A Local Non-Negative Pursuit Method for Intrinsic Manifold Structure Preservation

Dongdong Chen and **Jian Cheng Lv** and **Zhang Yi**
Machine Intelligence Laboratory
College of Computer Science, Sichuan University
Chengdu 610065, P. R. China

Abstract

The local neighborhood selection plays a crucial role for most representation based manifold learning algorithms. This paper reveals that an improper selection of neighborhood for learning representation will introduce negative components in the learnt representations. Importantly, the representations with negative components will affect the intrinsic manifold structure preservation. In this paper, a local non-negative pursuit (LNP) method is proposed for neighborhood selection and non-negative representations are learnt. Moreover, it is proved that the learnt representations are sparse and convex. Theoretical analysis and experimental results show that the proposed method achieves or outperforms the state-of-the-art results on various manifold learning problems.

Introduction

Manifold learning is generally to learn a data representation which can uncovers the intrinsic manifold structure. It is extensively used in machine learning, pattern recognition and computer vision (Wright et al. 2009; Lv et al. 2009; Liu, Lin, and Yu 2010; Cai et al. 2011).

To learn a data representation, the neighbors of a data point should be selected in advance. There are many neighborhood selection methods, which can be divided into two categories: K nearest neighbors (KNN) methods and ℓ_1 norm minimization methods. Accordingly, the representation learning is divided into: KNN-based learning algorithms, such as Locally Linear Embedding (LLE, (Roweis and Saul 2000)), Laplacian eigenmaps (LEM, (Belkin and Niyogi 2003)); and ℓ_1 based learning algorithms, such as Sparse Manifold Clustering Embedding (SMCE, (Wright et al. 2009)).

However, the Knn method is heuristic. It is not easy to select proper neighbors of a data point in practical applications. On the other hand, the working mechanism of ℓ_1 based methods has not been fully elucidated (Zhang, Yang, and Feng 2011). The solution of ℓ_1 norm minimization does't indicate the space distribution feature of the samples.

More importantly, the representations learnt by these algorithms cannot avoid the existence of negative components. The representations with negative components cannot correctly reflect the essential relations between data pairs. The intrinsic structure of the data would be broken. Hoyer (Hoyer 2002) proposed Non-negative Sparse Coding (NSC) to learn sparse non-negative representations; recently, Zhuang *et al*. (Zhuang et al. 2012) also proposed non-negative low rank and sparse representations for semi-supervised learning. Unfortunately, the following concerns haven't been discussed. Why do the negative components of learnt representation exist and what is the influence of it on the intrinsic manifold structure preservation?

In this paper, we firstly reveals that an improper neighborhood selection will result in the existence of negative components of learnt representations. It is illustrated that the representations with negative components will destroy the intrinsic manifold structure. In order to avoid the existence of negative components and well preserve the intrinsic structure of the data, a local non-negative pursuit (LNP) is proposed to select the neighbors and learn the non-negative representations. The selected neighbors construct a convex set so that non-negative affine representations are learnt. Further, we have proves the representations are sparse and non-negative, which are useful to manifold dimensionality estimation and intrinsic manifold structure preservation. Theoretical analysis and experimental results show that the proposed method achieves or outperforms the state-of-the-art results on various manifold learning problems.

Notations and Preliminaries

$\mathcal{A} = \{\mathbf{a}_i \in \mathbb{R}^m\}_{i=1}^n$ is the data set, which lies on a manifold $\mathcal{M}$ of intrinsic dimension $d(\ll m)$. $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n]$ is the matrix form of $\mathcal{A}$. Generally, the representation learning for a manifold involves to solve the following problem:

$$\mathbf{A} = \mathbf{AX}, \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{n\times n}$ is the representation matrix of $\mathcal{A}$. Accordingly, $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in}]^\top$ is the representation of $\mathbf{a}_i \in \mathcal{A}$ $(i = 1, 2, \cdots, n)$. $\mathbf{X}$ should preserves the intrinsic structure of $\mathcal{M}$. With different purposes, various constraints could be added on $\mathbf{X}$ so that some particular representations can be obtained, such as sparse representation.

In this paper, three famous representation learning algorithms (LLE, LEM, and SMCE) will be used to compare

with our algorithm in various manifold learning problems. LLE learns the local affine representations with an assumption that the local neighborhood of a point on the manifold can be well approximated by the affine subspace spanned by the neighbors of the point. LEM learns the heat kernel representations that best preserves locality instead of local linearity in LLE. SMCE learns a sparse affine representation by solving an ℓ_1 minimization problem which can achieve datum-adaptive neighborhood.

Given a particular representation $\mathbf{X}$, the corresponding low-dimensional embeddings are obtained by solving the trace minimization problem below (Yan et al. 2007).

$$\min_{\mathbf{Y}\mathbf{Y}^\top=\mathbf{I}_d} \text{trace}(\mathbf{Y}\Phi(\mathbf{X})\mathbf{Y}^\top), \tag{2}$$

where $\mathbf{I}_d$ is an identity matrix and $\mathbf{Y} \in \mathbb{R}^{d\times n}$ is the final embedding of $\mathbf{A}$. $\Phi(\mathbf{X})$ is a symmetric and positive definite matrix w.r.t $\mathbf{X}$.

