LARGEST CENTER-SPECIFIC MARGIN FOR DIMENSION REDUCTION

Jian'an Zhang *Yuan Yuan* *Feiping Nie* *Qi Wang*

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, PR China

ABSTRACT

Dimensionality reduction plays an important role in solving the "curse of the dimensionality" and attracts a number of researchers in the past decades. In this paper, we proposed a new supervised linear dimensionality reduction method named largest center-specific margin (LCM) based on the intuition that after linear transformation, the distances between the points and their corresponding class centers should be small enough, and at the same time the distances between different unknown class centers should be as large as possible. On the basis of this observation, we take the unknown class centers into consideration for the first time and construct an optimization function to formulate this problem. In addition, we creatively transform the optimization objective function into a matrix function and solve the problem analytically. Finally, experiment results on three real datasets show the competitive performance of our algorithm.

Index Terms— Dimensionality Reduction, LCM, Center-specific Method

1. INTRODUCTION

Dimensionality reduction plays an important role in solving the "curse of the dimensionality". Directly working on high dimensional data is not only time consuming but also computationally unreliable. So a great effort has been put in the past decades and many classical algorithms have been proposed. A good review of these algorithms can be referenced from [1] [2] [3]. In addition, new ideas and methods can be further referenced from [4] [5] [6] [7] [8] [9] [10] [11] [12]

Traditional dimensionality reduction algorithms can be grouped into two classes, unsupervised ones and supervised ones. A great number of these methods belong to unsupervised ones such as principal components analysis (PCA [13]), however, compared with supervised methods, unsupervised methods cannot make full use of the samples' potential. On the other hand, most of traditional dimensionality reduction methods do not utilize the information of class centers. Therefore, supervised method with class centers' information can be taken into consideration and applied into dimensionality reduction.

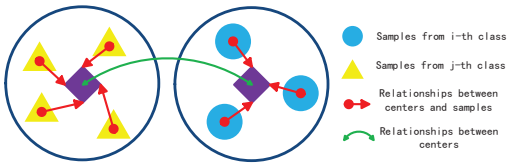

Fig. 1. Diagram for the intuition, circles and triangles should be close to their diamond centers and the distances between different diamond centers should be as large as possible.

In this work, we proposed a new linear dimensionality reduction method on the basis of the observation that after linear transformation the distances between the points and their corresponding class center should be small enough, and at the same time the distances between different unknown class centers should be as large as possible. It will be clearer to understand the above idea from Fig. 1.

From Fig.1, it can be found that the class centers' information is of importance to the dimensionality reduction. So in this work, for the first time, we take the unknown class centers that generate after linear transformation into consideration. And based on the relationships showed in Fig.1, we construct an optimization objective function with the variables of the transformation matrix $\mathbf{A}$ and the unknown class centers $\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_n}$ to formulate this intuition. Furthermore, we creatively convert the initial objection function into a matrix function which is more prone to analysing and solving the problem. Moreover, we study the objective function intensively and improve the objective function by imposing two regular terms making it a convex function, and at the same time the meanings of the formulation are reserved. At last, we get the transformation matrix by solving the optimization problem and the low-dimensional transformed data can be acquired by multiplying the transformation matrix.

Qi Wang and Feiping Nie are the corresponding authors. This work is supported by the National Natural Science Foundation of China under Grant 61379094 and Natural Science Foundation Research Project of Shaanxi Province under Grant 2015JM6264.

2. METHODOLOGY

In this section, we introduce a new linear dimensionality reduction method named largest center-specific margin(LCM). As a definition, linear dimensionality means, given n d-dimensional data points $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}] \in R^{d\times n}$ and a choice of dimensionality $r < d$, optimize some objective $f_{\mathbf{X}}(\cdot)$ to produce a linear transformation $\mathbf{A} \in R^{r\times d}$, and call $\mathbf{Y} = \mathbf{AX} \in R^{r\times n}$ the low-dimensional transformed data. Next,we introduce a new method to optimize the linear transformation matrix $\mathbf{A}$

