Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Active Manifold Learning via Gershgorin Circle Guided Sample Selection

Hongteng Xu[1], **Hongyuan Zha**[2,3], **Ren-Cang Li**[4], **Mark A. Davenport**[1]
[1]School of ECE, Georgia Institute of Technology, Atlanta, GA, USA
[2]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[3]Software Engineering Institute, East China Normal University, Shanghai, China
[4]Department of Mathematics, University of Texas at Arlington, Arlington, TX, USA

Abstract

In this paper, we propose an interpretation of active learning from a pure algebraic view and combine it with semi-supervised manifold learning. The proposed active manifold learning algorithm aims to learn the low-dimensional parameter space of the manifold with high accuracy from smartly labeled samples. We demonstrate that this problem is equivalent to a condition number minimization problem of the alignment matrix. Focusing on this problem, we first give a theoretical upper bound for the solution. Then we develop a heuristic but effective sample selection algorithm with the help of the Gershgorin circle theorem. We investigate the rationality, the feasibility, the universality and the complexity of the proposed method and demonstrate that our method yields encouraging active learning results.

Introduction

Many learning tasks need to model high-dimensional data via a low-dimensional parametric model. Manifold learning provides us with an approach for finding the low-dimensional structure of data from high-dimensional observations, which has yielded encouraging results in a number of unsupervised learning tasks (Roweis and Saul 2000; Belkin and Niyogi 2003; Zhang and Zha 2005; Zhang, Wang, and Zha 2012) and semi-supervised learning tasks (De Silva and Tenenbaum 2004; Ham, Ahn, and Lee 2006; Yang et al. 2006; Zhang, Zha, and Zhang 2008).

However, one important issue has remained (at least not explicitly addressed): the success of semi-supervised manifold learning relies on the distribution of labeled samples in the sample space. Facing a large number of unlabeled samples, we can only label a small subset of them because of the limitations in money, time and other resources. *Specifying which samples to label to obtain the best learning results, as the essential problem of active learning, is still a challenging problem from the view point of manifold learning*.

In this paper, we propose an active manifold learning algorithm, combining active learning with semi-supervised manifold learning in a novel manner. Specifically, we unify various manifold learning algorithms into the same framework, which corresponds to an eigenvalue problem related to the alignment matrix. In this framework, we formulate the active learning problem as follows: *given the alignment matrix (which contains all the geometrical information of the manifold), delete L rows/columns (corresponding to label L samples) so that the remaining principal submatrix has the smallest condition number (the learning error is bounded well)*. We analyze the problem theoretically and propose an effective algorithm, which relaxed the original problem rationally and solve it based on a heuristic process.

Our algorithm is based on two key points. First, the logarithmic condition number of a normal symmetric matrix is equal to the difference between the maximum eigenvalue and the minimum one of the logarithmic function of the matrix. Based on this fact, the original condition number minimization problem is rewritten as an eigenvalue dynamic range minimization problem. Second, the Gershgorin circle theorem provides us with an upper bound for the dynamic range of eigenvalues. Deleting the rows/columns of matrix is then equal to deleting the corresponding Gershgorin circles. Applying a sequential strategy, we delete circles and the corresponding rows/columns one at a time, which approximately minimizes the corresponding condition number.

The contributions of our method include the following three points. 1) We come up with an algebraic interpretation of active learning, formulating active learning as a condition number minimization problem in semi-supervised manifold learning. 2) We give an upper bound for the solution based on results from matrix theory. 3) Focusing on the active manifold learning problem, we propose an effective sample selection method guided by the Gershgorin circle theorem. Comparing with existing methods, we show the superiority of the proposed method.

Related Work

Manifold Learning. Many manifold learning approaches have been proposed to find the intrinsic structure of data. Among them, the methods based on local analysis, e.g., the locally linear embedding (LLE) (Roweis and Saul 2000), the Laplacian Eigenmap (LE) (Belkin and Niyogi 2003) and the local tangent space alignment (LTSA) (Zhang and Zha 2005) are widely used. Based on the fact that each data point and its neighbors lie on a locally linear patch of the manifold, LLE finds the linear coefficients that reconstruct each data point from its neighbors and align them glob-

ally. LE describe the manifold by a Laplacian graph matrix, which ensures that similar samples have similar latent variables. Similarly, LTSA constructs an approximation for the tangent space at each data point, and aligns these tangent spaces to give the global coordinates of the data points. In the field of semi-supervised learning, semi-supervised manifold learning is proposed in (Yang et al. 2006; Zhang, Zha, and Zhang 2008). Moreover, manifold assumption has been widely used to regularize other models in semi-supervised learning (Belkin, Niyogi, and Sindhwani 2006), e.g., manifold regularized sparse coding (Zheng et al. 2011; Long et al. 2013) for image classification. However, the sample selection problem for semi-supervised manifold learning has not been addressed in these works.

Active Learning. The sample selection problem has been studied in the field of active learning for several years (Settles 2010). In (Beygelzimer, Dasgupta, and Langford 2009), the importance of data is measured, and a weighted active learning algorithm is proposed. In (Vijayanarasimhan, Jain, and Grauman 2010), a batch active learning algorithm for image and video recognition is proposed. In (Hu et al. 2013), active learning is achieved by neighborhood reconstruction. An active transfer learning method is proposed in (Zhao et al. 2013) for cross-system recommendation. These works only focus on classification, and they do not combine active learning with manifold learning. Additionally, none of the works above study active learning from the algebraic view. Two manifold-related active learning methods are the harmonic function method in (Zhu, Lafferty, and Ghahramani 2003) and the landmark method in (De Silva and Tenenbaum 2004). The harmonic function method is suitable for classification tasks while its extension for regression tasks is not proposed. The landmark method labels samples one at a time. Each new labeled sample maximizes the minimum geodesic distance to any of the existing labeled samples. In both of these two methods, the first labeled sample is chosen arbitrarily, so the stability is not guaranteed.