Four kinds of subspace w.r.t. $\mathcal{A}$ that used in this paper are defined as:

- non-negative subspace: $\mathcal{S}\{\mathcal{A}\}^+ = \{\sum_i c_i\mathbf{a}_i | \forall c_i \geq 0\}$
- negative subspace: $\mathcal{S}\{\mathcal{A}\}^- = \{\sum_i c_i\mathbf{a}_i | \forall c_i < 0\}$
- affine subspace: $\mathcal{S}_a\{\mathcal{A}\} = \{\sum_i c_i\mathbf{a}_i | \sum_i c_i = 1\}$
- convex set: $\mathcal{S}_a\{\mathcal{A}\}^+ = \{\sum_i c_i\mathbf{a}_i | \sum_i c_i = 1, \forall c_i \geq 0\}$

where c_i is the combinational coefficient w.r.t $\mathbf{a}_i$. In addition, $\mathbb{P}_{\mathcal{S}}$ and $\mathbb{P}_{\mathcal{S}_a}$ denotes the projection operator on the subspace $\mathcal{S}$ and $\mathcal{S}_a$, respectively.

Motivation

Three methods mentioned above and their extensions have been wildly studied and applied in many fields (He and Niyogi 2003; Lv, Zhang, and Kok Kiong 2007; Elhamifar, Sapiro, and Vidal 2012). The performance of these methods suffers due to negative components of the learnt representations, and the intrinsic manifold structure cannot be preserved well.

From Eqs.(1), $\mathbf{a}_i$ is the i-th column of $\mathbf{A}$, $\mathbf{x}_i$ is the i-th column of $\mathbf{X}$ ($i = 1, 2, \cdots, n$). The point $\mathbf{a}_i^* = \mathbf{A}\mathbf{x}_i$, is the reconstruction of $\mathbf{a}_i$. Suppose there are t non-zero components in $\mathbf{x}_i$, define a subset $\mathcal{A}_i = \{\mathbf{a}_{\lambda_j} | x_{i\lambda_j} \neq 0\}_{j=1}^t \subseteq \mathcal{A}$, where λ_j denotes the index of the j-th non-zero component in $\mathbf{x}_i$. Thus, $\mathbf{a}_i^*$ is a linear combination of points in $\mathcal{A}_i$, where the combination should involves both additive and subtractive interactions in terms of the components of $\mathbf{x}_i$.

Clearly, from the definition of $\mathcal{S}\{\cdot\}^+$, if $\mathbf{a}_i^* \notin \mathcal{S}\{\mathcal{A}_i\}^+$, then $\exists x_{ij} < 0 (j = 1, \cdots, n)$. The negative components will broke the intrinsic manifold structure. The following example will illustrate this problem.

Suppose $\mathcal{A} = \{\mathbf{a}_1 = (9.8, 15.4)^\top, \mathbf{a}_2 = (12.35, 13.70)^\top, \mathbf{a}_3 = (11.75, 8.2)^\top, \mathbf{a}_4 = (4.90, 1.95)^\top\}$ which is sampled from a 1-d manifold as shown in Figure 1(a). For $\forall \mathbf{a}_i \in \mathcal{A} (i = 1, 2, 3, 4)$, denotes $\mathcal{A}_i = \mathcal{A} \setminus \{\mathbf{a}_i\}$. LLE is used to learn the affine representation of $\mathcal{A}$ with the neighborhood size is three. The local affine representation $\mathbf{X}$ can be obtained as: $\mathbf{X} = [$ (0,1.7408,**-1.1470**,0.4062)$^\top$, (0.5541,0,0.6871,**-0.2412**)$^\top$, **(-0.7562**,1.4021,0,0.3541)$^\top$, (1.8100,**-3.4338**,2.6238,0)$^\top$].

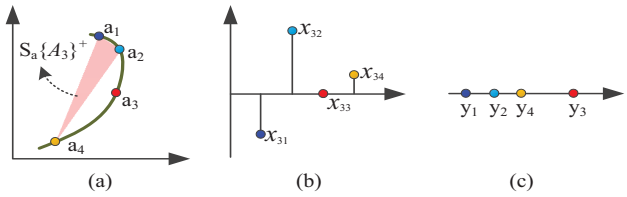


Figure 1: (a): Data manifold and $\mathcal{S}_a\{\mathcal{A}_3\}^+$; (b): The affine representation of $\mathbf{a}_3$ learnt by LLE; (c): The corresponding 1-d embedding of $\mathcal{A}$, denoted by $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}$.

It is easy to see that each $\mathcal{S}_a\{\mathcal{A}_i\}$ is the whole plane of $\mathbb{R}^2$, thus $\mathbf{a}_i \in \mathcal{S}_a\{\mathcal{A}_i\}$ and the reconstruction error is zero, i.e., $\|\mathbf{a}_i - \mathbf{A}\mathbf{x}_i\|_2 = 0$. However, $\forall \mathbf{a}_i \notin \mathcal{S}_a\{\mathcal{A}_i\}^+$. Thus, the corresponding representation $\mathbf{x}_i$ must contains negative components. Figure 1(b) shows the representation of $\mathbf{a}_3$.

By solving Eqs.(2) with $\Phi(\mathbf{X}) = (\mathbf{I} - \mathbf{X})^\top(\mathbf{I} - \mathbf{X})$, where $\mathbf{I}$ is the identity matrix, the corresponding 1-d embedding of $\mathcal{A}$ can be obtained (Figure 1(c)). It shows that the intrinsic manifold structure is changed.

Furthermore, non-negative representations are required for some clustering algorithms (Von Luxburg 2007). For example, spectral clustering algorithm requires to construct a non-negative graph weights matrix. Generally, let $\mathbf{X} = |\mathbf{X}|$. Clearly, the graph weights matrix $|\mathbf{X}|$ is generally not same with original $\mathbf{X}$, the weights in $|\mathbf{X}|$ would no longer reflect the inherent relationships between data points if there exist negative components in $\mathbf{X}$.

Clearly, an improper subset $\mathcal{A}_i$ will result in includes some negative components in the learnt representation so that the local manifold structure is broken. Generally, the subset $\mathcal{A}_i$ is determined according to the neighborhood of $\mathbf{a}_i$. Thus, to select a proper neighborhood and learn a non-negative representation is crucial to preserve the intrinsic manifold structure.

