})$ is defined above.*

The above condition is used as a foundation in manifold learning algorithms. Typically, for a positive integer k , the S_i is taken as the index subset consisting of the $k - 1$ nearest neighbors of the observation y_i in addition to y_i itself.

The above condition can be viewed as a statement of “sufficient sampling density”. It is not trivial to determine when a sample is dense enough. It will depend on the topological

properties of f as well as the local sampling rate. This is not the focus of this paper, and we will not pursue this question any further.

2.5 What Exactly Do We Solve?

For observations $\{Y_i : 1 \leq i \leq n\}$, we want to find a set of ‘parameters’ $\{X_i : 1 \leq i \leq n\}$, such that:

- (1) Within each ‘patch’ P_i , $1 \leq i \leq n$, set $\{X_j : j \in P_i\}$ maximally represents the set of observations $\{Y_j : j \in P_i\}$; (Technically speaking, the rows of $\{X_j\}_{j \in P_i}$ span the subspace that is associated with the d rows of $\{Y_j\}_{j \in P_i}$ with largest variance.)
- (2) The x_i ’s solve the following:

$$\min_{x_i\text{'s}} \sum_{i=1}^n \sum_{j \in P_i} \|Y_j - f_i(x_j)\|_2^2,$$

where $f_i(\cdot)$ is a distance-preserving mapping.

2.6 Solutions as Invariant Subspaces and a Related Metric

We now give a more detailed discussion on invariant subspaces. Let $\mathcal{R}(X)$ denote the subspace spanned by the columns of matrix 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 a measure of distance between two invariant subspaces, and letting X and \tilde{X} denote the true and estimated parameters, respectively, how do we evaluate $\|\tan(\mathcal{R}(X), \mathcal{R}(\tilde{X}))\|_2$?

2.7 More Notation

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

local isometry of f implies that for each $x_0 \in \Omega$ and for x sufficiently close to x_0 , we have

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

$\forall x \in N_\varepsilon(x_0)$. The above in fact implies the following lemma.

Lemma 2.4 *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 (2004).

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

2.8 LTSA: Local Tangent Space Alignment

We now review LTSA (Section 2.8.1), and propose an equivalent version (Section 2.8.2) which is more amenable to perturbation analysis.

2.8.1 Original LTSA

There are two main steps in the original LTSA algorithm (Zhang and Zha, 2004).

- (1) The first step is to compute the local representation on the manifold. 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 $d < k$. 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.

LTSA solves the following optimization problem: for $1 \leq i \leq n$,

$$\min_{U, \Theta} \|Y_i \bar{P}_k - U \Theta\|_F,$$

where $U \in \mathbb{R}^{D \times d}$, $U^T U = I_d$, $\Theta \in \mathbb{R}^{d \times k}$, and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The sum of the columns of $Y_i \bar{P}_k$ is centered at zero. Let U_i and Θ_i denote the solution to the above optimization problem. It is shown in the original LTSA paper that U_i is made by the d left singular vectors of matrix $Y_i \bar{P}_k$ corresponding

to the largest singular values. The columns of Θ_i give local representations of the k points in the local tangent space.

- (2) In the second step, a global alignment is computed. The solution to this step is given by the d eigenvectors corresponding to the 2nd to the $(d + 1)$ st smallest eigenvalues of the matrix

$$(S_1, S_2, \dots, S_n) \bar{P}_k \begin{pmatrix} I_k - \Theta_1^+ \Theta_1 & & & \\ & I_k - \Theta_2^+ \Theta_2 & & \\ & & \ddots & \\ & & & I_k - \Theta_n^+ \Theta_n \end{pmatrix} \bar{P}_k (S_1, S_2, \dots, S_n)^T, \quad (2.3)$$

where $S_i \in \mathbb{R}^{n \times k}$ is the selection matrix associated with Y_i (i.e.,

$$Y_i = (y_1, y_2, \dots, y_n) S_i,$$

$\forall 1 \leq i \leq n$), and Θ_i^+ is the Moore-Penrose pseudo-inverse (Golub and van Loan, 1996) of Θ_i .

As mentioned earlier, the subspace spanned by the eigenvectors associated with the 2nd to the $(d + 1)$ st eigenvalues of the matrix in (2.3) is an invariant subspace, which will be analyzed under perturbation.

2.8.2 Modified LTSA

We will need the following condition to prove the equivalence of the modified version to the original.

Condition 2.5 (Local Linear Independence Condition) *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 reader can easily verify that the following two steps are equivalent to the two steps in the original LTSA. The modified version eliminates the need to consider the pseudo-inverse, and therefore is more amenable to perturbation analysis.