Note that our work is different from **domain adaptation** (Gong et al. 2012; Gong, Grauman, and Sha 2013). In domain adaptation, labeled source samples are given to guide the labeling of target samples. In our work, source samples are unavailable, we decide which target samples to be labeled according to their own information.

Background

Let $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N] \in \mathbb{R}^{D\times N}$ be a set of samples from a manifold $\mathcal{X}$ given as $\boldsymbol{x}_i = f(\boldsymbol{y}_i) + \boldsymbol{n}_i$, $i = 1, .., N$. Here $\boldsymbol{y}_i \in \mathbb{R}^d$ represents the unknown low-dimensional parameter ($d \ll D$) corresponding to $\boldsymbol{x}_i$, and $\boldsymbol{n}_i$ represents sampling noise. Assuming that the manifold is well-sampled, we can achieve manifold learning by various algorithms, e.g., ISOMAP, LLE, LE, LTSA, etc. As shown in (Yang et al. 2006; 2010), these algorithms can be formulated in a unified manner: construct a K-NN graph and then solve the following problem.

$$\min_{\boldsymbol{Y}} \text{tr}(\boldsymbol{Y}\boldsymbol{\Phi}\boldsymbol{Y}^T), \;\; s.t.\; \boldsymbol{Y}\boldsymbol{Y}^T = \boldsymbol{I},\; \boldsymbol{Y} = [\boldsymbol{y}_1, ..., \boldsymbol{y}_N], \quad (1)$$

where $\text{tr}(\cdot)$ is the trace operator. $\boldsymbol{\Phi}$ is the Laplacian graph matrix in LE or the alignment matrix in ISOMAP, LLE or LTSA, which is computed based on the K-NN graph. For convenience, we call it the alignment matrix in this paper.

Semi-supervised manifold learning extends basic manifold learning — parameter estimation depends on both geometric information (the alignment matrix) and semantic information (labels). Given $\boldsymbol{X} = [\boldsymbol{X}_l, \boldsymbol{X}_u]$, where $\boldsymbol{X}_l = [\boldsymbol{x}_1^{(l)}, .., \boldsymbol{x}_L^{(l)}]$ are labeled with parameters $\boldsymbol{Z}_l = [\boldsymbol{z}_1^{(l)}, .., \boldsymbol{z}_L^{(l)}]$, we want to determine $\boldsymbol{Z}_u = [\boldsymbol{z}_1^{(u)}, .., \boldsymbol{z}_{N-L}^{(u)}]$ for unlabeled samples $\boldsymbol{X}_u = [\boldsymbol{x}_1^{(u)}, .., \boldsymbol{x}_{N-L}^{(u)}]$. A direct way is the **Least Squares (LS)** method, minimizing the following objective function (Yang et al. 2006),

$$\text{tr}(\boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}^T) = \text{tr}\left([\boldsymbol{Z}_l, \boldsymbol{Z}_u]\begin{bmatrix}\boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12}\\ \boldsymbol{\Phi}_{12}^T & \boldsymbol{\Phi}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{Z}_l^T\\ \boldsymbol{Z}_u^T\end{bmatrix}\right), \quad (2)$$

where $\boldsymbol{Z}_l$ is known. It is equivalent to find the least squares solution for a linear system of equations, which will be discussed further in the next section.

Besides the direct way above, we can also achieve semi-supervised manifold learning by the **spectral** method (Zhang, Zha, and Zhang 2008). Instead of finding the mapping $f_{z\to x}$ directly, the spectral method considers two manifolds: $\mathcal{X} = h(\mathcal{Y})$ and $\mathcal{Z} = g(\mathcal{Y})$, which share the same latent space $\mathcal{Y}$, $\boldsymbol{y}_i \in \mathcal{Y}$. We assume that the mapping g: $\mathcal{Y} \to \mathcal{Z}$ is an affine transformation. Then, we learn the mapping h: $\mathcal{Y} \to \mathcal{X}$ by manifold learning algorithm, which is regularized by the label information. Specifically, the error between $\boldsymbol{Y}_l = [\boldsymbol{y}_1^{(l)}, .., \boldsymbol{y}_L^{(l)}]$ and the affine transformation of $\boldsymbol{Z}_L$ should be minimized. According to this constraint, we add a regularization term of $\boldsymbol{Y}_l$ to Eq. (1) and estimate $\boldsymbol{Y}$ by minimizing following objective function:

$$\begin{aligned}&\text{tr}(\boldsymbol{Y}\boldsymbol{\Phi}\boldsymbol{Y}^T) + \lambda\text{tr}(\boldsymbol{Y}_l\boldsymbol{G}\boldsymbol{Y}_l^T), \quad (3)\\ =&\text{tr}\left([\boldsymbol{Y}_l, \boldsymbol{Y}_u]\begin{bmatrix}\boldsymbol{\Phi}_{11}+\lambda\boldsymbol{G} & \boldsymbol{\Phi}_{12}\\ \boldsymbol{\Phi}_{12}^T & \boldsymbol{\Phi}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{Y}_l^T\\ \boldsymbol{Y}_u^T\end{bmatrix}\right),\end{aligned}$$

where $\boldsymbol{G}$ is the orthogonal projection whose null space is spanned by $[\boldsymbol{1}, \boldsymbol{Z}_l^T]$. After getting $\boldsymbol{Y}$, we learn the affine transformation between $\boldsymbol{Y}_l$ and $\boldsymbol{Z}_l$.