Local Non-negative Pursuit

Manifold assumption (Saul and Roweis 2003): If sufficient sampling can be obtained from the manifold, then the local structure of the manifold is linear which can be well approximated by the local affine subspace. Based on this assumption, we suppose that for each data point, there exists an optimal neighborhood which satisfies: 1) the size of this neighborhood is less than $d + 1$; 2) the corresponding affine projection of the point is located inside the non-negative affine subspace spanned by its neighboring points. More precisely:

***Assumption* 1** *Given a data set $\mathcal{A}$ that is sampled from a sufficient dense sampled manifold $\mathcal{M}$, for $\forall \mathbf{a}_i \in \mathcal{A}$, consider a linear patch $\mathcal{U} \subseteq \mathcal{A}$ that contains $K(\geq d + 1)$ nearest neighbors of $\mathbf{a}_i$. We assume that: $\exists \mathcal{A}_{opt} \subseteq \mathcal{U}$ such that $\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{opt}\}}\mathbf{a}_i \in \mathcal{S}_a\{\mathcal{A}_{opt}\}^+$, where $\mathcal{A}_{opt} \in \mathbb{R}^{m\times k}$, $k \leq d + 1$.*

Such $\mathcal{S}_a\{\mathcal{A}_{opt}\}^+$ is called as a sparse convex patch of $\mathcal{M}$ at $\mathbf{a}_i$. Generally, if $\mathbf{a}_i$ is a boundary point, $|\mathcal{A}_{opt}| < d + 1$, otherwise, $|\mathcal{A}_{opt}| = d + 1$ where $|.|$ is the set cardinality.

Clearly, $\mathcal{U}$ is the K nearest neighbors of $\mathbf{a}_i$. Rewrite $\mathcal{U} = \mathcal{A}_k \cup \overline{\mathcal{A}_k}$ where $\mathcal{A}_k$ is a set that contains any $k(\leq K)$ neighbors of $\mathbf{a}_i$ and $\overline{\mathcal{A}_k}$ is the complementary set of $\mathcal{A}_k$. Denote $\mathcal{G}_k = \{\mathbf{g}_t = \mathbf{a}_i - \mathbf{a}_t | \forall \mathbf{a}_t \in \mathcal{A}_k\}$, we have the following result.

***Theorem* 1** *For $\forall \mathbf{a}_j \in \overline{\mathcal{A}_k}$, if $\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i \in \mathcal{S}_a\{\mathcal{A}_k\}^+$ and $\mathbb{P}_{\mathcal{S}\{\mathcal{G}_k\}}\mathbf{g}_j \in \mathcal{S}\{\mathcal{G}_k\}^-$, it holds that:*

$$\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k+1}\}}\mathbf{a}_i \in \mathcal{S}_a\{\mathcal{A}_{k+1}\}^+.$$

where $\mathbf{g}_j = \mathbf{a}_i - \mathbf{a}_j$ and $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{\mathbf{a}_j\}$.

The proof can be found in Appendix A. According to the theorem, a Local Non-negative Pursuit (LNP) method is proposed to select the proper neighboring points from $\mathcal{U}$ one by one. The algorithm is as follows.

Algorithm 1 Find $\mathcal{A}_{opt}$: Local Non-negative Pursuit

1: Input: $\mathbf{a}_i \in \mathbb{R}^m, \mathcal{U} \in \mathbb{R}^{m\times K}$.
2: Initialize: $k = 0, \mathcal{A}_k, \mathcal{G}_k = \varnothing$.
3: Let $k = 1$, select the the nearest neighbor from $\mathcal{U}$ by

$$\mathbf{a}_{\lambda_1} = \arg\min_{\mathbf{a}_j \in \mathcal{U}} \|\mathbf{a}_i - \mathbf{a}_j\|_2. \tag{3}$$

4: Updating: $\mathcal{A}_1 = \{\mathbf{a}_{\lambda_1}\}, \mathcal{G}_1 = \{\mathbf{g}_{\lambda_1}\}, k = 2$.
5: **while** true

$$\begin{cases} \mathbf{a}_{\lambda_k} & = \arg\min_{\mathbf{a}_j \in \mathcal{U}} \|\mathbf{a}_i - \mathbf{a}_j\|_2, \\ \text{s.t.} & \mathbb{P}_{\mathcal{S}\{\mathcal{G}_{k-1}\}}\mathbf{g}_j \in \mathcal{S}\{\mathcal{G}_{k-1}\}^-. \end{cases} \tag{4}$$

6: **if** Eqs.(4) has solution, updating:

$$\begin{cases} \mathcal{A}_k & = \mathcal{A}_{k-1} \cup \{\mathbf{a}_{\lambda_k}\}; \\ \mathcal{G}_k & = \mathcal{G}_{k-1} \cup \{\mathbf{g}_{\lambda_k}\}; \\ k & = k + 1. \end{cases} \tag{5}$$

7: **else** Break.
8: **endwhile**
9: Output: $\mathcal{A}_{opt} = \mathcal{A}_k = \{\mathbf{a}_{\lambda_1}, \mathbf{a}_{\lambda_2}, \cdots, \mathbf{a}_{\lambda_k}\}$.

$\lambda_j (j = 1, 2, \cdots, k)$ above denotes the index of the j-th point that is selected by LNP. Clearly, the points in $\mathcal{A}_{opt}$ are from the local of $\mathcal{M}$ at $\mathbf{a}_i$ and are affine independent of each other, which is very useful properly for manifold learning.

The algorithm converges fast with computational complexity of $O(k(K-k)/2)$, where $k = |\mathcal{A}_{opt}|$ and $K = |\mathcal{U}|$. The following theorem describes the problem.