- (1') We solve the minimization problem:

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

where $\Lambda \in \mathbb{R}^{D \times d}$, $V \in \mathbb{R}^{d \times k}$, 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 the matrix $Y_i \bar{P}_k$.

- (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

$$(S_1, \dots, S_n) \begin{pmatrix} \bar{P}_k - V_1^T V_1 & & & \\ & \bar{P}_k - V_2^T V_2 & & \\ & & \ddots & \\ & & & \bar{P}_k - V_n^T V_n \end{pmatrix} (S_1, \dots, S_n)^T. \quad (2.4)$$

In the modified version, we consider the right singular vectors in step 1', instead of the left singular vectors in step 1. By doing so, we eliminate the need to consider pseudo-inverses. The verification of the equivalence is a standard exercise in linear algebra. We relegate the proof to Appendix A.1.

3 Perturbation Analysis

We now carry out a perturbation analysis on the modified 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

The following condition ensures that f is locally smooth. We impose a bound on all the components of the Hessians: $[H_i(f; x)]_{jk}, i = 1, 2, \dots, D, j, k = 1, 2, \dots, d$, which are defined in Section 2.7.

Condition 3.1 (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.

Recall the selection matrices $S_i, i = 1, 2, \dots, n$, which are defined right after (2.3). Let X_i be the true parameter set. 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 . Suppose the strong covering condition (Condition 2.3) holds with the subsets $P_i, i = 1, 2, \dots, n$, being reflected in the selection matrices. Moreover, we require a global upper bound on the size of the neighborhoods.

Condition 3.2 (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 defined in Condition

2.3.

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 $X_i \overline{P}_k$. 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 3.3 *For the k nearest neighbors and the τ that is defined in Condition 3.2, we have*

$$d_{\min} \leq d_{\max} \leq \tau \sqrt{k}. \quad (3.5)$$

For the proof, see Appendix A.2.

Condition 3.4 (Local Tangent Space) *There exists a constant $C_2 > 0$, such that*

$$C_2 \cdot \tau \leq d_{\min}. \quad (3.6)$$

The above condition implies that the parametrization is sufficiently “well-behaved” that it is recoverable. This condition is not satisfied by topological structures that have different dimensions (e.g., Hausdorff dimension) at different places.

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

Condition 3.5 (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. Notice that this condition is also somewhat nonstandard, since the magnitude of the errors is assumed to depend on n , but it seems to be necessary to ensure the consistency of LTSA.

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

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

be the singular value decomposition of the matrix $X_i \overline{P}_k$; 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 $B_i \in \mathbb{R}^{d \times k}$ is made by the right singular vectors

$(B_i B_i^T = I_d)$. It is not hard to verify that

$$B_i = B_i \bar{P}_k. \quad (3.7)$$

Let \tilde{B}_i denote the matrix made by the perturbed d right singular vectors (i.e., the perturbed version of B_i). 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 3.6 *Given invariant subspaces $\mathcal{R}(B_i^T)$ and $\mathcal{R}(\tilde{B}_i^T)$ as defined above, we have*

$$\|\tan(\mathcal{R}(B_i^T), \mathcal{R}(\tilde{B}_i^T))\|_2 \leq C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right),$$

where C_3 is a constant that depends on k , D and C_2 , and $\|\tan(\cdot, \cdot)\|_2$ is the distance between two invariant subspaces that has been defined in matrix perturbation theory.

The proof is presented in Appendix A.3. The above gives an upper bound on the deviation of the local invariant subspace in step (1') of the modified LTSA. It will be called later to prove a global result.

3.2 Global Alignment

Condition 3.7 (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 constant C_4 depends on the covering $\{P_i\}$. The next condition (Condition 3.8) 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 matrix $\sum_{i=1}^n S_i$. 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 show an upper bound on the deviation of the invariant subspace between the result of LTSA working on the noisy observations and the true set of parameters.

