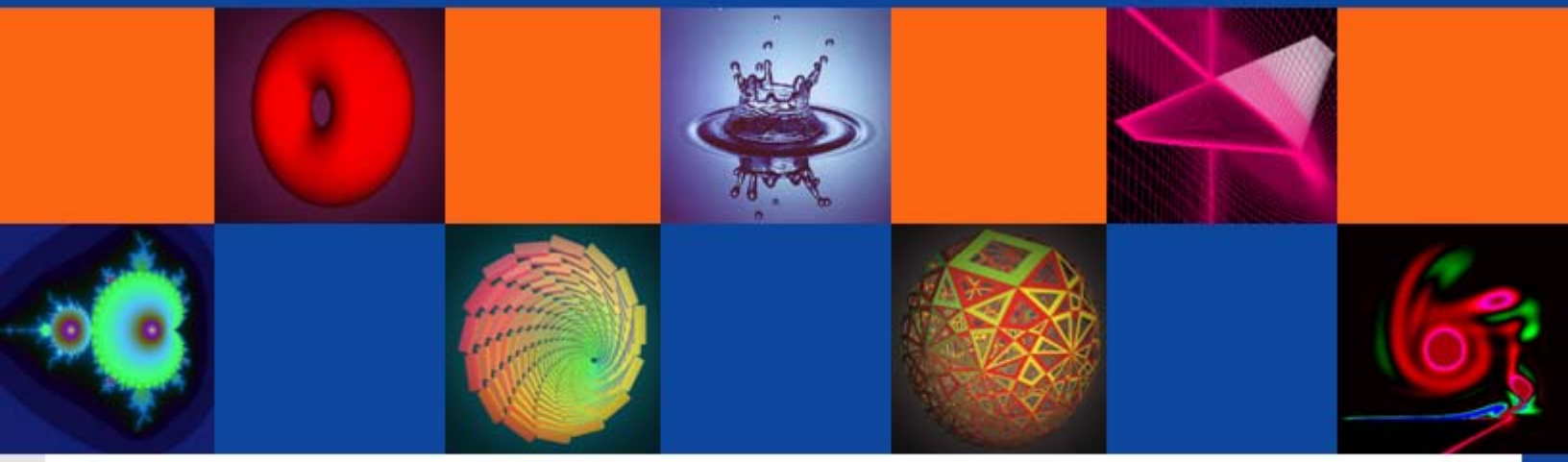
The University of Texas ARLINGTON

MATHEMATICS
Preprint Series



# Active Learning Methods for Semi-supervised Manifold Learning

**Xin Yang**
**Ren-Cang Li**
**Hongyuan Zha**

**Technical Report 2009-06**

http://www.uta.edu/math/preprint/

# Active Learning Methods for Semi-supervised Manifold Learning

Xin Yang [*] Ren-Cang Li [†] Hongyuan Zha [‡]

**Abstract**

The objective of this paper is to propose a principled approach for selecting the data points for labeling used in semi-supervised manifold learning. We postulate that the data points should be chosen so that the alignment matrix for the remaining data points will have the best condition number possible. We also proposed an efficient algorithm for selecting principal submatrices of the alignment matrices with good condition numbers. Experiments on both synthetic and real-world problems show that our proposed method can substantially improve the accuracy of the computed global parameterizations over several alternative methods.

## 1 Introduction

A nonlinear manifold can be parameterized in infinite many different ways. Labeled points in the form of known parameter vectors for certain data points can be exploited to identify the parameterization that is physically meaningful to the problem under investigation. This gives rise to the so-called semi-supervised manifold learning methods which have been applied to several interesting problems such as video annotation and pose estimation [13, 2, 17]. However, no systematic investigation has been done on how to best select the data points for labeling used in semi-supervised manifold learning. The main objective of this paper is to propose a principled method for selecting the data points for labeling and to analyze the performance of the resulting algorithms. Our proposed methods can be considered as a form of active learning for semi-supervised manifold learning.

We motivate our methods using local manifold learning methods such as LLE and LTSA as concrete examples. But they can also be applied to other manifold learning methods. LLE and LTSA are based on solving an eigenvalue problem of the following form [14, 15, 18, 19],

$$\min_{YY^{\mathrm{T}}=I_{d+1}} \Phi(Y) = \mathrm{trace}(YMY^{\mathrm{T}}), \tag{1}$$

[*]Department of Computer Science and Engineering, The Pennsylvania State University.

[†]Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019-0408 (rcli@uta.edu). Supported in part by the National Science Foundation under Grant No. DMS-0510664. and DMS-0702335.

[‡]Division of Computational Science and Engineering, College of Computing, Georgia Institute of Technology, Atlanta, GA 30332-0280.

where $M$ is the so-called alignment matrix. Specifically, the parameter vectors (for some parameterization) are computed from the $d$ eigenvectors of $M$ corresponding to its 2nd to $(d+1)$st smallest eigenvalues. Here $d$ is the intrinsic dimension of the manifold and is assumed to be known.