***Theorem* 2** *The LNP never selects the same data point twice. Thus, it must converges in no more than K steps.*

The proof can be found in Appendix A. The LNP will stopped when no appreciate neighbors can be selected. e.g. if no neighbor can be selected at step $t(< K)$, the algorithm will stopped since it never select the same point twice.

Clearly, $\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i$ is the affine reconstruction of $\mathbf{a}_i$ at step k. Denote $\mathbf{r}_k = \mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i$. The following theorem shows the reconstruction error is decreasing.

***Theorem* 3** *For $\forall k \geqslant 1$, $\|\mathbf{r}_k\|_2 < \|\mathbf{r}_{k-1}\|_2$.*

The proof can be found in Appendix A.

Based on the selected neighbors set $\mathcal{A}_{opt}$, a non-negative sparse affine representation of $\mathbf{a}_i$ is given by minimizing the following objective function:

$$\begin{cases} \min\limits_{\mathbf{x}_i} & \frac{1}{2}\|\mathbf{a}_i - \mathbf{A}\mathbf{x}_i\|_2^2, \\ \text{s.t.} & \sum\limits_{j=1}^{k} x_{i\lambda_j} = 1; \quad x_{it} = 0 \text{ if } \mathbf{a}_t \notin \mathcal{A}_{opt}. \end{cases} \tag{6}$$

Denote $\mathbf{G} = [\mathbf{a}_i - \mathbf{a}_{\lambda_1}, \mathbf{a}_i - \mathbf{a}_{\lambda_2}, \cdots, \mathbf{a}_i - \mathbf{a}_{\lambda_k}]$ and $\mathbf{M} = (\mathbf{G}^\top\mathbf{G})^{-1}$, where $\forall \mathbf{a}_{\lambda_j} \in \mathcal{A}_{opt} (j = 1, 2, \cdots, k)$. A closed form solution is:

$$\begin{cases} x_{i\lambda_j} = & \dfrac{\mathbf{1}^\top\mathbf{m}_j}{\mathbf{1}^\top\mathbf{M}\mathbf{1}}, \quad \text{for } \mathbf{a}_{\lambda_j} \in \mathcal{A}_{opt}; \\ x_{it} = & 0, \qquad \text{for } \mathbf{a}_t \notin \mathcal{A}_{opt}. \end{cases} \tag{7}$$

where $\mathbf{m}_j$ is the j-th column of $\mathbf{M}$ and $\mathbf{1} \in \mathbb{R}^k$ is the vector of all ones.

***Theorem* 4** *The representation $\mathbf{x}_i$ learnt by Eqs.(6) is sparse and convex.*

The proof can be found in Appendix A. Such a non-negative sparse affine representation is referred as Sparse Convex Representation (SCR) in this paper.

Experiments and Discussions

In this section, a series of synthetic and real-world data sets are used for the experiments. The proposed method is compared with the three methods, LLE, LEM and SMCE. All the experiments were carried out using MATLAB on a 2.2 GHz machine with 2.0GB RAM.

Manifold dimensionality estimation. By using the proposed method, the learnt SCRs are sparse with $\|x_i\|_0 \leq d + 1 (i = 1, \cdots, n)$. Based on the sparsity of SCR, an intrinsic dimensionality estimation method is given. We firstly sort the components of $\mathbf{x}_i$ in descending order and denote $\mathbf{x}'_i = [x'_{i1}, x'_{i2}, \cdots, x'_{in}]$ with $x'_{i1} \geq x'_{i2} \geq \cdots \geq x'_{in}$. Then, a manifold dimensionality estimation vector (***dev***) of $\mathcal{M}$ is defined as:

$$\textbf{\textit{dev}}(\mathcal{M}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}'_i = [\rho_1, \rho_2, \cdots, \rho_n]^\top. \tag{8}$$

Thus, the intrinsic dimension d of $\mathcal{M}$ can be determined by:

$$d = \arg\max_l \{\rho_l - \rho_{l+1}\}_{l=1}^{k-1} - 1. \tag{9}$$

Multi-manifold clustering and embedding. The SCRs of a data set can be learnt by using the proposed method. The SCRs, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{n\times n}$, can be used for multi-manifold clustering and embedding. Firstly, a similarity matrix $\mathbf{W} = \max(\mathbf{X}^\top, \mathbf{X})$ is constructed. Then, the normalized Laplacian matrix is calculated by $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is a diagonal matrix with the diagonal element $d_{ii} = \sum_{j=1}^{n} w_{ij}$. Finally, the standard spectral clustering (Von Luxburg 2007) is used to obtain the n_c clusters, denoted by $\{\mathcal{M}_i\}_{i=1}^{n_c}$. For each cluster $\mathcal{M}_i$, the intrinsic dimension d_i is estimated by Eqs.(9) and the d_i-dimension embedding is obtained by Eqs.(2).

Evaluate metric. It is important to preserve the intrinsic structure of a manifold after embedding. Some method, such as MMRE (Lee and Verleysen 2009) and LCMC (Chen and Buja 2009) have been proposed to assess the intrinsic structure preservation by computing the neighborhood overlap. MMRE not only computes the neighborhood alignment accuracy, but also considers the order of the neighbors. Based on the MMRE method, this paper assesses the quality of the intrinsic structure preservation with different method, LLE, LEM, SMCE and our method.