Given (3.7), 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 the matrix

$$(\mathbf{1}_n, X)^T \quad (3.8)$$

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

$$(S_1, \dots, S_n) \bar{P}_k \begin{pmatrix} I - B_1^T B_1 & & & \\ & I - B_2^T B_2 & & \\ & & \ddots & \\ & & & I - B_n^T B_n \end{pmatrix} \bar{P}_k (S_1, \dots, S_n)^T. \quad (3.9)$$

Assume that the matrix

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

is unitary, where $X^c \in \mathbb{R}^{n \times (n-1-d)}$. Based on the previous paragraph, we have

$$\begin{pmatrix} \frac{\mathbf{1}_n^T}{\sqrt{n}} \\ X^T \\ (X^c)^T \end{pmatrix} M_n \begin{pmatrix} \mathbf{1}_n \\ \sqrt{n} \\ X \\ X^c \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{(d+1) \times (d+1)} & \mathbf{0}_{(d+1) \times (n-d-1)} \\ \mathbf{0}_{(n-d-1) \times (d+1)} & L_2 \end{pmatrix} \quad (3.10)$$

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 3.8 (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

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

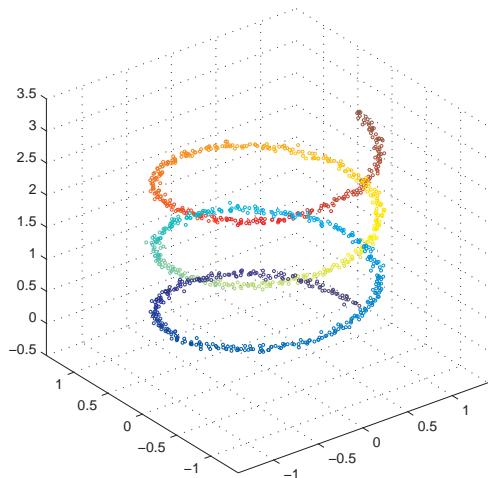
Theorem 3.9 (Main Theorem)

$$\|\tan(\mathcal{R}(\tilde{X}), \mathcal{R}(X))\|_2 \leq 4 \cdot \frac{C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau\right) \cdot \|\sum_{i=1}^n S_i\|_{\infty}}{\ell_{\min}}. \quad (3.11)$$

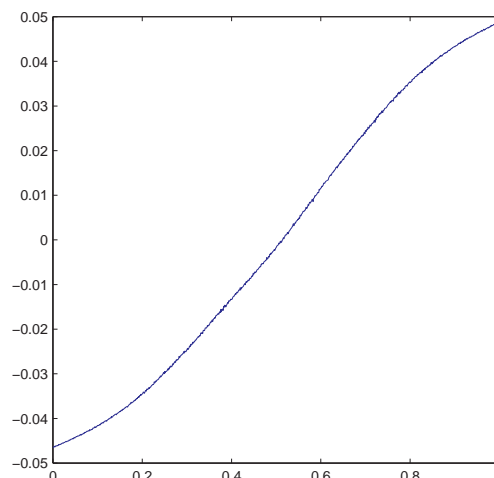
As mentioned in Introduction, the above theorem gives a worst-case bound on the performance of LTSA. A discussion on when Condition 3.8 is satisfied will be long and beyond the scope of this paper. We leave it to future investigation.

4 Simulations

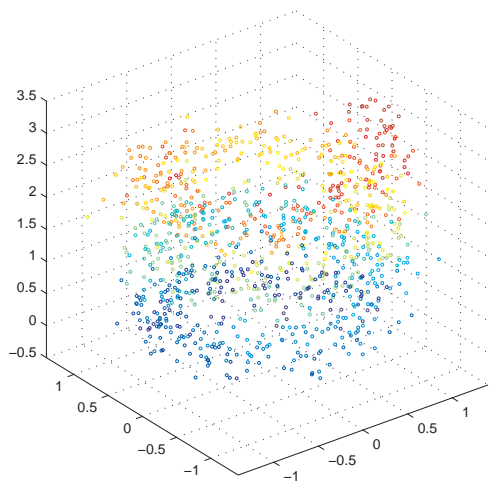
(a) Noisy Observations when $\sigma = 0.025$



(b) Result of LTSA



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



(d) Result of LTSA

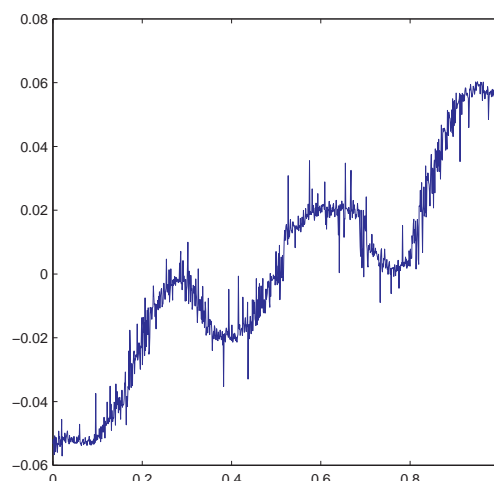


Fig. 2. Reruns of the illustrative example in Section 2.2, with different noise standard deviations.