In semi-supervised manifold learning [9, 17], we assume that the parameter vectors for a subset of the $m$ data points are already known. Let us partition $Y$ as $[Y_1, Y_2]$, and assume $Y_1$ is known, corresponding to the set of known parameter vectors. Conformally partition $M$ as $M = [M_{ij}]_{i,j=1}^{2}$ with $M_{11} \in \mathcal{R}^{m \times m}$. Problem (1) becomes

$$\min_{Y_2} \text{trace}\left(Y_1 M_{11} Y_1^{\mathrm{T}} + 2 Y_1 M_{12} Y_2^{\mathrm{T}} + Y_2^{\mathrm{T}} M_{22} Y_2^{\mathrm{T}}\right). \tag{2}$$

Since $Y_1$ is known, it follows that $Y_2$ can be obtained by solving the following linear system of equations

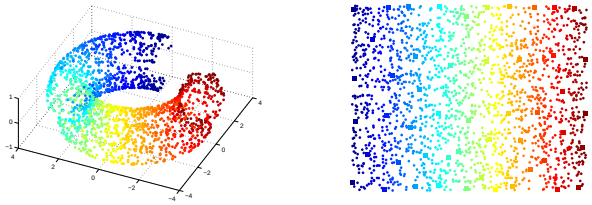
$$M_{22} Y_2^{\mathrm{T}} = -M_{12}^{\mathrm{T}} Y_1^{\mathrm{T}}. \tag{3}$$

In some applications, $Y_1$ is given and we may not have the freedom to choose which data points to select for labeling. In many applications, we do have the freedom to choose which data points to label (which we will call prior points) and we should explore it to our advantage. For example, in the video annotation application discussed in Section 6, we want to annotate the positions of the elbow and the wrist of a person in each frame of a video clip. In this case, we can certainly choose the set of frames to annotate manually so that the rest of the frames can be annotated as accurately as possible by the semi-supervised manifold learning algorithm. Generally, we seek to select the data points so that the resulting semi-supervised manifold learning problem is more robust and gives more accurate parameter vectors for the remaining data points. This can be considered as a form of *active learning* for the manifold learning problems. One can always randomly select a subset for labeling, but we shall demonstrate that we can do better with more sophisticated methods.

# 2 An Illustrative Example

It was already shown that using the so-called landmark points as the prior points [5], considerably improved the solution than that with randomly selected prior points [17]. The landmark points was originally used for reducing the computational complexity of Isomap method [5, 16]. It is a greedy algorithm and it also involves the computation of approximate pairwise geodesic distances between the sample points. Here we illustrate its use for a data set with 2000 data points sampled from the "incomplete tire" (a section of the torus with a slice and a strip cut out, shown in Figure 1(a) ), Generated by Matlab code:

```
s = pi*5*rand(1,N)/3;
t = pi*5*rand(1,N)/3;
X = [(3+cos(s)).*cos(t);
     (3+cos(s)).*sin(t);
     sin(s);]
```



(a) (b)

Figure 1: (a) The incomplete tire; (b) The true $(s, t)$-coordinates for the incomplete tire.

In Figure 3 on Page 6, we compare the average errors of the landmark point method and the random selection method using LLE (SSLLE) and LTSA (SSLTSA). We define the average error as $\|Y - Y_{\text{true}}\|/\|Y_{\text{true}}\|$, where $Y$ is the computed coordinates and $Y_{\text{true}}$ is the true parameter vectors.

For this example, the landmark point method is certainly better than random selection. However, its performance improvement is not consistent, and we will see examples where it performs even worse than random selection! Moreover, a rigorous mathematical analysis is yet to be done for the landmark point method. This motivates us to develop more robust methods discussed below.

# 3 A Data Selection Principle

Equation (3) shows that $Y_2^{\mathrm{T}}$ can be computed by solving a linear system with $M_{22}$ as the coefficient matrix. Standard perturbation theory [6, 8] for linear systems indicates that the degree of accuracy with which we can numerically compute $Y_2^{\mathrm{T}}$ is roughly proportional to the condition number of $M_{22}$. This observation leads us to the following data selection principle:

> Given $m$, select a subset of $m$ data points to label so that the condition number of the resulting $M_{22}$ is as small as possible.

Since $M$ is a symmetry positively semidefinite matrix, so is its principal submatrix $M_{22}$. The problem reduces to symmetrically deleting $m$ rows and columns of $M$ so as to minimize the condition number $\kappa(M_{22}) = \lambda_{\max}(M_{22})/\lambda_{\min}(M_{22})$, where $\lambda_{\max}(M_{22})$ and $\lambda_{\min}(M_{22})$ are the largest and smallest eigenvalues of $M_{22}$, respectively. This is equivalent to finding an $(n-m) \times (n-m)$ principle submatrix of $M$ with the smallest possible condition number.

How small can the condition numbers of the principal submatrices be? We answer this fundamental question in the next section using matrix analysis techniques. We then present numerical methods that can deliver principal submatrices with small condition numbers.

# 4 Data Selection: Theory

Rank-revealing QR-factorizations (RRQR) [3, 7, 10] are designed to reveal the numerical rank of a matrix, here we introduce it to prove an existence theorem of a good principle submatrix.