We build on the simple intuition that after linear transformation the distances between the points of the same label and their corresponding class center should be small enough, and at the same time the distances among unknown centers of different classes should be as large as possible. As shown in Fig. (1), we can take the class centers' information into consideration and establish the relationships between points and their unknown centers as well as the relationships among the corresponding centers. From the basic intuition, we can formulate the idea as below

$$\min_{\mathbf{A},\mathbf{y_1},\mathbf{y_2},...,\mathbf{y_c}} \sum_i \sum_j \|\mathbf{Ax_i} - \mathbf{y_j}\|^2 - \sum_j \sum_{\mathbf{c_j}\neq\mathbf{c_i}} \|\mathbf{y_i} - \mathbf{y_j}\|^2, \quad (1)$$

where $\mathbf{x_i}, i = 1, 2, ..., n$ are the feature representations of instances of different classes and $\mathbf{y_j}, j = 1, 2, ..., c$ are the unknown class centers. In Eq. (1), the first term implies that after linear transformation $\mathbf{A}$, the distances between the points and their corresponding unknown center of the same class, and the second term represents the distance between two unknown centers of different class. In intuition, in order to acquire an effective transformation matrix, the first term should be as small as possible and in contrast the second term should be as larger as possible. So we transform the second term to the minus term making it a unified optimal problem.

Eq. (1) can be transformed into the following matrix form after permutation and combination of the terms

$$\min_{\mathbf{A},\mathbf{Y}} \|\mathbf{AX} - \mathbf{YC}\|_F^2 - tr(\mathbf{YLY}^T), \quad (2)$$

where $\|\cdot\|_F$ is Frobenius norm, $tr(\cdot)$ stands for the trace operator. m is the dimensionality of feature vector and n represents the number of all training samples. $\mathbf{A} \in R^{m\times m}$ is a linear transform matrix and $\mathbf{X} \in R^{d\times n}$ stands for the sample matrix which consist of all the training samples and each column stands for a feature vector of an instance of one class. $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_c}] \in R^{m\times c}$ is the matrix synthesized by the centers of c classes. $\mathbf{C} \in R^{c\times n}$ is a matrix with the following form

$$\mathbf{C} = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}$$

$\mathbf{L} = \mathbf{I} - \frac{1}{c}\mathbf{1}\mathbf{1}^T$, which is the centering matrix, I stands for identity matrix with dimensionality of c, and $\mathbf{1}$ stands for the c-dimensional vector with all elements being 1. Note that, $\|\mathbf{AX} - \mathbf{YC}\|_F^2 = tr(\mathbf{AX} - \mathbf{YC})(\mathbf{AX} - \mathbf{YC})^T$, so Eq. (2) can be simplified into the following form

$$\begin{aligned}\min_{\mathbf{A},\mathbf{Y}} & tr(\mathbf{AXX}^T\mathbf{A}^T) - 2tr(\mathbf{Y}^T\mathbf{AXC}^T) \\ & + tr(\mathbf{Y}(\mathbf{CC}^T - \mathbf{I} + \frac{1}{c}\mathbf{1}\mathbf{1}^T)\mathbf{Y}^T).\end{aligned} \quad (3)$$

Nevertheless, the optimal problem described in Eq.(1) is not guaranteed to be convex. For the convenience of tracking, the original problem can be reformulated as follow, i.e. add two regular terms to the objective function

$$\begin{aligned}\min_{\mathbf{A},\mathbf{y_1},\mathbf{y_2},...,\mathbf{y_n}} & \sum_i \sum_j \|\mathbf{Ax_i} - \mathbf{y_j}\|^2 - \sum_j \sum_{\mathbf{c_j}\neq\mathbf{c_i}} \|\mathbf{y_i} - \mathbf{y_j}\|^2 \\ & + \gamma\|\mathbf{A}\|_F^2 + \eta\|\mathbf{Y}\|_F^2.\end{aligned} \quad (4)$$

Note that, after modifying, the new optimal Eq. (4) is jointly convex with regard to $\mathbf{A}$ and $\mathbf{Y}$, hence this optimal problem has globally optimal solution. Moreover, even if adding two regular terms to the original optimal Eq. (1), the significance of the problem is not changed because the above regular terms are equivalent to imposing constraint to $\mathbf{A}$ and $\mathbf{Y}$ so that the norms of $\mathbf{A}$ and $\mathbf{Y}$ are not too large.