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 Figure 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 \widetilde{X} are one-dimensional, we have

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

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

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

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

In Figure 2 (b), when σ is small, we observe a nearly straight line; while in Figure 2 (d), 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 DR 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.

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.

We assume that the covering P_i is given, such that $\tau \rightarrow 0$ holds. 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 with Existing Works

The error analysis in the original paper about LTSA is the closest to our result. However, Zhang and Zha (2004) 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 (Roweis and Saul, 2000), Laplacian eigenmaps (Belkin and Niyogi, 2003), and Hessian eigenmaps (Donoho and Grimes, 2003) 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 (Huo et al., 2005) 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.

The DR problem considered here is an unsupervised learning problem. There are supervised learning problems that are of similar flavor, e.g., contour regression and inverse regression Li et al. (2005). As mentioned in the Introduction, the rich literature in the supervised counterpart (e.g., the concepts of *exhaustiveness*, *sufficiency*, *central subspaces*, etc) gives motivation to derive corresponding results in the supervised framework.

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, as long as their solutions are invariant subspaces.

A Proofs

A.1 Proof of Equivalence between Two LTSA Algorithms

It suffices to verify that

$$\overline{P}_k \Theta^+ \Theta \overline{P}_k = V^T V. \quad (\text{A.1})$$

Let $Y_k \overline{P}_k = ADB$ be the singular value decomposition of the matrix $Y_k \overline{P}_k$. Suppose the singular vectors are order at decreasing order in the diagonal of matrix D . Let A_k (resp., B_k) denote the submatrix of A (resp., B) by taking the first k columns (resp., rows) of that matrix. We have $A_k \in \mathbb{R}^{D \times d}$ and $B_k \in \mathbb{R}^{d \times k}$. Recall $d < k$. Let D_k be the upper-left d by d submatrix of D . Given Condition 2.5, the matrix D_k is invertible. We can easily verify $\Theta = D_k B_k$ and $V = B_k$. It is not hard to see $\Theta^+ = B_k^T D_k^{-1}$. Moreover, the rows of B_k span a subspace that is a subset of the subspace spanned by the rows of \overline{P}_k . Considering that \overline{P}_k is a projection matrix and that the rows of B_k form a subset of an orthonormal basis, we have $B_k \overline{P}_k = B_k$. Combining all the above, we have

$$\begin{aligned} \text{left hand side of (A.1)} &= \overline{P}_k B_k^T D_k^{-1} D_k B_k \overline{P}_k \\ &= B_k^T B_k \\ &= \text{right hand side of (A.1)}. \end{aligned}$$

A.2 Proof of Lemma 3.3

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

$$\begin{aligned} d_{\max, i} &= \|X_i \overline{P}_k\|_2 = \|(X_i - x_0 \cdot \mathbf{1}_k^T) \overline{P}_k\|_2 \\ &\leq \|X_i - x_0 \cdot \mathbf{1}_k^T\|_2 \end{aligned} \quad (\text{A.2})$$

$$\leq \sqrt{k} \cdot \max_{j \in P_i} \|x_j - x_0\|_2 \quad (\text{A.3})$$

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

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

In the above, inequality (A.2) is true because in general, for two matrices A and B , we have $\|AB\|_2 \leq \|A\|_2 \cdot \|B\|_2$ (Stewart and Sun (1990, page 69)). The inequality (A.3) is also standard linear algebra (Stewart and Sun (1990, page 71)).

A.3 Proof of Theorem 3.6

The following two equations will be used:

$$Y_i \bar{P}_k = (Y_i - f(x_i^{(0)}) \cdot \mathbf{1}_k^T) \bar{P}_k \quad (\text{A.4})$$

and

$$X_i \bar{P}_k = (X_i - x_i^{(0)} \cdot \mathbf{1}_k^T) \bar{P}_k, \quad (\text{A.5})$$

where $x_i^{(0)}$ is defined right before Condition 2.3. Readers can easily verify them 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 $\|E_i\|_2$:

$$\begin{aligned} \|E_i\|_2 &= \|Y_i \bar{P}_k - J(f; x_i^{(0)}) X_i \bar{P}_k\|_2 \\ &\leq \sqrt{k} \cdot \sup_{j \in P_i} \|(y_j - f(x_i^{(0)})) \cdot \bar{P}_k - J(f; x_i^{(0)})(x_j - x_i^{(0)}) \cdot \bar{P}_k\|_2 \\ &\leq \sqrt{k} \cdot \sup_j \|(y_j - f(x_i^{(0)})) - J(f; x_i^{(0)})(x_j - x_i^{(0)})\|_2 \\ &\leq \sqrt{kD} \cdot \sup_j \|(y_j - f(x_i^{(0)})) - J(f; x_i^{(0)})(x_j - x_i^{(0)})\|_\infty \\ &\leq \sqrt{kD} \cdot [\sigma + \frac{1}{2} C_1 \tau^2]. \end{aligned} \quad (\text{A.6})$$

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.