Given an $n \times n$ matrix $B$ with singular values $\sigma_1(B) \geq \sigma_2(B) \geq \cdots \geq \sigma_n(B) \geq 0$. If there is a reasonable gap between $\sigma_k(B)$ and $\sigma_{k+1}(B)$, and $\sigma_{k+1}(B)$ is small enough, it is reasonable to regard that $B$ has a numerical rank $k$. In this case, any RRQR attempts to find a permutation matrix $\Pi$ such that the QR factorization

$$B\Pi = QR, \quad R = \begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix}, \quad R_{11} \text{ is } k \times k \tag{4}$$

satisfies $R_{11}$'s smallest singular value $\sigma_{\min}(R_{11}) \approx \sigma_k(B)$ and $R_{22}$'s largest singular value $\sigma_{\max}(R_{22}) \approx \sigma_{k+1}(B)$, where $Q$ is orthogonal, $R$ is upper triangular. In essence, $R_{11}$ captures the well-conditioned part of $B$.

**Lemma 1** *Let $B$ be $n \times n$ and let $1 \leq k < n$. Then there exists an RRQR* (4) *such that*

$$\sigma_{\min}(R_{11}) \geq \frac{\sigma_k(B)}{\sqrt{k(n-k)+1}}, \tag{5}$$

$$\sigma_{\max}(R_{22}) \leq \sqrt{k(n-k)+1}\,\sigma_{k+1}(B). \tag{6}$$

**Proof** Let $B = U\Sigma V^{\mathrm{T}}$ be its singular value decomposition [8, p.70], where $\Sigma = \mathrm{diag}(\sigma_1(B), \sigma_2(B), \ldots, \sigma_n(B))$. Partition

$$\Pi^{\mathrm{T}} V = \Pi^{\mathrm{T}}[V_1,\, V_2] = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix},$$

where $V_1$ is $n \times k$ and $V_{11}$ is $k \times k$, and the sizes of other blocks are determined accordingly. Then [10, Theorem 1.5] (see also [8, Theorem 12.2.1]) says

$$\sigma_{\min}(R_{11}) \geq \sigma_{\min}(V_{11}) \cdot \sigma_k(B),$$
$$\sigma_{\max}(R_{22}) \leq [\sigma_{\min}(V_{11})]^{-1} \cdot \sigma_{k+1}(B).$$

Therefore it suffices to show that there is a permutation matrix $\Pi$ such that

$$\sigma_{\min}(V_{22}) \equiv \sigma_{\min}(V_{11}) \geq [k(n-k)+1]^{-1/2}. \tag{7}$$

Such a permutation does exist. In fact by the proof of [4, Theorem 2], there is a $\Pi$ whose construction will be outlined in the following remark such that

$$k(n-k+1) \geq \|V_{11}^{-1}\|_{\mathrm{F}}^2 \geq \frac{1}{[\sigma_{\min}(V_{11})]^2} + k - 1,$$

which leads to (7). □

**Remark 1** The permutation matrix $\Pi$ in Lemma 1 can be constructed at the cost of no more than $O(n^2k^2)$ if $V_1$ is known and $O(n^2(n-k)^2)$ if $V_2$ is known. Assume that $V_1$ is available. The construction goes by deleting $n-k$ rows from $V_1$, one at a time, and at the end the remaining $k$ rows give the desired $V_{11}$. Let $W$ be $V_1$ after $n-\ell$ rows deleted, where $\ell > k$. We now look for the next row to delete. This next row is the solution to

$$\arg\min_{w} \text{trace}((W^{\mathrm{T}}W - w^{\mathrm{T}}w)^{-1}), \tag{8}$$

where the minimization is taken over all $\ell$ rows $w$ of $W$. This idea is due to the constructive proof of [4, Theorem 2]. We now comment on how (8) should be solved. A straightforward way would involve inverting $\ell$ matrices $W^{\mathrm{T}}W - w^{\mathrm{T}}w$. This is too costly. A better way is to just invert $W^{\mathrm{T}}W$ and then use Sherman-Morrison formula [6] to get

$$\begin{aligned} &\left(W^{\mathrm{T}}W - w^{\mathrm{T}}w\right)^{-1} \\ &= \left(W^{\mathrm{T}}W\right)^{-1} + \frac{\left(W^{\mathrm{T}}W\right)^{-1} w^{\mathrm{T}}w \left(W^{\mathrm{T}}W\right)^{-1}}{1 - w\left(W^{\mathrm{T}}W\right)^{-1} w^{\mathrm{T}}} \end{aligned} \tag{9}$$

which gives

$$\begin{aligned} &\text{trace}((W^{\mathrm{T}}W - w^{\mathrm{T}}w)^{-1}) \\ &= \text{trace}((W^{\mathrm{T}}W)^{-1}) + \frac{\left\| \left(W^{\mathrm{T}}W\right)^{-1} w^{\mathrm{T}} \right\|^2}{1 - w\left(W^{\mathrm{T}}W\right)^{-1} w^{\mathrm{T}}}. \end{aligned}$$

Its significant parts in computation are 1) forming $W^{\mathrm{T}}W$, 2) inverting $W^{\mathrm{T}}W$, and 3) solving linear systems to get $\left(W^{\mathrm{T}}W\right)^{-1} w^{\mathrm{T}}$ for each $w$. Therefore its cost[1] is $O((n-\ell)\ell^2 + k^3)$. Thus the cost overall to obtain $V_{11}$ and at the same time $\Pi$ is

$$\sum_{\ell=0}^{n-k+1} O((n-\ell)k^2 + k^3) = O(n^2k^2).$$

Similar argument works for the case when $V_2$ is available. When both $V_1$ and $V_2$ are available, one should work with the one having the least number of columns.