Besides manifold learning, **manifold based regularization** is widely used in classification problems, describing the structure of (labeled and unlabeled) samples (Zheng et al. 2011; Long et al. 2013). Take the graph regularized sparse coding model (GraphSC) in (Zheng et al. 2011) as an example, the objective function is

$$\underbrace{\|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{Y}\|_F^2 + \lambda_1\|\boldsymbol{Y}\|_1}_{\text{sparse representation}} + \underbrace{\lambda_2\text{tr}(\boldsymbol{Y}\boldsymbol{\Phi}\boldsymbol{Y}^T)}_{\text{manifold regularization}}. \quad (4)$$

After getting $\boldsymbol{Y}$, various classifiers, e.g., SVM, Logistic regression, can be trained according to $\{\boldsymbol{Y}_l, \boldsymbol{Z}_l\}$.

The Algebraic Interpretation of Active Manifold Learning

Proposed Model

Without loss of generality, we can rewrite the objective function in Eq. (2) as follows,

$$\min_{\boldsymbol{Z}_u} 2\boldsymbol{Z}_l\boldsymbol{\Phi}_{12}\boldsymbol{Z}_u^T + \boldsymbol{Z}_u\boldsymbol{\Phi}_{22}\boldsymbol{Z}_u^T. \quad (5)$$

Similarly, Eq. (3) can be rewritten as follows,

$$\min_{\boldsymbol{Y}} \boldsymbol{Y}_l(\boldsymbol{\Phi}_{11}+\boldsymbol{G})\boldsymbol{Y}_l^T + 2\boldsymbol{Y}_l\boldsymbol{\Phi}_{12}\boldsymbol{Y}_u^T + \boldsymbol{Y}_u\boldsymbol{\Phi}_{22}\boldsymbol{Y}_u^T. \quad (6)$$

If the mapping g: $\mathcal{Y} \to \mathcal{Z}$ is an affine transformation, we can merely preserve the terms involving $\boldsymbol{Y}_u$ so that Eq. (6) is equivalent to Eq. (5). Here, the active learning problem arises: given limited resources, we need to select L samples from $\boldsymbol{X}$ and label them accordingly with $\boldsymbol{Z}_l$, such that the estimation error of $\boldsymbol{Y}_u$ (or $\boldsymbol{Z}_u$) is minimized.

Setting the gradient of the objective function Eq. (6) (or Eq. 5) with respect to $\boldsymbol{Y}_u$ (or $\boldsymbol{Z}_u$) to zero, we get

$$\boldsymbol{\Phi}_{22}\boldsymbol{Y}_u^T = \boldsymbol{\Phi}_{12}\boldsymbol{Y}_l^T. \quad (7)$$

The error in Eq. (7) only exists in $\boldsymbol{\Phi}_{12}$ and $\boldsymbol{\Phi}_{22}$, so we consider the parameterized system:

$$(\boldsymbol{\Phi}_{22}+\epsilon\boldsymbol{E}_2)\widehat{\boldsymbol{Y}}_u^T = (\boldsymbol{\Phi}_{12}+\epsilon\boldsymbol{E}_1)\boldsymbol{Y}_l^T. \quad (8)$$

Here $\boldsymbol{E}_1$, $\boldsymbol{E}_2$ are two noise matrices, ϵ bounds their energy. $\widehat{\boldsymbol{Y}}_u$ is the estimation of the ground truth $\boldsymbol{Y}_u$. According to a result in (Golub and Van Loan 2012), the relative estimation error satisfies

$$\frac{\widehat{\boldsymbol{Y}}_u - \boldsymbol{Y}_u}{\|\boldsymbol{Y}_u\|} \le \kappa(\boldsymbol{\Phi}_{22})|\epsilon|\left(\frac{\|\boldsymbol{E}_1\|}{\|\boldsymbol{\Phi}_{12}\|}+\frac{\|\boldsymbol{E}_2\|}{\|\boldsymbol{\Phi}_{22}\|}\right). \quad (9)$$

Here $\kappa(\boldsymbol{\Phi}_{22})$ is the condition number of $\boldsymbol{\Phi}_{22}$. The definition of condition number is $\kappa(\boldsymbol{\Phi}_{22}) = \|\lambda_{\max}/\lambda_{\min}\|$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and the smallest eigenvalues of $\boldsymbol{\Phi}_{22}$. We can find that the relative error is bounded by $\kappa(\boldsymbol{\Phi}_{22})$ times the relative errors in $\boldsymbol{\Phi}_{12}$ and $\boldsymbol{\Phi}_{22}$. It means that $\kappa(\boldsymbol{\Phi}_{22})$ decides the sensitivity of the estimation directly.

From this view, we formulate the active learning problem in a purely algebraic setting: given the alignment matrix $\boldsymbol{\Phi}$, delete L rows/columns so that the remaining principal submatrix $\boldsymbol{\Phi}_{22}$ has the smallest condition number. Let $\boldsymbol{v} \in \{0,1\}^N$ be the index vector for the rows/columns of $\boldsymbol{\Phi}$ corresponding to $\boldsymbol{\Phi}_{22}$, the active learning problem in manifold learning is solving following problem.

$$\min_{\boldsymbol{v}} \kappa(\boldsymbol{\Phi}(\boldsymbol{v},\boldsymbol{v})), \quad s.t.\ \|\boldsymbol{v}\|_0 = N-L. \quad (10)$$

Here $\boldsymbol{\Phi}(\boldsymbol{v},\boldsymbol{v})$ is the principal submatrix corresponding to the rows/columns indexed by $\boldsymbol{v}$. $\|\boldsymbol{v}\|_0$ counts the number of nonzero elements in $\boldsymbol{v}$ (the number of unlabeled samples).

Theoretical Bound for The Solution

There are many algorithms for condition number minimization (Greif and Varah 2006; Chen, Womersley, and Ye 2011). Unfortunately, all of these algorithms require the feasible domain to be continuous. In Eq. (10), however, the feasible domain is discrete, which contains $\binom{L}{N}$ possible solutions.