It is also easy to see that

$$(Y_i \bar{P}_k)^T (Y_i \bar{P}_k) = B_i^T D_i^2 B_i + E_i^T J(f; x_i^{(0)}) A_i D_i B_i + B_i^T D_i A_i^T J^T(f; x_i^{(0)}) E_i + E_i^T E_i.$$

Let $R_i = (Y_i \bar{P}_k)^T (Y_i \bar{P}_k) - B_i^T D_i^2 B_i$. Note that $(X_i \bar{P}_k)^T (X_i \bar{P}_k) = B_i^T D_i^2 B_i$. As a preparation to invoke Theorem V.2.7 in Stewart and Sun (1990), we assume that the square matrix

$$\begin{pmatrix} B_i \\ B_i^c \end{pmatrix} \in \mathbb{R}^{k \times k}$$

is an orthogonal matrix, i.e., $B_i^c \in \mathbb{R}^{(k-d) \times k}$, $B_i^c (B_i^c)^T = I_{(k-d)}$, and $B_i^c B_i^T = \mathbf{0}_{(k-d) \times d}$. The following is self-evident:

$$\begin{pmatrix} B_i \\ B_i^c \end{pmatrix} (X_i \bar{P}_k)^T (X_i \bar{P}_k) \begin{pmatrix} B_i^T \\ (B_i^c)^T \end{pmatrix} = \begin{pmatrix} D_i^2 & \mathbf{0}_{d \times (k-d)} \\ \mathbf{0}_{(k-d) \times d} & \mathbf{0}_{(k-d) \times (k-d)} \end{pmatrix}.$$

On the other hand, we have

$$\begin{aligned} \begin{pmatrix} B_i \\ B_i^c \end{pmatrix} R_i \begin{pmatrix} B_i^T \\ (B_i^c)^T \end{pmatrix} &= \begin{pmatrix} B_i \\ B_i^c \end{pmatrix} E_i^T J(f; x_i^{(0)}) A_i D_i (I_d, \mathbf{0}_{d \times (k-d)}) \\ &\quad + \begin{pmatrix} I_d \\ \mathbf{0}_{(k-d) \times d} \end{pmatrix} D_i A_i^T J^T(f; x_i^{(0)}) E_i (B_i^T, (B_i^c)^T) \\ &\quad + \begin{pmatrix} B_i \\ B_i^c \end{pmatrix} E_i^T E_i (B_i^T, (B_i^c)^T) \\ &\triangleq \begin{pmatrix} R_{11} & R_{12} \\ R_{12}^T & R_{22} \end{pmatrix}, \end{aligned}$$

where $R_{11} \in \mathbb{R}^{d \times d}$, $R_{12} \in \mathbb{R}^{d \times (k-d)}$, $R_{22} \in \mathbb{R}^{(k-d) \times (k-d)}$, and

$$\begin{aligned} R_{11} &= B_i E_i^T J(f; x_i^{(0)}) A_i D_i + D_i A_i^T J^T(f; x_i^{(0)}) E_i B_i^T + B_i E_i^T E_i B_i^T, \\ R_{12} &= D_i A_i^T J^T(f; x_i^{(0)}) E_i (B_i^c)^T + B_i E_i^T E_i (B_i^c)^T, \\ R_{22} &= B_i^c E_i^T E_i (B_i^c)^T. \end{aligned}$$

Letting $\|\cdot\|$ be a *consistent family of matrix norms* (Stewart and Sun, 1990, page 69), we have

$$\begin{aligned} \|R_{11}\| &\leq 2\|E_i\| \cdot \|D_i\| + \|E_i\|^2, \\ \|R_{12}\| &\leq \|D_i\| \cdot \|E_i\| + \|E_i\|^2, \\ \|R_{22}\| &\leq \|E_i\|^2. \end{aligned}$$

To verify the conditions in Theorem V.2.7 in Stewart and Sun (1990), it is necessary to

specify the range of the singular values of the matrices $X_i \overline{P}_k$, $1 \leq i \leq n$. Assumption 3.4 is a sufficient condition for LTSA to recover the local linear structure of the manifold. If we have

$$d_{\min}^2 \geq 4(\|D_i\| \cdot \|E_i\| + \|E_i\|_2^2), \quad (\text{A.7})$$

one can verify that all the conditions in Theorem V.2.7. in Stewart and Sun (1990) are satisfied. Note that d_{\min}^2 corresponds to, in the language of Stewart and Sun (1990), $\text{sep}(D_i^2, \mathbf{0}_{(k-d) \times (k-d)})$.