Making use of the RRQR bounds, we obtain the following result on the possible condition number of a principal submatrix of $M$.

**Theorem 1** *Let the eigenvalues of $M$ be $\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_n(M) \geq 0$. There exists an $(n-m)\times(n-m)$ principal submatrix $M_{22}$ of $M$ such that*

$$\kappa(M_{22}) \leq [m(n-m)+1]\frac{\lambda_1(M)}{\lambda_{n-m}(M)}. \tag{10}$$

[1] Careful implementation can reduce the cost somewhat for the first two parts by taking advantage of (9) for the next $\left(W^{\mathrm{T}}W\right)^{-1}$ through updating. Note also that there is no need to compute trace$((W^{\mathrm{T}}W)^{-1})$ in solving (8).

**Proof** Since $M$ is positive semidefinite, there is an $n \times n$ $B$ such that $M = B^{\mathrm{T}}B$. Let $B$ have an RRQR (4) satisfying (5) with $k = n - m$. Now notice

$$\begin{aligned}\Pi^{\mathrm{T}} M \Pi &= (B\Pi)^{\mathrm{T}}(B\Pi) = R^{\mathrm{T}}R \\ &= \begin{bmatrix} R_{11}^{\mathrm{T}}R_{11} & R_{12}^{\mathrm{T}}R_{12} \\ R_{12}^{\mathrm{T}}R_{12} & R_{12}^{\mathrm{T}}R_{12} + R_{22}^{\mathrm{T}}R_{22} \end{bmatrix}\end{aligned}$$

to see that $M$ has an $(n-m) \times (n-m)$ principle submatrix $M_{22} = R_{11}^{\mathrm{T}}R_{11}$ whose smallest eigenvalue is

$$[\sigma_{\min}(R_{11})]^2 \geq \left[\frac{\sigma_{n-m}(B)}{\sqrt{m(n-m)+1}}\right]^2 = \frac{\lambda_{n-m}(M)}{m(n-m)+1},$$

because $\lambda_{n-m}(M) = [\sigma_{n-m}(B)]^2$. The result follows by noting that $\lambda_{\max}(M_{22}) \leq \lambda_1(M)$. □

The above result shows that in order to have $M_{22}$ with small condition number, $\lambda_{n-m}$ should not be close to zero. For how large an $m$ is this possible? The following two examples illustrate the possibility.

EXAMPLE 1. This is an example for which, roughly speaking, $\kappa(M_{22})$ of any principal submatrix $M_{22}$ is at least as big as the right-hand side of (10), asides from the factor $[m(n-m)+1]$, and thus in general the bound (10) is nearly best possible. Consider one-dimensional manifold learning problem with data points sitting equidistantly on a straight line. Then the alignment matrix is given by [11, Example 4.4],

$$M = \frac{1}{6}\begin{bmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{bmatrix} = \frac{1}{6}B^{\mathrm{T}}B,$$

where $B = \begin{bmatrix} e_1 & T_{n-2} & e_n \end{bmatrix}$, $e_1$ and $e_n$ are the first and last column of the $n \times n$ identity matrix, and $T_k$ is the $k$-by-$k$ famous tridiagonal Toeplitz matrix with diagonal entries $-2$ and off-diagonal entries 1. This immediately implies that

$$\lambda_1(M) \geq \cdots \geq \lambda_{n-2}(M) > 0 = \lambda_{n-1}(M) = \lambda_n(M).$$

Deleting the first and last row and column of $M$ gives $\frac{1}{6}T_{n-2}^2$. By Cauchy's interlacing theorem [12], we have

$$6\lambda_i(M) \geq \lambda_i(T_{n-2}^2) \geq 6\lambda_{i+2}(M).$$

Interpret that $\lambda_i(\cdot) \equiv 0$ for $i$ bigger than the size of the matrix. Since the eigenvalues of $T_{n-2}^2$ are known exactly [11], we have for $2 \leq j \leq n-3$

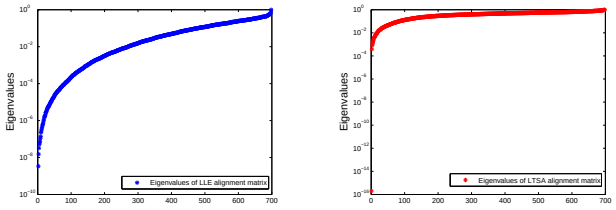
$$\begin{aligned}&\frac{8}{3}\left[\sin\frac{(j-1)\pi}{2(n-1)}\right]^4 = \frac{1}{6}\lambda_{n-j}(T_{n-2}^2) \\ &\leq \lambda_{n-j}(M) \leq \frac{1}{6}\lambda_{n-j-2}(T_{n-2}^2) = \frac{8}{3}\left[\sin\frac{(j+1)\pi}{2(n-1)}\right]^4.\end{aligned}$$

This, together with Theorem 1, imply that $M$ has an $(n-m)\times(n-m)$ principle submatrix $M_{22}$ for $2 \le m \le n-3$ such that

$$\kappa(M_{22}) \le [m(n-m)+1]\lambda_1(M)\frac{3}{8}\left[\sin\frac{(m+1)\pi}{2(n-1)}\right]^{-4}. \tag{11}$$

On the other hand for small $m$, $M_{22}$ obtained by deleting rows and corresponding columns of $M$ "evenly" has the condition number about at least as much as the bound (11) indicates, asides from the factor $[m(n-m)+1]$.