Eq. (10) can be solved approximately by the Rank-revealing QR-factorizations (RRQR) in (Hong and Pan 1992). Mathematically, given an $N \times N$ matrix $\boldsymbol{B}$ with singular values $\sigma_1(\boldsymbol{B}) \ge \sigma_2(\boldsymbol{B}) \ge \cdots \ge \sigma_N(\boldsymbol{B}) \ge 0$, if there is a reasonable gap between $\sigma_n(\boldsymbol{B})$ and $\sigma_{n+1}(\boldsymbol{B})$, and $\sigma_{n+1}(B)$ is small enough, it is reasonable to assume that $\boldsymbol{B}$ has a numerical rank n. In this case, any RRQR attempts to find a permutation matrix $\boldsymbol{\Pi}$ such that the QR factorization

$$\boldsymbol{B\Pi} = \boldsymbol{QR},\ \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{0} & \boldsymbol{R}_{22} \end{bmatrix},\ \boldsymbol{R}_{11} \text{ is } n \times n \quad (11)$$

satisfies $\boldsymbol{R}_{11}$'s smallest singular value $\sigma_{\min}(\boldsymbol{R}_{11}) \approx \sigma_n(\boldsymbol{B})$ and $\boldsymbol{R}_{22}$'s largest singular value $\sigma_{\max}(\boldsymbol{R}_{22}) \approx \sigma_{n+1}(\boldsymbol{B})$, where $\boldsymbol{Q}$ is orthogonal and $\boldsymbol{R}$ is upper triangular. In essence, $\boldsymbol{R}_{11}$ captures the well-conditioned part of $\boldsymbol{B}$. Readers can refer to (Hong and Pan 1992) for the details of RRQR.

An important property of RRQR is that there exists an RRQR such that

$$\sigma_{\min}(\boldsymbol{R}_{11}) \ge \frac{\sigma_n(\boldsymbol{B})}{\sqrt{n(N-n)+1}}. \quad (12)$$

Making use of this property, we obtain an upper bound for the condition number of a principal submatrix of $\boldsymbol{\Phi}$.

Theorem 1 *Let the eigenvalues of $\boldsymbol{\Phi}$ be $\lambda_1(\boldsymbol{\Phi}) \ge \lambda_2(\boldsymbol{\Phi}) \ge \cdots \ge \lambda_N(\boldsymbol{\Phi}) \ge 0$. There exists an $(N-L)\times(N-L)$ principal submatrix $\boldsymbol{\Phi}_{22}$ of $\boldsymbol{\Phi}$ such that*

$$\kappa(\boldsymbol{\Phi}_{22}) \le [L(N-L)+1]\frac{\lambda_1(\boldsymbol{\Phi})}{\lambda_{N-L}(\boldsymbol{\Phi})}. \quad (13)$$

Proof Since $\boldsymbol{\Phi}$ is positive semidefinite, there is an $N\times N$ $\boldsymbol{B}$ such that $\boldsymbol{\Phi} = \boldsymbol{B}^T\boldsymbol{B}$. Let $\boldsymbol{B}$ have an RRQR (11) satisfying (12) with $n = N-L$. Now notice

$$\begin{aligned}\boldsymbol{\Pi}^T\boldsymbol{\Phi}\boldsymbol{\Pi} &= (\boldsymbol{B\Pi})^T(\boldsymbol{B\Pi}) = \boldsymbol{R}^T\boldsymbol{R} \\ &= \begin{bmatrix} \boldsymbol{R}_{11}^T\boldsymbol{R}_{11} & \boldsymbol{R}_{12}^T\boldsymbol{R}_{12} \\ \boldsymbol{R}_{12}^T\boldsymbol{R}_{12} & \boldsymbol{R}_{12}^T\boldsymbol{R}_{12}+\boldsymbol{R}_{22}^T\boldsymbol{R}_{22} \end{bmatrix}\end{aligned}$$

to see that $\boldsymbol{\Phi}$ has an $(N-L)\times(N-L)$ principle submatrix $\boldsymbol{\Phi}_{22} = \boldsymbol{R}_{11}^T\boldsymbol{R}_{11}$ whose smallest eigenvalue is

$$[\sigma_{\min}(\boldsymbol{R}_{11})]^2 \ge \left[\frac{\sigma_{N-L}(\boldsymbol{B})}{\sqrt{L(N-L)+1}}\right]^2 = \frac{\lambda_{N-L}(\boldsymbol{\Phi})}{L(N-L)+1},$$

because $\lambda_{N-L}(\boldsymbol{\Phi}) = [\sigma_{N-L}(\boldsymbol{B})]^2$. The result follows by noting that $\lambda_{\max}(\boldsymbol{\Phi}_{22}) \le \lambda_1(\boldsymbol{\Phi})$. □

Proposed Active Manifold Learning

Reformulation Based on Logarithmic Transform

Although RRQR provides us with an upper bound for the solution of Eq. (10), the bound is too loose for practical application. We need to find a more practical method to solve this problem. The direct way to solve Eq. (10) is enumerating all the possible submatrices, which involves $\binom{L}{N}$ solving eigen-problems. Obviously, this method is impractical for the large-scale case. Here we come up with an alternative way, *reformulating this problem from minimizing condition number to minimizing the dynamic range of the eigenvalues*: the logarithmic version of the objective function in Eq. (10) is

$$\min_{\boldsymbol{v}} |\ln\lambda_{\max}(\boldsymbol{\Phi}(\boldsymbol{v},\boldsymbol{v})) - \ln\lambda_{\min}(\boldsymbol{\Phi}(\boldsymbol{v},\boldsymbol{v}))|. \quad (14)$$

Here, $\ln\lambda_{\max}$ and $\ln\lambda_{\min}$ are the largest and the smallest eigenvalues of $\ln(\boldsymbol{\Phi}(\boldsymbol{v},\boldsymbol{v}))$.