We validate inequality (A.7). Recall that $\|E_i\|_2 \leq \sqrt{kD}(\sigma + \frac{1}{2}C_1\tau^2)$ (cf. (A.6)). In addition, we have $\|D_i\|_2 = d_{\max,i} \leq \sqrt{k}\tau$ (Lemma 3.3), and $d_{\min} \geq C_2\tau$ (cf. Condition 3.4). When $\tau \rightarrow 0$, which is the case that is of interest to us, (A.7) holds; because

$$\begin{aligned} \frac{\text{L. H. S. of (A.7)}}{\text{R. H. S. of (A.7)}} &\geq \frac{C_2^2\tau^2}{4\|E_i\|_2(\|D_i\|_2 + \|E_i\|_2)} \\ &\geq \frac{C_2^2\tau^2}{4\sqrt{kD}(\sigma + \frac{1}{2}C_1\tau^2)[\sqrt{k}\tau + \sqrt{kD}(\sigma + \frac{1}{2}C_1\tau^2)]} \\ &= \frac{C_2^2}{4k\sqrt{D}} \cdot \frac{1}{(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau)[1 + \sqrt{D}(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau)]} \\ &\rightarrow +\infty, \end{aligned}$$

as $\tau \rightarrow 0$, which (by Condition 3.5) implies $\frac{\sigma}{\tau} \rightarrow 0$.

Hence by invoking Theorem V.2.7 in Stewart and Sun (1990), we have

$$\begin{aligned} \|\tan(\mathcal{R}(B_i^T), \mathcal{R}(\tilde{B}_i^T))\|_2 &\stackrel{\text{Th. V.2.7.}}{\leq} 2 \cdot \frac{\|D_i\|_2 \cdot \|E_i\|_2 + \|E_i\|_2^2}{d_{\min}^2 - 2\|D_i\|_2 \cdot \|E_i\|_2 - 2\|E_i\|_2^2} \\ &\leq 4 \cdot \frac{d_{\max}\|E_i\|_2 + \|E_i\|_2^2}{d_{\min}^2} \\ (3.5) \ \&\ (3.6) \quad &\leq 4 \cdot \frac{\sqrt{k} \cdot \tau \|E_i\|_2 + \|E_i\|_2^2}{C_2^2 \cdot \tau^2} \\ &= \frac{4}{C_2^2\tau^2} \|E_i\|_2 (\sqrt{k}\tau + \|E_i\|_2) \\ (A.6) \quad &\leq \frac{4\sqrt{kD}}{C_2^2\tau^2} (\sigma + \frac{1}{2}C_1\tau^2) [\sqrt{k}\tau + \sqrt{kD}(\sigma + \frac{1}{2}C_1\tau^2)] \\ &= \frac{4k\sqrt{D}}{C_2^2} (\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau) [1 + \sqrt{D}(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau)] \\ &\leq \frac{8k\sqrt{D}}{C_2^2} (\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau). \end{aligned}$$

Take $C_3 = \frac{8k\sqrt{D}}{C_2^2}$, we prove the theorem.

A.4 Proof of Theorem 3.9

Now we consider the step of global alignment. Recall that the columns of $(\mathbf{1}_n, X)$, where X is defined in (3.8), are eigenvectors associated with the zero eigenvalue of (3.9). We choose a matrix $U \in \mathbb{R}^{(n-1) \times n}$, such that $UU^T = I_{(n-1)}$ and $U\mathbf{1}_n = \mathbf{0}_{(n-1) \times 1}$. The purpose of introducing U is to convert an invariant space problem spanned by the columns of $(\mathbf{1}_n, X)$ to a problem with the invariant space spanned by the columns of UX .