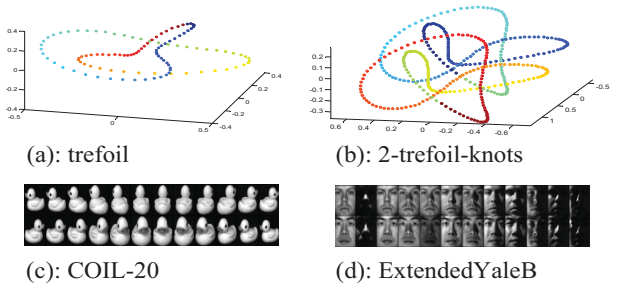
For the data set $\mathcal{A} = \{\mathbf{a}_i\}_{i=1}^n$, its corresponding low-dimensional embedding is denoted by $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^n$. Clearly, $\mathbf{y}_i$ is the embedding of $\mathbf{a}_i$. $\Lambda_i = \{\lambda_{ij}\}_{j=1}^{s_i}$ is the indices set of the selected s_i neighbors of $\mathbf{a}_i$ with $\|\mathbf{a}_i - \mathbf{a}_{\lambda_{ij}}\|_2 \leq \|\mathbf{a}_i - \mathbf{a}_{\lambda_{i,j+1}}\|_2 (j = 1, \cdots, s_i)$. $\Gamma_i = \{\gamma_{ij}\}_{j=1}^{t_i}$ is the indices set of the selected t_i neighbors of $\mathbf{y}_i$ with $\|\mathbf{y}_i - \mathbf{y}_{\gamma_{ij}}\|_2 \leq \|\mathbf{y}_i - \mathbf{y}_{\gamma_{i,j+1}}\|_2 (j = 1, \cdots, t_i)$. The neighborhood alignment accuracy (*nalac*) is given as:

$$nalac \triangleq \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{\min(s_i,t_i)} \frac{\delta(\lambda_{ij} = \gamma_{ij})}{\min(s_i, t_i)}, \tag{10}$$

where $\delta(\cdot)$ is the delta function. Another average alignment accuracy is formulated as follows:

$$soft - nalac \triangleq \frac{1}{n} \sum_{i=1}^{n} \frac{|\Lambda_i \cap \Gamma_i|}{\max(s_i, t_i)}. \tag{11}$$

It is clear that the bigger the value of *nalac* (*soft-nalac*), the change of neighborhood is smaller. Thus, a bigger value of *nalac* (*soft-nalac*) indicates a better behaviour of intrinsic manifold structure preservation .



(a): trefoil (b): 2-trefoil-knots

(c): COIL-20 (d): ExtendedYaleB

Figure 2: The employed four data sets.

Synthetic Data. 110 data points are sampled from a trefoil that is embedded in $\mathbb{R}^{100}$ with small Gaussian white noises (Figure 2(a)). LLE, LEM, SMCE and LNP scheme are used to learn a low dimensional manifold with $K = 2, 3, \cdots, 100$ and $\lambda = \frac{60}{2}, \frac{60}{3}, \cdots, \frac{60}{100}$ for SMCE. Figure (6) a reports the corresponding time costs. Figure 3 shows the final embedding results and the corresponding ***dev***s. It is shown that the SCRs learnt by our method are always non-negative and sparse. The ***dev***s of our method always reflect the intrinsic dimension of the data. For different neighborhood sizes K (λ for SMCE), only our method can always provide a good embedding result.

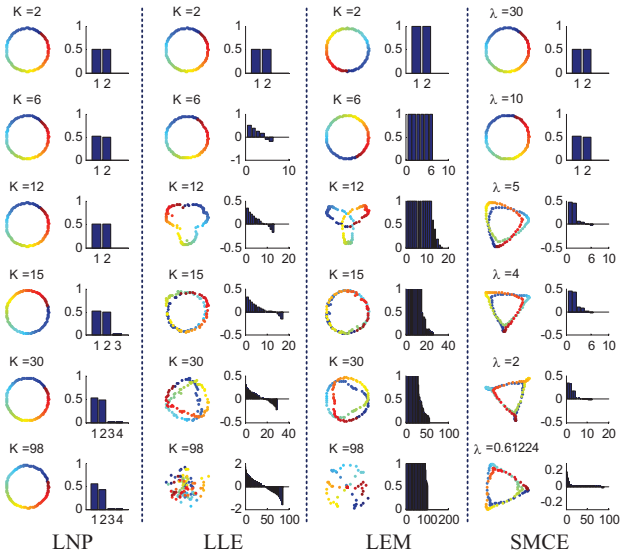
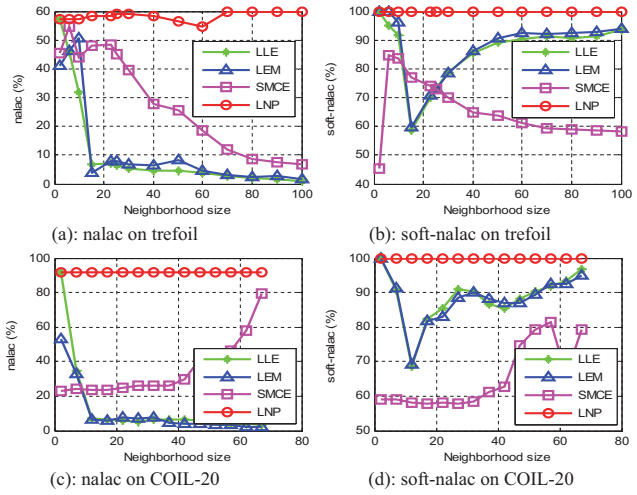


Figure 3: Embeddings of trefoil. For each method, the left is the obtained embeddings while the right is the corresponding ***dev***s.

The corresponding *nalac* and *soft-nalac* are shown in Figure 4 (a) and (b) respectively. Clearly, by using the proposed method, the neighborhood alignment accuracy is highest. That means the intrinsic structure is preserved better compared with the other methods.