Inspired by Eq. (14), we transfer the optimization problem about $\mathbf{\Phi}$ to the one about $\ln \mathbf{\Phi}$: Given $\mathbf{\Lambda} = \ln \mathbf{\Phi}$, delete L rows/columns so that the remaining principal submatrix $\mathbf{\Lambda}(\boldsymbol{v}, \boldsymbol{v})$ has the narrowest dynamic range of eigenvalues. Formally, we solve the following optimization problem[1],

$$\min_{\boldsymbol{v}} |\lambda_{\max}(\mathbf{\Lambda}(\boldsymbol{v}, \boldsymbol{v})) - \lambda_{\min}(\mathbf{\Lambda}(\boldsymbol{v}, \boldsymbol{v}))|, \quad (15)$$
$$s.t. \quad \|\boldsymbol{v}\|_0 = N - L.$$

Sample Selection Guided by Gershgorin Circle

Replacing Eq. (10) with Eq. (15) does not simplify the problem directly — we still need to solve $\binom{L}{N}$ eigen-problems to attain the optimal $\boldsymbol{v}$. However, it allows us to further relax the problem and solve it very effectively with the help of the Gershgorin circle theorem (Geršgorin 1931).

Theorem 2 *For an $N \times N$ matrix $\boldsymbol{M}$, each eigenvalue satisfies: $|\lambda - m_{ii}| \leq \sum_{j \neq i} |m_{ij}|$ for $i = 1, .., N$. Here m_{ij}'s are the elements of $\boldsymbol{M}$. Let $r_i = \sum_{j \neq i} |m_{ij}|$, then the set $\mathcal{C}_i = \{x \in \mathbb{C} \mid |x - m_{ii}| \leq r_i\}$ is called the i^{th} Gershgorin circle of $\boldsymbol{M}$.*

According to the Gershgorin circle theorem, every eigenvalue of $\boldsymbol{M}$ must fall into the union of its Gershgorin circles $\bigcup_{i=1}^{N} \mathcal{C}_i$. As a result, the dynamic range of the eigenvalues is bounded by the boundary of the set $\bigcup_{i=1}^{N} \mathcal{C}_i$ — denoting $\lambda_l = \min\{m_{ii} - r_i\}_{i=1}^{N}$, $\lambda_u = \max\{m_{ii} + r_i\}_{i=1}^{N}$, we have

$$|\lambda_{\max} - \lambda_{\min}| \leq |\lambda_u - \lambda_l|. \quad (16)$$

Here we further relax the problem: *rather than minimizing the dynamic range of the eigenvalues, we minimize its upper bound.* So, the final problem we want to solve is

$$\min_{\boldsymbol{v}} |\lambda_u(\mathbf{\Lambda}(\boldsymbol{v}, \boldsymbol{v})) - \lambda_l(\mathbf{\Lambda}(\boldsymbol{v}, \boldsymbol{v}))|, \quad (17)$$
$$s.t. \quad \|\boldsymbol{v}\|_0 = N - L.$$

Eq. (17) is easy to solve because the influence of deleting matrix's rows/columns on the upper bound is reflected by Gershgorin circles directly and quantitatively. For example, if we delete the ith row/column of $\boldsymbol{M}$, in the complex plane of eigenvalues the ith Gershgorin circle will be deleted, the radius of the jth circle will be reduced to $r_j - |m_{ji}|$, and the upper bound $|\lambda_u - \lambda_l|$ will become tight accordingly. For minimizing the upper bound, we need to delete the rows/columns whose corresponding circles are far from the center of $\bigcup_{i=1}^{N} \mathcal{C}_i$ and have large radius. Consider the special case deleting 1 row/column ($L = 1$), the algorithm for circle deletion is shown below.

The principle of the algorithm above is simple. We first find the Gershgorin circles containing λ_l and λ_u. Then, we delete the circle (the row/column of matrix accordingly) having larger radius because after deleting it the rest of circles shrink more. For convenience, we denote $CDA(\cdot)$ as an operator, whose input is an matrix and output is the index of the deleted row/column.

[1]It should be noted that Eq. (15) is different from Eq. (10) because generally $\ln(\mathbf{\Phi})(\boldsymbol{v}, \boldsymbol{v}) \neq \ln(\mathbf{\Phi}(\boldsymbol{v}, \boldsymbol{v}))$. However, experimental results show that such a relaxation indeed improves the final learning results.

Circle Deletion Algorithm (CDA)
1. For $N \times N$ matrix $\boldsymbol{M}$, compute its Gershgorin circles $\{\mathcal{C}_i\}_{i=1}^{N}$.
2. The upper bounds and the lower bounds for all circles, denoted as $\boldsymbol{\alpha} = [m_{11} + |r_1|, ..., m_{NN} + |r_N|]^T$ and $\boldsymbol{\beta} = [m_{11} - |r_1|, ..., m_{NN} - |r_N|]^T$, respectively.
3. Find the minimum in $\boldsymbol{\alpha}$ and the maximum in $\boldsymbol{\beta}$, record the corresponding indices as a and b.
4. If $r_a \leq r_b$, delete $\mathcal{C}_b$ and the bth row/column from $\boldsymbol{M}$, else, delete $\mathcal{C}_a$ and the ath row/column from $\boldsymbol{M}$.

The Scheme of Algorithm

With the help of the Gershgorin circle theorem, we propose an effective active manifold learning algorithm, which smartly selects samples to label. In summary, given the alignment matrix, we select one sample at a time. In each iteration, the logarithmic function of the remaining submatrix is calculated, and one row/column is deleted by our circle deletion algorithm. Repeating the process above L times, the indices of the deleted rows/columns are the indices of the samples requiring labels.