The following describes how to remove $\mathbf{1}_n$ from the invariant subspace. It is easy to verify that

$$U^TUX = X, \quad U^TUX^c = X^c;$$

because U^TU forms a projection matrix of the column vectors of U^T (i.e., $U^TUU^T = U^T$) and the columns of X and X^c form a basis of the subspace spanned by the columns of U^T . Hence we have

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

The reader may compare the above with equation (3.10). Removing the first column and first row of the above matrix equation, we have

$$\begin{pmatrix} X^T \\ (X^c)^T \end{pmatrix} U^TUM_nU^TU(X, X^c) = \begin{pmatrix} \mathbf{0}_{d \times d} & \mathbf{0}_{d \times (n-d-1)} \\ \mathbf{0}_{(n-d-1) \times d} & L_2 \end{pmatrix}.$$

The above states that the submatrices UX form the invariant subspaces of UM_nU^T that is associated with the null space. Similar to the analysis in the local step, we define

$$\tilde{R} = U\tilde{M}_nU^T - UM_nU^T,$$

where

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

and \tilde{B}_i is defined right before Theorem 3.6. Note that $\tilde{R} \in \mathbb{R}^{(n-1) \times (n-1)}$. We further define the following:

$$\begin{pmatrix} \tilde{R}_{11} & \tilde{R}_{12} \\ \tilde{R}_{12}^T & \tilde{R}_{22} \end{pmatrix} = \begin{pmatrix} X^T \\ (X^c)^T \end{pmatrix} U^T \tilde{R} U(X, X^c),$$

where $\tilde{R}_{11} \in \mathbb{R}^{d \times d}$, $\tilde{R}_{12} \in \mathbb{R}^{d \times (n-d-1)}$, and $\tilde{R}_{22} \in \mathbb{R}^{(n-1-d) \times (n-d-1)}$. We can easily verify the following:

$$\tilde{R}_{11} = X^T U^T U \tilde{M}_n U^T U X, \quad (\text{A.8})$$

$$\tilde{R}_{12} = X^T U^T U \tilde{M}_n U^T U X^c, \quad (\text{A.9})$$

$$\tilde{R}_{22} = (X^c)^T U^T U \tilde{M}_n U^T U X^c - L_2. \quad (\text{A.10})$$

We now consider the norm $\|\tilde{R}_{11}\|_2$. For $y \in \mathbb{R}^d$ and $\|y\|_2 = 1$, we have

$$y^T \tilde{R}_{11} y = y^T X^T U^T U \tilde{M}_n U^T U X y.$$

Recall Theorem 3.6. Given that $\sin \theta \leq \tan \theta$, for $0 < \theta < \frac{\pi}{2}$, the singular values of the matrix

$$(I_k - \tilde{B}_i^T \tilde{B}_i) - (I_k - B_i^T B_i)$$

are no larger than $C_3(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau)$. (See Stewart and Sun (1990, Theorem I.5.5); Note matrices $\tilde{B}_i^T \tilde{B}_i$ and $B_i^T B_i$ are projection matrices.) Hence we have

$$\begin{aligned} y^T \tilde{R}_{11} y &\leq \sum_{i=1}^n C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \cdot \|\bar{P}_k S_i^T U^T U X y\|_2^2 \\ &\leq C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \cdot \sum_{i=1}^n \|S_i^T U^T U X y\|_2^2 \\ &\leq C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \left\| \sum_{i=1}^n S_i \right\|_{\infty} \|U^T U X y\|_2^2 \\ &\leq C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \left\| \sum_{i=1}^n S_i \right\|_{\infty} \|y\|_2^2 \\ &= C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \left\| \sum_{i=1}^n S_i \right\|_{\infty}. \end{aligned}$$

In the above, $\|\sum_{i=1}^n S_i\|_{\infty}$ is the maximum row sum of the matrix $\sum_{i=1}^n S_i$ (Theorem II.2.10 in Stewart and Sun (1990).) The third inequality is a consequence of the fact that S_i 's are selection matrices – we omit the detailed verification. The above is equivalent to the following:

$$\|\tilde{R}_{11}\|_2 \leq C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \left\| \sum_{i=1}^n S_i \right\|_{\infty}. \quad (\text{A.11})$$

A nearly identical argument will prove the following:

$$\|\tilde{R}_{22}\|_2 \leq C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \left\| \sum_{i=1}^n S_i \right\|_{\infty}. \quad (\text{A.12})$$

Also, a slightly modified derivation will yield:

$$\|\tilde{R}_{12}\|_2 \leq C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \left\| \sum_{i=1}^n S_i \right\|_{\infty}. \quad (\text{A.13})$$