The bounds in this example show that we can reduce the condition number from $h^{-4}$ to $h^{-2}$ by labeling $\sqrt{n}$ data points, here $h = 1/n$.




(a) (b)

Figure 2: (a) Eigenvalues of LLE alignment matrix for face dataset; (b) Eigenvalues of LTSA alignment matrix for face dataset

.

EXAMPLE 2. We use the face data set in [16]. Figure 2 plots the eigenvalues of the alignment matrix for LLE and LTSA. We can see that the eigenvalues drops sharply, and Theorem 1 implies that deleting a small number of rows and columns of the alignment matrix will result in a matrix with moderate condition number.

# 5 Data Selection: Algorithm

Suppose we have computed an approximate eigenspace associated with $M$'s $m$ smallest eigenvalues. Namely we have an $n \times m$ $V$ such that its columns are orthonormal and span approximately the eigenspace. This can be done by, e.g., the symmetric Lanczos method [12] at cost about $O(mN)$, where $N$ is the number of nonzero entries in $M$. Next we look for a permutation matrix $\Pi$ so that the top $m \times m$ submatrix $V_{11}$ of $\Pi^{\mathrm{T}}V$ is well conditioned in the sense that its smallest singular value is at least modest. There are two ways to do this:

1. Perform the procedure outline in Remark 1. This is more expensive than the next one but guarantees that $\sigma_{\min}(V_{11}) \ge [m(n-m)+1]^{-1}$. Thus if $V$ is accurately computed, then the later selected $M_{22}$ will satisfy the inequality in Theorem 1.
2. Let $\Pi$ be the one arising from the QR with column pivoting $V^{\mathrm{T}}\Pi = QR$. This is less expensive than the previous one, but offers no guarantee, either. Nevertheless it usually works pretty well in practice.

Now let $\Pi = (e_{i_1}, e_{i_2}, \cdots, e_{i_n})$, where $e_i$ is the $i$th column of the identity matrix. Then $M_{22}$ can be gotten by deleting $M$'s rows $i_1, i_2, \ldots, i_m$ and the corresponding columns. This is because if we write $M = B^{\mathrm{T}}B$, then $V$'s column space is an approximate singular subspace associated with the $m$ smallest singular values of $B$.

The row and column index set of $M_{22}$ is $p = \{i_{m+1}, i_{m+2}, \cdots, i_n\}$, and the index set of the prior information points is complementary to $p$ with respect to the full index set of the sample points $\mathbb{X}$, denoted by $\bar{p}$. Our algorithm for the active learning of semi-supervised manifold learning can be described as follows.

**Algorithm 1** ACTIVELEARN:

1. Calculate the alignment matrix $M$ for LLE or LTSA from the sample data points $\mathbb{X}$;
2. Calculate the index set $p$ for a good principle submatrix of $M$ as just described;
3. Get the complementary index set $\bar{p}$ to $p$. This $\bar{p}$ is the index set of the prior points;
4. Query the prior information for the selected prior points, named it $Y_1$;
5. Calculate the low dimensional coordinates $Y$ by solving linear system (3).

According to the different alignment matrix in Step 1, we name our proposed active learning methods as: LLE-AE and LTSA-AE. Those based on the landmark point method are named as LLE/LTSA-landmark, and those based on random selection are named as LLE/LTSA-random.

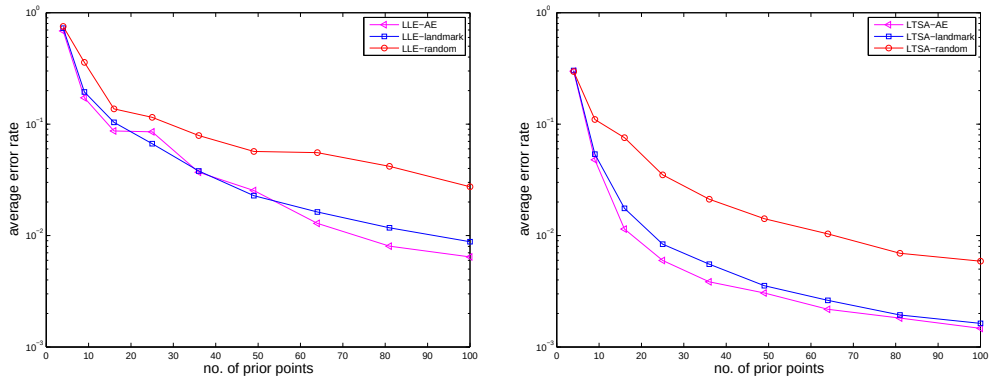
We now briefly comment on the computational complexity of our algorithm and the landmark point method since the random selection idea is not really an intelligent choice. Both need to construct the alignment matrix $M$ and solve $M_{22}Y_2^{\mathrm{T}} = M_{12}^{\mathrm{T}}Y_1^{\mathrm{T}}$, but differ only at the selection of prior information points. For Algorithm 1, getting $V$ costs $O(mN)$ as mentioned above; getting $\Pi$ can cost either $O(m^2n^2)$ by the more stable method with guarantee or $O(m^2n)$ by QR with column pivoting that usually works well. The landmark method, on the other hand, is a greedy algorithm which can take up to $O(n^2)$ operations while offers no mathematical guarantee to always produce well-conditioned $M_{22}$. So if $m$ is small and $N$ is nowhere near $n^2$ since $M$ is likely sparse, our algorithm with using QR with column pivoting for $\Pi$ could turn out to be much cheaper, while our algorithm with the most stable method for $\Pi$ costs at the same order as the landmark method and yet offers a mathematical guarantee.