Active Manifold Learning
Input: Sample set $\boldsymbol{X} \in \mathbb{R}^{D \times N}$;
The number of labeled samples we can select, L;
Algorithm:
1. Given the K-NN graph of $\boldsymbol{X} \in \mathbb{R}^{D \times N}$, compute the alignment matrix $\mathbf{\Phi}$ by a certain method.
2. Initialize the index set $\boldsymbol{v} = \{1, .., N\}$ for unlabeled samples, the index set $\boldsymbol{w} = \emptyset$ for labeled samples.
3. For $i = 1 : L$
   $\mathbf{\Lambda} = \ln(\mathbf{\Phi}(\boldsymbol{v}, \boldsymbol{v}))$.
   $l_i = CDA(\mathbf{\Lambda})$, $\boldsymbol{v} = \boldsymbol{v} \setminus \{l_i\}$, $\boldsymbol{w} = \boldsymbol{w} \cup \{l_i\}$.
4. Select $\boldsymbol{X}_l = \boldsymbol{X}(:, \boldsymbol{w})$ as the samples requiring labels. Label them with $\boldsymbol{Z}_l$.
Semi-supervised manifold learning:
apply manifold learning method to get $\boldsymbol{Z} = [\boldsymbol{Z}_l, \boldsymbol{Z}_u]$.
Manifold regularized learning task:
learn feature $\boldsymbol{Y}$ and train model by $\{\boldsymbol{Y}_l, \boldsymbol{Z}_l\}$.

Like the landmark method in (De Silva and Tenenbaum 2004) and the harmonic function based method in (Zhu, Lafferty, and Ghahramani 2003), our method is heuristic as well. However, our method is more deterministic because the first labeled sample is chosen definitely based on the numerical property of alignment matrix. Our algorithm involves computing a matrix logarithm iteratively, so in the worst case the computational complexity of our algorithm is $\mathcal{O}(LN^3)$, where L is the number of labeled samples and N is the total number of samples. Fortunately, when the alignment matrix is sparse[2], the computational complexity can be substantially reduced using Krylov-subspace based iterative methods (Al-Mohy and Higham 2012). On the other

[2]Using K-NN graph to model the geometrical information of samples, the alignment matrix can be sparse. In fact, the sparse situation is common when we use manifold assumption to regularize the training process of classifier.

hand, the landmark algorithm computes geodesic distance between samples, whose complexity is $\mathcal{O}(LN^3)$. The harmonic function based method compute inverse of matrix iteratively, whose complexity is $\mathcal{O}(LN^3)$ as well. As a result, our method has comparable complexity to the competitors and the performance of our algorithm is more stable.

Another advantage of our algorithm is its universality — as long as the learning model involves an alignment matrix, we can apply our sample selection method to improve the learning result. Because most of manifold learning algorithms and learning tasks regularized by manifold involves computing an alignment matrix, it is natural to introduce our algorithm into existing learning model.

Experiments

We test our active manifold learning algorithm on both regression and classification tasks. The active learning methods include: randomly labeling; (**Random**), two manifold related methods: the landmark method in (De Silva and Tenenbaum 2004) (**Landmark**) and the harmonic function method[3] in (Zhu, Lafferty, and Ghahramani 2003) (**Harmonic**); and our Gershgorin circle guided method (**GC**). The semi-supervised manifold learning algorithms for regression include the least squares method in (Yang et al. 2006) (**LS**) and the spectral method in (Zhang, Zha, and Zhang 2008) (**Spectral**). Like (Yang et al. 2006; Zhang, Zha, and Zhang 2008; Zhang, Wang, and Zha 2012), we compute the alignment matrix by LTSA. For classification, graph regularized sparse coding model (**GraphSC**, Eq. (4)) is applied. After learning sparse codes of samples, we label some samples by various active learning methods and train SVM classifier by labeled sparse codes.

Regression Case

The first data set we use is the face data from (Tenenbaum, De Silva, and Langford 2000). The data set contains $N = 698$ faces with size 64×64 ($D = 4096$) synthesized under various lighting conditions and poses ($d = 3$, 1 for lighting and 2 for pose), as shown in Fig. 1(c). Specifically, in each trial, we randomly select 600 samples and construct a K-NN graph, $K = 8$. The parameter of the spectral method is set to $\lambda = \frac{N}{L}$. We apply various algorithms to label $L = 10, 20, .., 100$ samples respectively, and then estimate the labels (parameters) for the rest of samples by semi-supervised manifold learning algorithms. Repeating the test 100 times in both noise-free and noisy cases (Gaussian noise with zero mean and variance $\sigma^2 = 0.01$), we obtain means of errors E's and variances of errors $\mathrm{Var}(E)$'s for various methods. Given the ground truth for the parameters of unlabeled samples $\boldsymbol{Z}_u$, the relative error E of the estimation result $\widehat{\boldsymbol{Z}}_u$ is measured as $E = \|\boldsymbol{Z}_u - \widehat{\boldsymbol{Z}}_u\|_F / \|\boldsymbol{Z}_u\|_F$.

Curves of $\log_{10}(E)$'s and $\log_{10}(\mathrm{Var}(E))$'s are shown in Fig. 1(a, b). Applying the sample selection algorithm reduces the mean and the variance of error greatly, and our method exhibits better performance than "Landmark" (De Silva and Tenenbaum 2004). In both the noise-free and noisy cases, the mean error obtained by our proposed method is better than that of "Landmark", which holds for various L's. Moreover, the variance obtained by our proposed method reduced rapidly in both noise-free and noisy cases. Although the variance obtained by "Landmark" is slightly better than ours when L is large and the samples are noise-free, in the noisy case "Landmark" is not stable — the variance does not always decay. The reason is that the randomness of the initial point leads to the instability of "Landmark", especially when L is small and the samples have noise. Table 1 further gives numerical results in the noise-free case.