The above can be derived from the following analysis: First of all, we have

$$\|\tilde{R}_{12}\|_2 = \max_{\|y\|_2=1, \|z\|_2=1} y^T \tilde{R}_{12} z.$$

Secondly, for $y \in \mathbb{R}^d$, $\|y\|_2 = 1$ and $z \in \mathbb{R}^{n-d-1}$, $\|z\|_2 = 1$, we have

$$\begin{aligned} y^T \tilde{R}_{12} z &= y^T X^T U^T U \tilde{M}_n U^T U X^c z \\ &= y^T X^T U^T U (\tilde{M}_n - M_n) U^T U X^c z \\ &= \sum_{i=1}^n (y^T X^T U^T U S_i \bar{P}_k) [(I_k - \tilde{B}_i^T \tilde{B}_i) - (I_k - B_i^T B_i)] [\bar{P}_k S_i^T U^T U X^c z] \\ &\leq \frac{1}{2} C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \sum_{i=1}^n [\|y^T X^T U^T U S_i \bar{P}_k\|_2^2 + \|\bar{P}_k S_i^T U^T U X^c z\|_2^2] \\ &\leq \frac{1}{2} C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \left\| \sum_{i=1}^n S_i \right\|_{\infty} (\|y\|_2^2 + \|z\|_2^2) \\ &= C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \left\| \sum_{i=1}^n S_i \right\|_{\infty}. \end{aligned}$$

The second equality takes advantage of the fact: $X^T U^T U M_n U^T U X^c = \mathbf{0}_{d \times (n-d-1)}$. From the above two, we have (A.13).

From the above, we have that the columns of the matrix $U\tilde{X}$ form a basis of the invariant subspace of the matrix $U\tilde{M}_n U^T$, and the subspace is spanned by the eigenvectors associated with the 2nd to the $(d+1)$ st smallest eigenvalues. By applying Theorem V.2.7. in Stewart and Sun (1990), we have

$$\|\tan(\mathcal{R}(U\tilde{X}), \mathcal{R}(UX))\|_2 \leq \frac{2\|\tilde{R}_{12}\|_2}{\ell_{\min} - \|\tilde{R}_{11}\|_2 - \|\tilde{R}_{22}\|_2}, \quad (\text{A.14})$$

where $\mathcal{R}(U\tilde{X})$ and $\mathcal{R}(UX)$ are the invariant subspaces spanned by the columns of the matrices $U\tilde{X}$ and UX , respectively. From (A.11), (A.12), (A.13), and Condition 3.8, together with the content of Section I.5 in Stewart and Sun (1990), we can verify the following:

$$\|\tan(\mathcal{R}(U\tilde{X}), \mathcal{R}(UX))\|_2 \leq 4 \cdot \frac{C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right) \cdot \left\| \sum_{i=1}^n S_i \right\|_{\infty}}{\ell_{\min}}. \quad (\text{A.15})$$

Recall that $\mathbf{1}_n^T X = \mathbf{1}_n^T \widetilde{X} = 0$, and

$$\begin{pmatrix} \mathbf{1}_n^T / \sqrt{n} \\ U \end{pmatrix}$$

is unitary (i.e., orthogonal). The inequality (A.15) is equivalent to

$$\|\tan(\mathcal{R}(\widetilde{X}), \mathcal{R}(X))\|_2 \leq 4 \cdot \frac{C_3(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau) \cdot \|\sum_{i=1}^n S_i\|_\infty}{\ell_{\min}}.$$

Thus, we have proved the theorem.

References

- [1] Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15 (6), 1373–1396.
- [2] Brand, M., March 2003. Charting a manifold. In: *Neural Information Processing Systems*. Vol. 15. Mitsubishi Electric Research Labs, MIT Press.
- [3] Cook, R. D., Ni, L., June 2005. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *Journal of the American Statistical Association* 100 (470), 410–428.
- [4] Donoho, D. L., Grimes, C. E., 2003. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Arts and Sciences* 100, 5591–5596.
- [5] Golub, G. H., van Loan, C. F., 1996. *Matrix computations*, 3rd Edition. Johns Hopkins University Press, Baltimore.
- [6] Huo, X., Ni, X. S., Smith, A. K., 2005. *Mining of Enterprise Data*. Springer, New York, Ch. A survey of manifold-based learning methods, invited book chapter, submitted.
- [7] Li, B., Zha, H., Chiaromonte, F., August 2005. Contour regression: a general approach to dimension reduction. *Annals of Statistics* 33 (4), 1580–1616.
- [8] Roweis, S. T., Saul, L. K., 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326.
- [9] Stewart, G. W., Sun, J.-G., 1990. *Matrix Perturbation Theory*. Academic Press, Boston, MA.
- [10] Tenenbaum, J. B., de Silva, V., Langford, J. C., 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323.
- [11] Wittman, T., April 2005. MANifold learning Matlab demo. URL: <http://www.math.umn.edu/~wittman/mani/index.html>.
- [12] Zhang, Z., Zha, H., 2004. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing* 26 (1), 313–338.