# 6 Experimental Results

In this section, we present several experimental results comparing the various algorithms for data point selections. Since both of the landmark point method (i.e., for selecting the first landmark point) and the random point selection depend on random number generation, we average the corresponding results over 10 runs.

EXAMPLE 1. We use the "incomplete tire" data in Section 2. Figure 3 shows the average errors of LLE/LTSA-AE, LLE/LTSA-landmark and LLE/LTSA-random using 18 nearest neighbors when generating the alignment matrices. From the results, it is obvious that the results of LLE/LTSA-AE and LLE/LTSA-landmark are overall better

than LLE/LTSA-random. But the average errors of LLE/LTSA-AE and LLE/LTSA-landmark are comparable with LLE/LTSA-AE being slightly better.



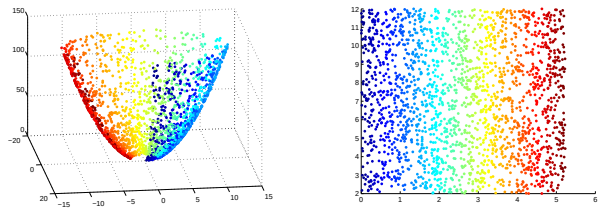

(a) LLE (b) LTSA

Figure 3: Average error rate with 18 nearest neighbors for "incomplete tire".

EXAMPLE 2. We sampled 2000 data points from the "paraboloid", generated by Matlab code:

```
 s = rand(1,N)*pi*5/3;
t = rand(1,N)*10+2;
X = [t.*cos(s);t.*sin(s);t.*t]
```

with the bottom and a strip cut out. shown in Figure 4(a). 2000 sample data points are generated by the following MATLAB code.



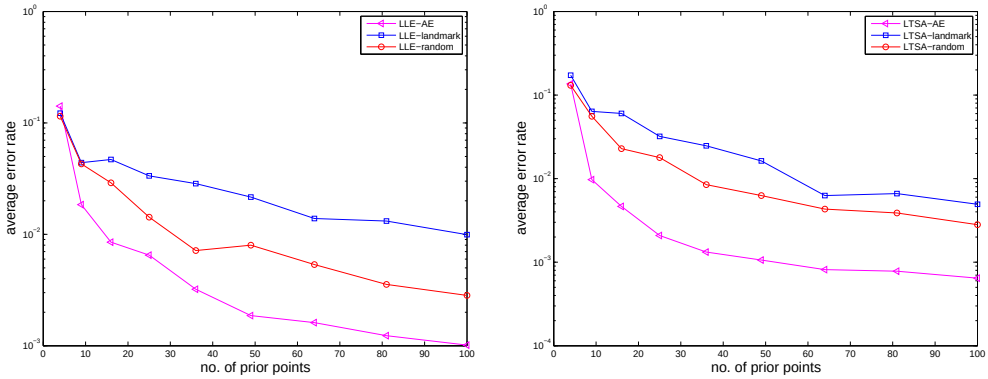
(a) (b)

Figure 4: (a) Paraboloid; (b) The generating coordinates of the paraboloid.

```
s = rand(1,N)*pi*5/3;
t = rand(1,N)*10+2;
X = [t.*cos(s);t.*sin(s);t.*t];
```

The generating coordinates are shown in Figure 4(b). The results of average errors are shown in Figure 5. In this example, the results of LLE/LTSA-landmark are even worse

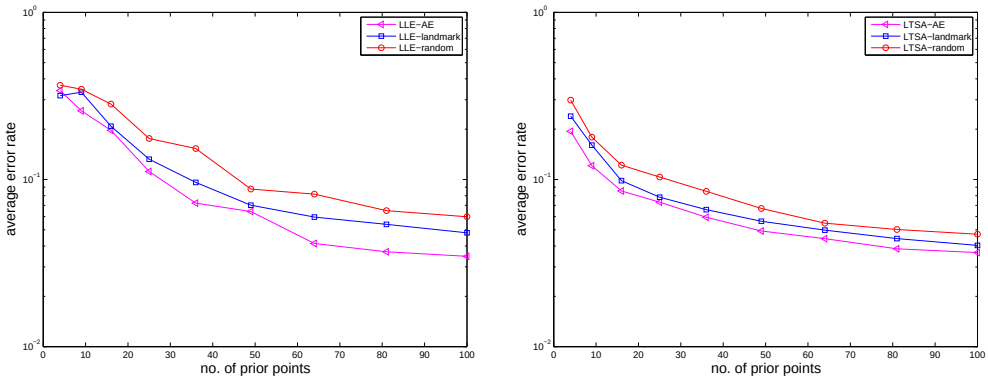
than those of LLE/LTSA-random. It is caused by highly non-isometric nature of the paraboloid, which destroyed the accuracy of the geodesic distance computation. But the proposed LLE/LTSA-AE works well regardless.