In the same manner, we can convert the Eq.(4) into the following matrix form based on the Eq.(3) and the property of trace operator.

$$\min_{\mathbf{A},\mathbf{Y}} tr(\mathbf{ANA}^T) - 2tr(\mathbf{Y}^T\mathbf{AXC}^T) + tr(\mathbf{YKY}^T), \quad (5)$$

where

$$\mathbf{N} = \mathbf{XX}^T + \gamma\mathbf{I}, \quad (6)$$

$$\mathbf{K} = \mathbf{CC}^T + (\eta - 1)\mathbf{I} + \frac{1}{c}\mathbf{1}\mathbf{1}^T, \quad (7)$$

and $\gamma > 0, \eta > 1$.

From now on, the optimization objective function has been established. So the next task is to solve the optimal Eq. (5). It is noted that Eq. (6) is continuous with regard to $\mathbf{A}$ and $\mathbf{Y}$, hence it can be solved by taking the derivation of one of the variables when fixed the other one and letting the derivation be $\mathbf{0}$. By fixing $\mathbf{Y}$, we get the derivation of Eq.(5) w.r.t $\mathbf{A}$, this is $\mathbf{AN} - \mathbf{YCX} = \mathbf{0}$, and hence we can get

$$\mathbf{A} = \mathbf{YCX}^T\mathbf{N}^{-1}. \quad (8)$$

By fixing A, we get the derivation of Eq.(5) w.r.t $\mathbf{Y}$, this is $\mathbf{YK} - \mathbf{AXC}^T = \mathbf{0}$, and hence we can get

$$\mathbf{Y} = \mathbf{AXC}^T\mathbf{K}^{-1}. \quad (9)$$

Algorithm 1 LCM Algorithm for Dimensionality Reduction

Input:

The n training samples with corresponding labels $(\mathbf{x_i}, \mathbf{y}_i)_{i=1}^n$

Output:

The transformation matrix $\mathbf{A}$

1: Initialize parameters γ, η, error bound ε, class numbers c, and set $k = 0$;
2: Execute PCA Algorithm and get the transformaztion matrix $\mathbf{P}$, set $\mathbf{A_k} \leftarrow \mathbf{P}$
3: Construct matrix $\mathbf{C}$, and calculate matrix $\mathbf{N}$, $\mathbf{K}$ from E.q.(7) and E.q.(8)
4: **while** true **do**
5: $k \leftarrow k + 1$;
6: update $\mathbf{Y_k} \leftarrow \mathbf{A_{k-1}} \mathbf{X} \mathbf{C}^T \mathbf{K}^{-1}$
7: update $\mathbf{A_k} \leftarrow \mathbf{Y}_{k-1} \mathbf{C} \mathbf{X}^T \mathbf{N}^{-1}$
8: **if** $\|\mathbf{A_k} - \mathbf{A}_{k-1}\|_F < \varepsilon$ **then**
9: return $\mathbf{A_k}$
10: break
11: **end if**
12: **end while**
13: Set $\mathbf{A} \leftarrow \mathbf{A_k}$

3. EXPERIMENT

We compare our algorithm with other dimensionality reduction methods, including PCA, MDS, LLE, LE, Isomap and LCM. Aside from the visualization on synthetic datasets, we also show results on three real datasets.

3.1. Visualization on Synthetic Dataset

We first show the visualization of our LCM algorithm on four synthetic datasets from 3D to 2D, they are Swiss roll dataset,vhelix dataset, twinpeaks dataset and broken Swiss roll dataset. Details of the synthetic datasets are shown in [1]. For every dataset, we generate 2000 data points. Fig.2 show the results. We can see that our LCM algorithm always projects the high dimensional data points into a linear manifold and at the same time maintains the property of clustering, which is a powerful tool for analysing the high dimensional data points. Compared with LCM, traditional dimensionality reduction methods such as PCA and LLE do not maintain the special shapes in low-dimensional space. It's worth pointing out that these four datasets are usually used to test the nonlinear dimensionality methods because of there non-cluster structure, and our algorithm also show good structure after embedding to low-dimensional space.