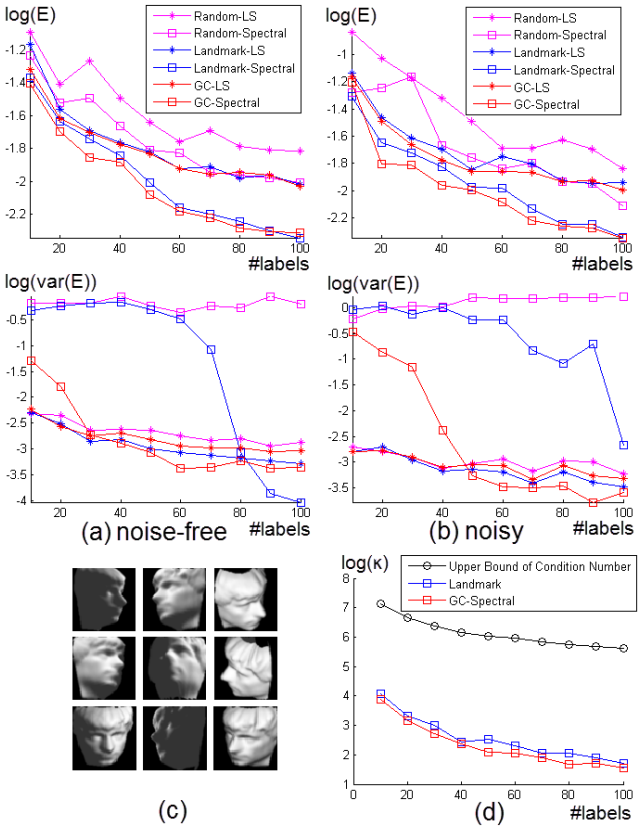

Figure 1: (a,b) Curves of $\log_{10}(E)$'s and $\log_{10}(\mathrm{Var}(E))$'s for various methods in clear and noisy cases. (c) Examples of data. (d) Curves of $\log_{10}(\kappa(\boldsymbol{\Phi}_{22}))$'s.

Fig. 1(d) provides the curves of $\log_{10}(\kappa(\boldsymbol{\Phi}_{22}))$'s with the increase of L. The condition number obtained by our method is not only much smaller than the theoretical bound but also smaller than that of "Landmark". According to Eq. (9), a small condition number leads to tight bound of error. This result also demonstrates the superiority of our method.

Another data set we used is from (Rahimi, Darrell, and Recht 2005), which shows a subject moving his arms. We choose 1200 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 construct a K-NN graph ($K = 50$) and pick the dimension d to be 8. Because the dimension of the image is very high,

[3]It is only applied for classification.

Table 1: Estimation Errors for Various Methods

Method\#labels	10	40	70	100
Random-LS	0.199	0.085	0.056	0.050
Random-Spectral	0.268	0.068	0.056	0.049
Landmark-LS	0.157	0.097	0.071	0.071
Landmark-Spectral	0.128	0.083	0.069	0.066
GC-LS	0.116	0.064	0.050	0.043
GC-Spectral (Proposed)	**0.085**	**0.060**	**0.048**	**0.041**

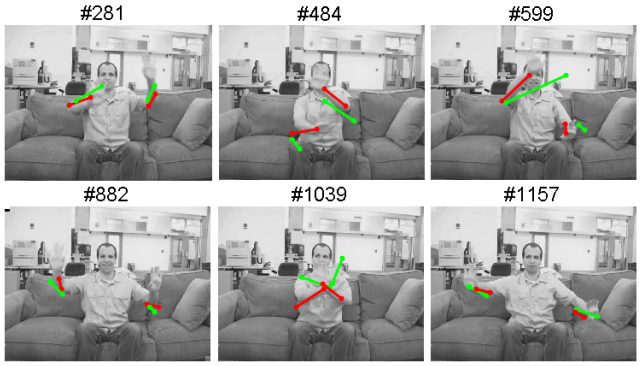

Figure 2: Comparison on tracking arms. In each subfigure, the index of frame is given at the top of image. The green arms correspond to Landmark-Spectral while the red ones correspond to GC-Spectral (Proposed).

we first apply PCA to original frames, reducing their dimension to $D = 500$ and treating these features as samples. $L = 60$ frames are labeled, which are selected by our method and "Landmark" respectively. Finally, applying the spectral method (Zhang, Zha, and Zhang 2008), we learn the locations of wrists and elbows in the unlabeled frames.

Fig. 2 illustrates our results. Compared with Landmark-Spectral, our method achieves better results: in the normal case where two arms do not obstruct with each other (i.e., the frames #281, #882, #1157), our algorithm is slightly better than Landmark-Spectral; in the challenging case where two arms overlap together (i.e., the frame #1039) or one arm is obstructed by hand (i.e., the frames #599), our algorithm achieves encouraging tracking results while Landmark-Spectral fails to learn the correct locations. The reason is that our sample selection algorithm makes a wiser selection than that of "Landmark". Some frames containing challenging cases are selected smartly to label.

Classification Case

Our active manifold learning method is also suitable for classification. The data sets we used include: 1) *The ARface data set* (Martinez 1998) contains over 4000 frontal view faces corresponding to 126 people's faces. We use a subset of the database consisting of 2600 images from 50 male subjects and 50 female subjects. 2) *The Extended YaleB data set* contains 2414 frontal face images of 38 persons. Following the settings in (Jiang, Lin, and Davis 2011), each face in the two data sets above is represented as a 540-dimensional random-face feature vector. 3) *The Caltech101 data set* (Fei-Fei, Fergus, and Perona 2007) contains 9144 images from 101 object classes and a "background" class.