(a) LLE (b) LTSA

Figure 5: Average errors with 18 nearest neighbors for "paraboloid".

EXAMPLE 3. We consider the face data set which was used in Section 4. The desirable parameters are the pose of the statue as well as the lighting condition. Figure 6 shows the average error rates of LLE/LTSA-AE, LLE/LTSA-Landmark, and LLE/LTSA-random with 15 nearest neighbors. The results of our active learning algorithms here are prominently better than those of LLE/LTSA-landmark and LLE/LTSA-random.



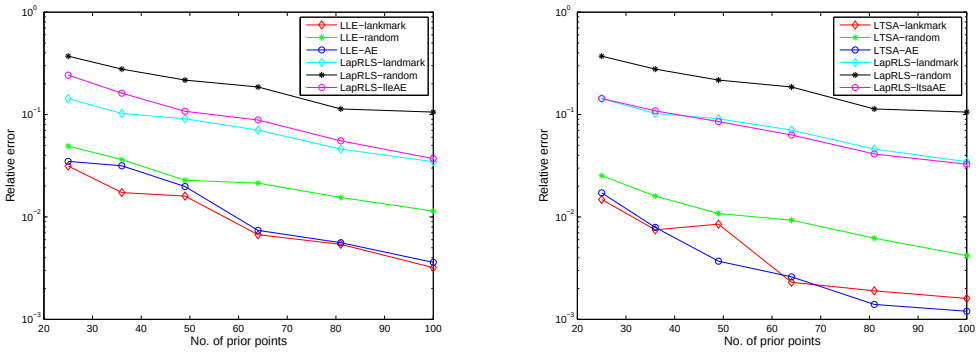

(a) LLE (b) LTSA

Figure 6: Average errors for Isomap face dataset with 15 nearest neighbors.

EXAMPLE 4. In this example, we applied our LLE-AE/LTSA-AE algorithms for prior points selection for Laplacian Regularized Least Squares (LapRLS), which is one of the semi-supervised learning methods derived from the semi-supervised geometric framework [1]. Even though LLE-AE/LTSA-AE is not specifically designed for

LapRLS, it still produces superior results than landmark point method and random selection.

1) We use the incomplete tire data again. In this experiment, we use LapRLS regression with different prior point selection methods: random selection, landmark point method and LLE-AE/LTSA-AE methods. Figure 7 demonstrates the semi-supervised LLE and LTSA give better results than LapRLS regression, the prior point selection methods are overall better than the random selection, and LTSA-AE gives better results than or comparable results to the landmark point method.




(a) SS-LLE compared with LapRLS

(b) SS-LTSA compared with LapRLS

Figure 7: Comparison of relative errors of the solutions for the incomplete tire data

2) We test the data selection methods on the semi-supervised classification problems. We use USPS dataset from [1], which is a collection of 2007 handwritten digit images. Each image is of size $16 \times 16$ gray corresponding to one of handwritten digits (between 0 and 9). For our test, we select 50 samples out of 2007 to label and apply LapRLS and compare it with their original results with random selection in [1]. Table 1 shows the results that LLE-AE/LTSA-AE do produce better accuracy for this multi-class semi-supervised classification problem.

| Prior point selection methods | Error rates (%) |
|---|---|
| Random[1] | 12.70 |
| LLE-AE (15NN) | 10.21 |
| LTSA-AE(30NN) | 9.96 |
| landmark (11NN) | 11.19 |

Table 1: USPS results of different prior point selection by LapRLS

EXAMPLE 5. In this video annotation example, we use the dataset from [13], which shows a subject moving his arms. Different from what was done in [13], we do not use the temporal information in the video frames. We choose the first 1000 frames from the video sequence, and manually determine the locations of the subject's wrists and

elbows as the parameter vectors that label the frames. We set the intrinsic dimension to be 8. Figure 8 shows the 10 selected frames by LTSA-AE. The $z$-axis is the frame number ordered along the temporal dimension, and the locations of the wrists and elbows are shown in the $xy$-plane for each frame. Figure 9 shows the average error rate as the number of prior points increases, LTSA-AE is slightly better than LTSA-landmark in this example.