Next, consider a case where the manifolds are intertwined and close to each other as shown in Figure 2 (b). 400 points are sampled from a 2-trefoil-knots, which are embedded in $\mathbb{R}^{100}$ and are corrupted with small Gaussian white noise. These mehtods, LLE, LEM, SMCE and LNP are used for manifold clustering with different neighborhood sizes. Figure 5 (a) reports the misclassification rates and Figure 6 (b) reports the corresponding time costs. The results shows that the proposed method can always achieve the stable and good clustering behaviour while the time costs are lowest among all other methods.

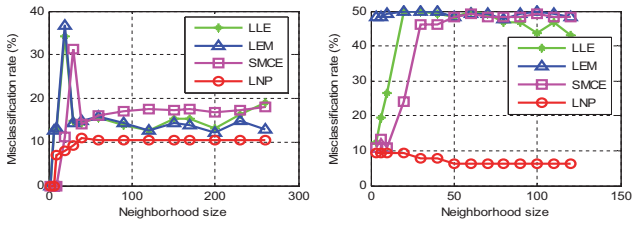
Real data. Two most-used real datasets, COIL-20 and ExtendedYaleB are used in this section, as shown in Figure 2 (c) and (d) respectively. The COIL-20 dataset includes 20 objects. Each object has 72 images of different positions and the size of each image is 192×168. Meanwhile, the 72 duck images is used for manifold embedding. These methods, LLE, LEM, SMCE and LNP, are used to learn the low dimension manifold with $K = 2, 3, \cdots, 71$ and $\lambda = \frac{60}{2}, \frac{60}{3}, \cdots, \frac{60}{71}$ for SMCE. Figure 6 (c) reports the corresponding time costs, which shows that the LNP is running fast. Figure 7 shows the final embedding results and the corresponding ***dev***s. The results shows that the SCRs learnt by LNP scheme are always non-negative and sparse. The ***dev***s of our method always reflect the intrinsic dimension of the data. The corresponding *nalac* and *soft-nalac* are reported in Figure 4 (c) and (d) respectively, which indicates that the proposed method has a good performance of intrinsic structure preservation.



(a): nalac on trefoil (b): soft-nalac on trefoil

(c): nalac on COIL-20 (d): soft-nalac on COIL-20

Figure 4: Intrinsic manifold structure preservation behaviour of various manifold learning methods.



(a): Misclassification rate on 2-trefoil-knots

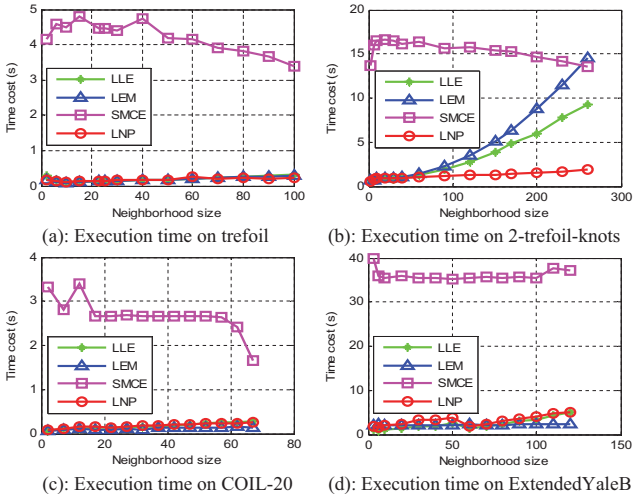
(b): Misclassification rate on ExtendedYaleB

Figure 5: Manifold clustering behaviour of various methods.

The ExtendedYaleB data set contains about $2,414$ frontal face images of 38 individuals with that the size of each image is 192×168. LLE, LEM, SMCE and LNP, are used for manifold clustering with different neighborhood sizes. Figure 5 (b) reports the misclassification rates and Figure 6 (d) reports the corresponding time costs. These results show that the proposed method can always obtain good performance of manifold clustering and the time costs are low. Figure 8 shows the embedding results and the corresponding ***dev***s of first two clusters which are obtained by using proposed LNP method.

Conclusion

This paper devotes to analyze the essence of the existence of negative components in the representation. A local Non-negative Pursuit (LNP) algorithm is proposed to select the neighbors of a data point and the Sparse Convex Representations (SCR) are learnt. It has been shown that the SCRs learnt by using the proposed method are sparse and non-negative. Experimental results show that the proposed method achieves or outperforms the state-of-the-art results.



(a): Execution time on trefoil (b): Execution time on 2-trefoil-knots

(c): Execution time on COIL-20 (d): Execution time on ExtendedYaleB

Figure 6: The execution time of various methods.

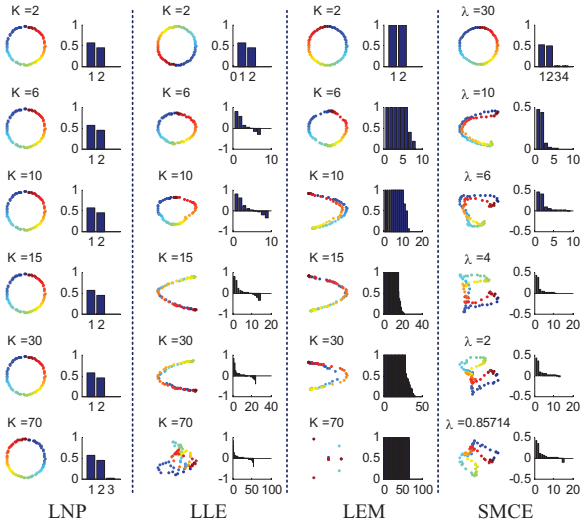


Figure 7: Embeddings of COIL-20. For each method, the left is the obtained embeddings while the right is the corresponding ***dev***s.