3.2. Real Datasets

In order test our algorithm on real datasets, we choose three datasets to perform classification tasks, i.e. (1) the ORL dataset [14], (2) the Yale dataset [15], (3) the UMIST dataset [16].

In experiments, firstly, we resize every image to the same size and convert it to a column vector as the original high-dimensional data representation. Next, dimensionality reduction algorithms including PCA [17], MDS [18], LLE [19], LE [20], Isomap [21], are used to project the high-dimensional data representations into a low-dimensional data representations. At last, we perform classification tasks on the low-dimensional data representations by randomly selecting train samples and test samples. Without loss of the generality, we utilize the simple k-NN classifier ($k = 1$ in our experiments) and evaluate our algorithm with the classification accuracy. For our LCM algorithm, we fixed our parameters $\eta = 1.5$, $\gamma = 0.5$.

In Fig.3, we present the accuracy of 1-nearest neighbor classifiers with different numbers of dimensionality which were trained and tested on the low-dimensional data representations obtained from the dimensionality reduction techniques. From Fig. 3, it is clear that in ORL dataset and UMIST dataset, our LCM algorithm achieved the best performance with nearly 100% accuracy for every dimensionality.

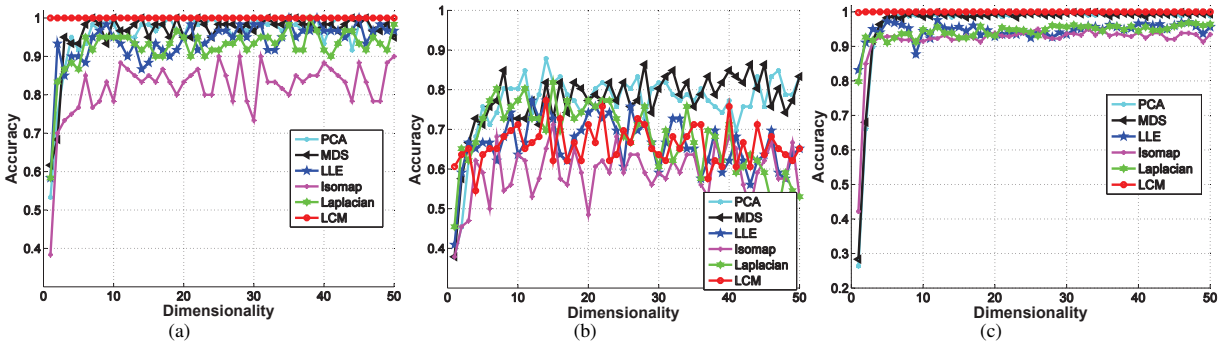

Fig. 3. (a) accuracy of 1-NN classifier on ORL dataset (b) accuracy of 1-NN classifier on Yale dataset (c) accuracy of 1-NN classifier on UMIST dataset