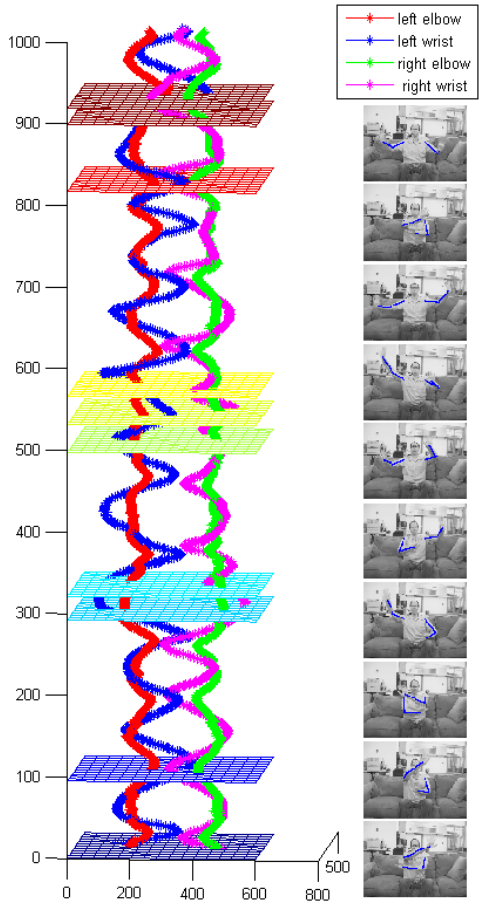



Figure 8: 10 selected frames by LTSA-AE for the video annotation example

We have also tried several other examples, our general conclusions are LLE-AE/LTSA-AE tends to have an edge over the landmark point method when the underlying manifold is close to being isometric, and landmark point method tends to perform even worse than random selection when the manifold is far from being isometric. Overall LLE-AE/LTSA-AE performs consistently better than both landmark and random

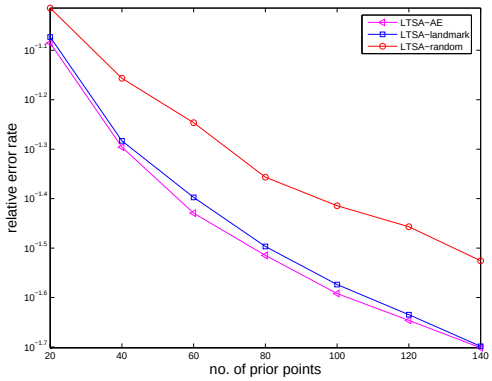



Figure 9: LTSA results with 50 nearest neighbors for the video annotation example

selection.

# 7 Conclusions and Remarks

In this paper, we proposed a method for selecting data points for labeling in semi-supervised manifold learning. Our method is based on selecting a submatrix from the alignment matrix with moderate condition number. We also used both synthetic and real-world problems to show that our proposed method can substantially improve the accuracy of the computed global parameterizations over existing methods. In future work, we plan to analyze LLE-AE/LTSA-AE in terms of its geometric implications, in particular, we observed that LLE-AE/LTSA-AE tends to first select points on the boundary of the parameter domains. We also plan to develop active learning algorithms for other semi-supervised manifold learning algorithm which better take advantage of the special structure of the semi-supervised manifold learning algorithms.

# References

[1] M. BELKIN, P. NIYOGI, AND V. SINDHWANI, *Manifold regularization: A geometric framework for learning from labeled and unlabeled examples.*, Journal of Machine Learning Research, 7 (2006), pp. 2399–2434.

[2] I. B.RAYTCHEV AND K.SAKAUE, *Head pose estimation by nonlinear manifold learning*, ICPR, (2004).

[3] T. F. CHAN, *Rank revealing qr factorizations*, Linear Algebra Appl., 88/89 (1987), pp. 67–82.

[4] F. R. DE HOOG AND R. M. M. MATTHEIJ, *Subset selection for matrices*, Linear Algebra Appl., 422 (2007), pp. 349–359.

[5] V. DE SILVA AND J. B. TENENBAUM, *Sparse multidimensional scaling using landmark points*, tech. report, Stanford Mathematics Technical Report, 2004.

[6] J. DEMMEL, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[7] L. V. FOSTER, *Rank and null space calculations using matrix decompositions without column interchanges*, Linear Algebra Appl., 74 (1986), pp. 47–71.

[8] G. H. GOLUB AND C. F. V. LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, 3rd edition ed., 1996.

[9] J. HAM, D. LEE, AND L. SAUL, *Semisupervised alignment of manifolds*, AISTATS, (2004).

[10] Y. P. HONG AND C. T. PAN, *Rank-revealing QR factorizations and the singular value decomposition*, Math. Comp., 58 (1992), pp. 213–232.

[11] C. LI, R. LI, AND Q. YE, *Eigenvalues of an alignment matrix in nonlinear manifold learning*, Communications in Mathematical Sciences, 5 (2007), pp. 313–329.

[12] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1998.

[13] A. RAHIMI, B. RECHT, AND T. DARRELL, *Learning appearance manifolds from video*, CVPR, (2005).

[14] S. T. ROWEIS AND L. K. SAUL, *Nonlinear dimensionality reduction by locally linear embedding*, Science, 90 (2000), pp. 2323–2326.

[15] L. K. SAUL AND S. T. ROWEIS, *Think globally, fit locally: Unsupervised learning of low dimensional manifolds*, Journal of Machine Learning Research, 4 (2003), pp. 119–155.

[16] J. B. TENENBAUM, V. DE SILVA, AND J. C. LANGFORD, *A global geometric framework for nonlinear dimension reduction*, Science, 90 (2000), pp. 2319–2323.

[17] X. YANG, H. FU, H. ZHA, AND J. BARLOW, *Semi-supervised nonlinear dimensionality reduction*, Proceedings of the 23rd international conference on Machine Learning, (2006).

[18] H. ZHA AND Z. ZHANG, *Spectral analysis of alignment in manifold learning*, IEEE International Conference on Acoustics, Speech, and Signal Processing, (2005).

[19] Z. ZHANG AND H. ZHA, *Principal manifolds and nonlinear dimension reduction via local tangent space alignment*, SIAM J. Scientific Computing, 26 (2004), pp. 313–338.