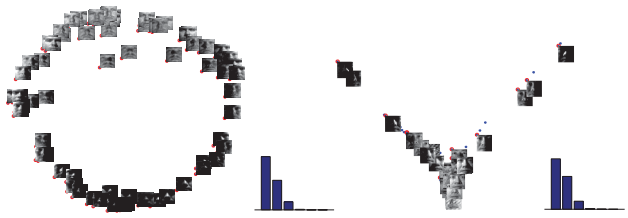


Figure 8: The embeddings of first two clusters of ExtendedYaleB, which are given by using the proposed LNP scheme, and the corresponding ***dev*** at the right bottom.

Acknowledgments

This work was supported by the National Science Foundation of China under grant 61375065, and National Program on Key Basic Research Project (973 Program) under Grant 2011CB302201.

Appendix A

The proof of Theorem 1: Since $\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i \in \mathcal{S}_a\{\mathcal{A}_k\}^+$, $\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i$ can be written as

$$\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i = \sum_{t=1}^{k} x_t\mathbf{a}_t, \tag{12}$$

with $\mathbf{a}_t \in \mathcal{A}_k, t = 1, \cdots, k$ and

$$\sum_{t=1}^{k} x_t = 1, x_t \geq 0. \tag{13}$$

Let $\mathbf{r}_k = \mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i$. Clearly,

$$\mathbf{r}_k = \mathbf{a}_i - \sum_{t=1}^{k} x_t\mathbf{a}_t = \sum_{t=1}^{k} x_t(\mathbf{a}_i - \mathbf{a}_t) = \sum_{t=1}^{k} x_t\mathbf{g}_t, \tag{14}$$

where $\mathbf{g}_t = \mathbf{a}_i - \mathbf{a}_t$ and $x_t \geq 0 (t = 1, \cdots, k)$. Thus, $\mathbf{r}_k \in \mathcal{S}\{\mathcal{G}_k\}^+$.

Since $\mathbb{P}_{\mathcal{S}\{\mathcal{G}_k\}}\mathbf{g}_j \in \mathcal{S}\{\mathcal{G}_k\}^-$, it holds that

$$90^\circ < \angle(\mathbf{r}_k, \mathbf{g}_j) \leq 180^\circ.$$

Let $\mathbf{o} = [0, 0, \cdots, 0]^\top \in \mathbb{R}^m$, we have

$$\mathbb{P}_{\mathcal{S}_a\{\mathbf{r}_k, \mathbf{g}_j\}}\mathbf{o} \in \mathcal{S}_a\{\mathbf{r}_k, \mathbf{g}_j\}^+.$$

It means that

$$\mathbb{P}_{\mathcal{S}_a\{\mathbf{a}_i - \mathbf{r}_k, \mathbf{a}_i - \mathbf{g}_j\}}(\mathbf{a}_i - \mathbf{o}) \in \mathcal{S}_a\{\mathbf{a}_i - \mathbf{r}_k, \mathbf{a}_i - \mathbf{g}_j\}^+.$$

Then, we can get

$$\begin{aligned} &\mathbb{P}_{\mathcal{S}_a\left\{\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i, \mathbf{a}_j\right\}}\mathbf{a}_i \in \mathcal{S}_a\left\{\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i, \mathbf{a}_j\right\}^+ \\ \Rightarrow &\mathbb{P}_{\mathcal{S}_a\left\{\sum_{t=1}^{k} x_t\mathbf{a}_t, \mathbf{a}_j\right\}}\mathbf{a}_i \in \mathcal{S}_a\left\{\sum_{t=1}^{k} x_t\mathbf{a}_t, \mathbf{a}_j\right\}^+. \end{aligned} \tag{15}$$

By the definition of $S_a\{\cdot\}^+$, we have

$$\begin{aligned} &\mathcal{S}_a\left\{\sum_{t=1}^{k} x_t\mathbf{a}_t, \mathbf{a}_j\right\}^+ \\ &= \left\{\eta_1(\sum_{t=1}^{k} x_t\mathbf{a}_t) + \eta_2\mathbf{a}_j\right\} \\ &= \left\{(\eta_1 x_1)\mathbf{a}_1 + \cdots + (\eta_1 x_k)\mathbf{a}_k + (\eta_2)\mathbf{a}_j\right\}, \end{aligned}$$

where $\eta_1 + \eta_2 = 1, \quad \eta_1, \eta_2 \geq 0$. Clearly,

$$\begin{aligned} &\eta_1 x_1 + \eta_1 x_2 + \cdots + \eta_1 x_k + \eta_2 \\ &= \eta_1(x_1 + x_2 + \cdots + x_k) + \eta_2 \\ &= \eta_1 + \eta_2 \\ &= 1. \end{aligned}$$

Thus,

$$\mathcal{S}_a\left\{\sum_{t=1}^{k} x_t\mathbf{a}_t, \mathbf{a}_j\right\}^+ \equiv \mathcal{S}_a\left\{\mathcal{A}_{k+1}\right\}^+. \tag{16}$$

By Eqs.(15) and Eqs.(16), we have

$$\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k+1}\}}\mathbf{a}_i \in \mathcal{S}_a\{\mathcal{A}_{k+1}\}^+$$

The proof is complete. ■

The proof of Theorem 2: Let $\mathbf{a}_t \in \mathcal{A}$ is the point selected at step k. If $\mathbf{a}_t \in \mathcal{A}_{k-1}$, it means that $\mathbf{a}_t \in \mathcal{S}\{\mathcal{A}_{k-1}\}^+$ and $\mathbb{P}_{\mathcal{S}\{\mathcal{A}_{k-1}\}}\mathbf{a}_t = \mathbf{a}_t$. Thus $\mathbb{P}_{\mathcal{S}\{\mathcal{A}_{k-1}\}}\mathbf{a}_t \in \mathcal{S}\{\mathcal{A}_{k-1}\}^+$. It follows that

$$\begin{aligned} &\mathbb{P}_{\mathcal{S}\{\mathcal{A}_{k-1}\}}\mathbf{a}_t \in \mathcal{S}\{\mathcal{A}_{k-1}\}^+ \\ \Leftrightarrow &\mathbb{P}_{\mathcal{S}\{\mathcal{G}_{k-1}\}}\mathbf{g}_t \in \mathcal{S}\{\mathcal{G}_{k-1}\}^+. \end{aligned} \tag{17}$$

where $\mathcal{G}_{k-1} = \{\mathbf{a}_i - \mathbf{a}_j | \mathbf{a}_j \in \mathcal{A}_{k-1}\}$ and $\mathbf{g}_t = \mathbf{a}_i - \mathbf{a}_t$.