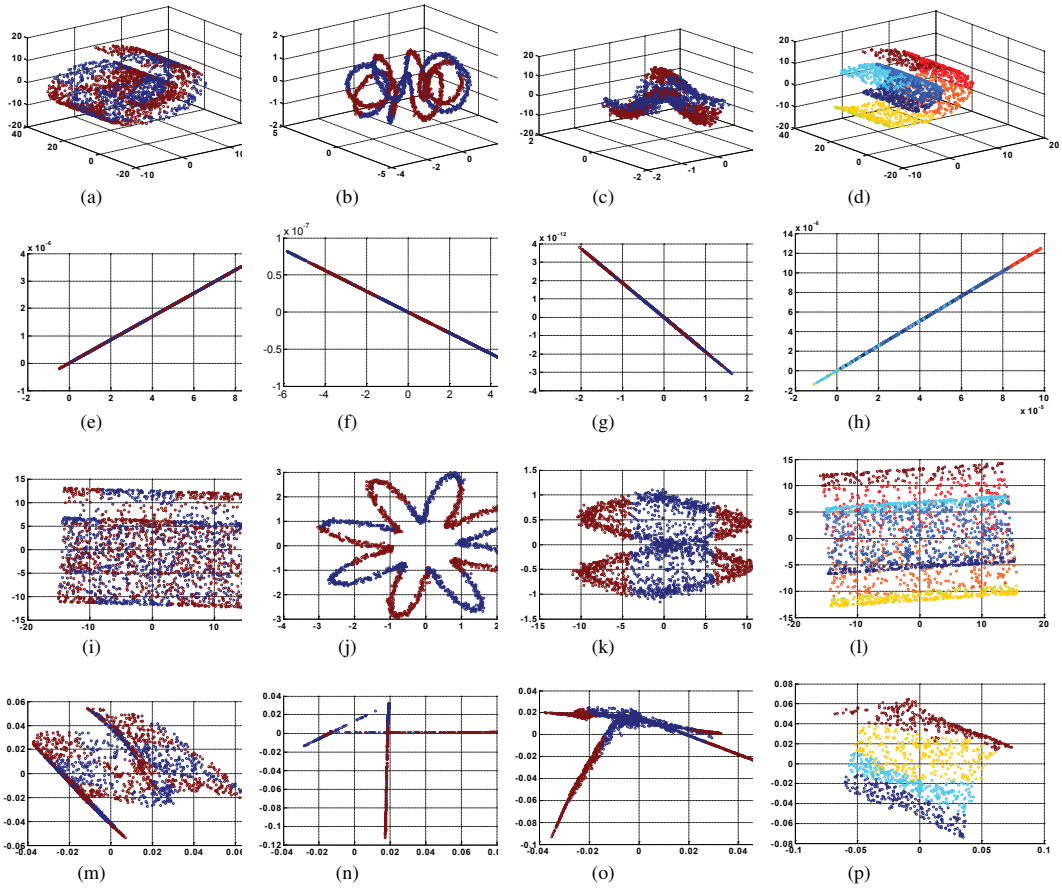

Fig. 2. 3D to 2D results on synthetic data. (a)-(d) original datasets from left to right:Swiss,Helix,twinpeaks and broken Swiss, (e)-(h) LCM (i)-(l) PCA (m)-(p) LLE

In Yale dataset, however, compared with traditional dimensionality reduction methods, LCM algorithm performed equivalent to PCA and Isomap though it did not achieve the best performance. Especially for UMIST dataset and ORL dataset , in which every class is of quite a number of samples , LCM achieved an astonishing performance. From the objective function, it can be found that our LCM algorithm is designed for the classes that are of quite a number of data points, this is why in UMIST dataset and ORL dataset, our LCM algorithm performed so good. In other hand, it also can be found in Yale dataset, our LCM also showed it's good performance, which demonstrates the competitive ability with traditional dimensionality reduction methods.

4. CONCLUSION

Dimensionality reduction algorithms play an significant role in solving the "curse of dimensionality". In this work, we proposed a new linear dimensionality reduction algorithm named Largest Center-specific Margin (LCM). Our algorithm is built upon the observation that after linear transformation, the distances between the points and their corresponding class centers should be small enough and the distances among unknown centers of different class should be large enough. For the first time, we take the unknown class centers into consideration. And based on the relationships showed in Fig.1, we construct an optimization objective function to formulate this intuition. Furthermore, we creatively convert the initial objection function into a matrix function which is more prone to analysing and solving the problem. We test our algorithm in classification tasks on three real datasets and experiment results showed that our LCM algorithm is competitive with traditional algorithms. In addition, visualization from 3D to 2D showed that our LCM algorithm always embedded the high dimensional data points into a linear manifold while other algorithms did not maintain special shapes. So it is more convenient to study the structure of high-dimensional manifolds.