Table 2: Comparison on Classification Accuracy (%)

ARface

Method\#labels	5/Class	10/Class	15/Class
Random-GraphSC	69.4	86.2	89.7
Harmonic-GraphSC	75.9	90.0	90.3
Landmark-GraphSC	76.5	90.2	90.7
GC-GraphSC (Proposed)	**79.7**	**92.7**	**93.9**

Extended YaleB

Method\#labels	5/Class	10/Class	15/Class
Random-GraphSC	56.2	68.3	78.6
Harmonic-GraphSC	62.8	83.1	88.2
Landmark-GraphSC	60.4	80.2	87.8
GC-GraphSC (Proposed)	**73.0**	**84.1**	**88.7**

Caltech101

Method\#labels	5/Class	10/Class	15/Class
Random-GraphSC	51.9	61.4	67.5
Harmonic-GraphSC	54.2	61.9	68.2
Landmark-GraphSC	53.3	62.1	68.3
GC-GraphSC (Proposed)	**56.0**	**64.2**	**69.5**

Applying GraphSC, we first learn dictionary and attain sparse codes of samples. The configuration of dictionary follows the set in (Jiang, Lin, and Davis 2011) as well. For the samples (codes) belonging to the same class, we select a part of them as training set by the "Random", the "Landmark", the "Harmonic" and our method respectively. Finally, given labeled samples, we train a SVM classifier. The classification results for various labeled samples per class are shown in Table 2. We find that our method improves classification accuracy greatly compared with the landmark method.

Further Analysis

Note that with the increase of selected samples the advantage of the proposed method seems to be limited in both regression and classification. This is the nature of active learning because the sufficiency of the selected sample can make up the robustness of sample selection. The more samples are labeled, the higher probability that they are sufficient to describe the sample space. In an extreme case, given a set of samples, if we require all samples to be labeled, our method will have no improvement because all active learning methods above select same samples to label. This phenomenon should not influence the evaluation of our method — our method obtains improved results over its alternatives when they label same number of samples, indeed, especially in the case labeling few samples and the case having noise.

Conclusions

We proposed a heuristic method to select samples for labeling in semi-supervised manifold learning. Our method can be viewed as an algebraic interpretation of active learning, which selects a principal submatrix from the alignment matrix with moderate condition number. In the future, we will further explore the theory for our method.

Acknowledgement: This work is supported in part by NSF grant DMS-1317424, DMS-1115834, CCF-1350616,

and CCF-1409261.

References

Al-Mohy, A. H., and Higham, N. J. 2012. Improved inverse scaling and squaring algorithms for the matrix logarithm. *SIAM Journal on Scientific Computing* 34(4):C153–C169.

Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* 15(6):1373–1396.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research* 7:2399–2434.

Beygelzimer, A.; Dasgupta, S.; and Langford, J. 2009. Importance weighted active learning. In *ICML*, 49–56. ACM.

Chen, X.; Womersley, R. S.; and Ye, J. J. 2011. Minimizing the condition number of a gram matrix. *SIAM Journal on Optimization* 21(1):127–148.

De Silva, V., and Tenenbaum, J. B. 2004. Sparse multidimensional scaling using landmark points. Technical report, Technical report, Stanford University.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2007. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding* 106(1):59–70.

Geršgorin, S. 1931. Uber die abgrenzung der eigenwerte einer matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na* 6:749–754.

Golub, G. H., and Van Loan, C. F. 2012. *Matrix computations*, volume 3. JHU Press.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2066–2073. IEEE.

Gong, B.; Grauman, K.; and Sha, F. 2013. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 222–230. ACM.

Greif, C., and Varah, J. M. 2006. Minimizing the condition number for small rank modifications. *SIAM Journal on Matrix Analysis and Applications* 29(1):82–97.

Ham, J.; Ahn, I.; and Lee, D. 2006. Learning a manifold-constrained map between image sets: applications to matching and pose estimation. In *CVPR*, volume 1, 817–824. IEEE.

Hong, Y. P., and Pan, C.-T. 1992. Rank-revealing qr factorizations and the singular value decomposition. *Mathematics of Computation* 58(197):213–232.

Hu, Y.; Zhang, D.; Jin, Z.; Cai, D.; and He, X. 2013. Active learning via neighborhood reconstruction. In *IJCAI*, 1415–1421. AAAI Press.

Jiang, Z.; Lin, Z.; and Davis, L. S. 2011. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *CVPR*. IEEE.

Long, M.; Ding, G.; Wang, J.; Sun, J.; Guo, Y.; and Yu, P. S. 2013. Transfer sparse coding for robust image representation. In *CVPR*, 407–414. IEEE.

Martinez, A. M. 1998. The ar face database. *CVC Technical Report* 24.

Rahimi, A.; Darrell, T.; and Recht, B. 2005. Learning appearance manifolds from video. In *CVPR*, volume 1, 868–875. IEEE.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Settles, B. 2010. Active learning literature survey. *University of Wisconsin, Madison* 52:55–66.

Tenenbaum, J. B.; De Silva, V.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323.

Vijayanarasimhan, S.; Jain, P.; and Grauman, K. 2010. Far-sighted active learning on a budget for image and video recognition. In *CVPR*, 3035–3042. IEEE.

Yang, X.; Fu, H.; Zha, H.; and Barlow, J. 2006. Semi-supervised nonlinear dimensionality reduction. In *ICML*, 1065–1072. ACM.

Yang, Y.; Nie, F.; Xiang, S.; Zhuang, Y.; and Wang, W. 2010. Local and global regressive mapping for manifold learning with out-of-sample extrapolation. In *AAAI*, volume 1, 649–654.

Zhang, Z., and Zha, H. 2005. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing* 26(1):313–338.

Zhang, Z.; Wang, J.; and Zha, H. 2012. Adaptive manifold learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34(2):253–265.

Zhang, Z.; Zha, H.; and Zhang, M. 2008. Spectral methods for semi-supervised manifold learning. In *CVPR*, 1–6. IEEE.

Zhao, L.; Pan, S. J.; Xiang, E. W.; Zhong, E.; Lu, Z.; and Yang, Q. 2013. Active transfer learning for cross-system recommendation. In *AAAI*.

Zheng, M.; Bu, J.; Chen, C.; Wang, C.; Zhang, L.; Qiu, G.; and Cai, D. 2011. Graph regularized sparse coding for image representation. *Image Processing, IEEE Transactions on* 20(5):1327–1336.

Zhu, X.; Lafferty, J.; and Ghahramani, Z. 2003. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, 58–65. ACM.