Clearly, it is contradict with the condition (4). Thus, $\mathbf{a}_t$ can not be selected twice. The proof is complete. ■

The proof of Theorem 3: The space spanned by $\mathcal{A}$ can be written in a direct sum:

$$\mathcal{S}\{\mathcal{A}\} = \mathcal{S}_a\{\mathcal{A}\} \oplus \{\mathcal{S}_a\{\mathcal{A}\}\}^\perp.$$

Clearly, $\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i \in \mathcal{S}_a\{\mathcal{A}_k\}$ and

$$\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i \in \{\mathcal{S}_a\{\mathcal{A}_k\}\}^\perp. \tag{18}$$

Since $\mathcal{S}_a\{\mathcal{A}_{k-1}\} \subsetneqq \mathcal{S}_a\{\mathcal{A}_k\}$, we have

$$\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k-1}\}}\mathbf{a}_i \in \mathcal{S}_a\{\mathcal{A}_k\},$$

thus,

$$(\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k-1}\}}\mathbf{a}_i) \in \mathcal{S}_a\{\mathcal{A}_k\}. \tag{19}$$

From Eqs.(18) and Eqs.(19), we have

$$(\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i) \perp (\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k-1}\}}\mathbf{a}_i).$$

Since

$$\begin{aligned} (\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k-1}\}}\mathbf{a}_i) = &(\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i) \\ &+ (\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k-1}\}}\mathbf{a}_i), \end{aligned}$$

From the Pythagorean theorem, it holds that:

$$\begin{aligned} \|\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k-1}\}}\mathbf{a}_i\|_2^2 = &\|\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i\|_2^2 \\ &+ \|\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k-1}\}}\mathbf{a}_i\|_2^2, \end{aligned}$$

i.e.,

$$\|\mathbf{r}_{k-1}\|_2^2 = \|\mathbf{r}_k\|_2^2 + \|\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k-1}\}}\mathbf{a}_i\|_2^2.$$

Clearly, $\|\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_k\}}\mathbf{a}_i - \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{k-1}\}}\mathbf{a}_i\|_2^2 > 0$, thus $\|\mathbf{r}_k\|_2 < \|\mathbf{r}_{k-1}\|_2$. The proof is complete. ■

The proof of Theorem 4: Clearly,

$$\mathbf{A}\mathbf{x}_i = \sum_{j=1}^{k} x_{i\lambda_j}\mathbf{a}_{\lambda_j} = \mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{opt}\}}\mathbf{a}_i,$$

where $\mathbf{a}_{\lambda_j} \in \mathcal{A}_{opt}$ and $\sum_{j=1}^{k} x_{i\lambda_j} = 1$.

It is easy to see that $\mathbb{P}_{\mathcal{S}_a\{\mathcal{A}_{opt}\}}\mathbf{a}_i \in \mathcal{S}_a\{\mathcal{A}_{opt}\}^+$. Thus, the learnt affine representation $\mathbf{x}_i$ is non-negative. Since $\|\mathbf{x}_i\|_0 = k (\leq d+1)$, $\mathbf{x}$ is also sparse (Elad 2010). The proof is complete. ■

References

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

Cai, D.; He, X.; Han, J.; and Huang, T. S. 2011. Graph regularized nonnegative matrix factorization for data representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33(8):1548–1560.

Chen, L., and Buja, A. 2009. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association* 104(485):209–219.

Elad, M. 2010. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer.

Elhamifar, E.; Sapiro, G.; and Vidal, R. 2012. See all by looking at a few: Sparse modeling for finding representative objects. 1600–1607. IEEE.

He, X., and Niyogi, P. 2003. Locality preserving projections. In *NIPS*, volume 16, 234–241.

Hoyer, P. O. 2002. Non-negative sparse coding. *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on* 557–565.

Lee, J. A., and Verleysen, M. 2009. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing* 72(7):1431–1443.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 663–670.

Lv, J. C.; Kok Kiong, T.; Zhang, Y.; and Huang, S. 2009. Convergence analysis of a class of hyvärinen–oja's ica learning algorithms with constant learning rates. *Signal Processing, IEEE Transactions on* 57(5):1811–1824.

Lv, J. C.; Zhang, Y.; and Kok Kiong, T. 2007. Global convergence of gha learning algorithm with nonzero-approaching learning rates. *Neural Networks, IEEE Transactions on* 18(6):1557–1571.

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

Saul, L. K., and Roweis, S. T. 2003. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *The Journal of Machine Learning Research* 4:119–155.

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and computing* 17(4):395–416.

Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(2):210–227.

Yan, S.; Xu, D.; Zhang, B.; Zhang, H.-J.; Yang, Q.; and Lin, S. 2007. Graph embedding and extensions: a general framework for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29(1):40–51.

Zhang, D.; Yang, M.; and Feng, X. 2011. Sparse representation or collaborative representation: Which helps face recognition? In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 471–478. IEEE.

Zhuang, L.; Gao, H.; Lin, Z.; Ma, Y.; Zhang, X.; and Yu, N. 2012. Non-negative low rank and sparse graph for semi-supervised learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2328–2335. IEEE.