References

[1] Postma EO van der MLJP and J van den HH, "Dimensionality reduction: A comparative review," Tech. Rep., Tilburg, Netherlands: Tilburg Centre for Creative Computing, Tilburg University, Technical Report: 2009-005, 2009.

[2] Carlos Oscar Sánchez Sorzano, Javier Vargas, and A Pascual Montano, "A survey of dimensionality reduction techniques," *arXiv preprint arXiv:1403.2877*, 2014.

[3] John P Cunningham and Zoubin Ghahramani, "Linear dimensionality reduction: Survey, insights, and generalizations," *Journal of Machine Learning Research*, vol. 16, pp. 2859–2900, 2015.

[4] Qi Wang and Xuelong Li, "Shrink image by feature matrix decomposition," *Neurocomputing*, vol. 140, pp. 162–171, 2014.

[5] Tanaya Mandal, QM Jonathan Wu, and Yuan Yuan, "Curvelet based face recognition via dimension reduction," *Signal Processing*, vol. 89, no. 12, pp. 2345–2353, 2009.

[6] Jayaraman J Thiagarajan, Peer-Timo Bremer, and Karthikeyan Natesan Ramamurthy, "Multiple kernel interpolation for inverting non-linear dimensionality reduction and dimension estimation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6751–6755.

[7] Olga Zoidi, Nikos Nikolaidis, and Ioannis Pitas, "Semi-supervised dimensionality reduction on data with multiple representations for label propagation on facial images," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6019–6023.

[8] Amir Najafi, Amir Joudaki, and Emad Fatemizadeh, "Nonlinear dimensionality reduction via path-based isometric mapping," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1452–1464, 2016.

[9] Xi Peng, Jiwen Lu, Zhang Yi, and Yan Rui, "Automatic subspace learning via principal coefficients embedding," *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–14, 2016.

[10] Dimitrios Bouzas, Nikolaos Arvanitopoulos, and Anastasios Tefas, "Graph embedded nonparametric mutual information for supervised dimensionality reduction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 951–963, 2015.

[11] Ruiping Wang, Shiguang Shan, Xilin Chen, Jie Chen, and Wen Gao, "Maximal linear embedding for dimensionality reduction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 9, pp. 1776–1792, 2011.

[12] Meng Meng, Jia Wei, Jiabing Wang, Qianli Ma, and Xuan Wang, "Adaptive semi-supervised dimensionality reduction based on pairwise constraints weighting and graph optimizing," *International Journal of Machine Learning and Cybernetics*, pp. 1–13, 2015.

[13] Yanwei Pang, Dacheng Tao, Yuan Yuan, and Xuelong Li, "Binary two-dimensional pca," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 4, pp. 1176–1180, 2008.

[14] Alan J Chaney, Ian D Wilson, and Andrew Hopper, "The design and implementation of a raid-3 multimedia file server," in *International Workshop on Network and Operating Systems Support for Digital Audio and Video*. Springer, 1995, pp. 306–317.

[15] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[16] Daniel B Graham and Nigel M Allinson, "Characterising virtual eigensignatures for general purpose face recognition," in *Face Recognition*, pp. 446–456. Springer, 1998.

[17] K. Allab, L. Labiod, and M. NADIF, "A semi-nmf-pca unified framework for data clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, no. 99, pp. 1–1, 2016.

[18] Trevor F Cox and Michael AA Cox, *Multidimensional scaling*, CRC press, 2000.

[19] Y. Zhang, H. Lv, Y. Liu, H. Wang, X. Wang, Q. Huang, X. Xiang, and Q. Dai, "Light field depth estimation via epipolar plane image analysis and locally linear embedding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2016.

[20] L. Shi, L. Zhang, L. Zhao, L. Zhang, P. Li, and D. Wu, "Adaptive laplacian eigenmap-based dimension reduction for ocean target discrimination," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 7, pp. 902–906, July 2016.

[21] Z. Zhang, T. W. S. Chow, and M. Zhao, "M-isomap: Orthogonal constrained marginal isomap for nonlinear dimensionality reduction," *IEEE Transactions on Cybernetics*, vol. 43, no. 1, pp. 180–191, Feb 2013.