ROBUST ADAPTIVE SPARSE LEARNING METHOD FOR GRAPH CLUSTERING

Mulin Chen[1] *Qi Wang*[1,2*] *Xuelong Li*[3,4]

[1]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China
[2]Unmanned System Research Institute (USRI),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China
[3]Xi'an Institute of Optics and Precision Mechanics,
Chinese Academy of Sciences, Xi'an, 710119, Shaanxi, P. R. China
[4]University of Chinese Academy of Sciences, Beijing, 100049, P. R. China

ABSTRACT

Graph clustering aims to group the data into clusters according to a similarity graph, and has received sufficient attention in computer vision. As the basis of clustering, the quality of graph affects the results directly. In this paper, a Robust Adaptive Sparse Learning (RASL) method is proposed to improve the graph quality. The contributions made in this paper are three fold: (1) the sparse representation technique is employed to enforce the graph sparsity, and the $\ell_{2,1}$ norm is introduced to improve the robustness; (2) the intrinsic manifold structure is captured by investigating the local relationship of data points; (3) an efficient optimization algorithm is designed to solve the proposed problem. Experimental results on various real-world benchmark datasets demonstrate the promising results of the proposed graph-based clustering method.

***Index Terms*—** Clustering, Manifold Structure, Graph Construction, Sparse Learning

1. INTRODUCTION

Data clustering partitions the data points into different categories, and is a hot research area in computer vision and machine learning. In the past decades, plenty of techniques have been proposed toward this topic, such as k-means clustering [1], multiview clustering [2–4], graph clustering [5–8], Non-negative Matrix Factorization (NMF) [9], and support vector clustering [10]. Among these methods, graph clustering achieves the state-of-the-art performance because of the utilization of data relationship, and has been used in many practical applications, such as document clustering [11] and image segmentation [8], *etc*.

*Corresponding author. This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, National Natural Science Foundation of China under Grant 61773316, Fundamental Research Funds for the Central Universities under Grant 3102017AX010, and the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences.

Graph clustering methods (e.g., spectral clustering [5], ratio cut [6] and normalized cut [8]) are based on the graphical representation of data relationship. These methods first transform the data points into a weighted graph according to their distances, and then perform graphical optimization to accomplish the clustering task. So the quality of graph determines the clustering performance directly. However, there are mainly three problems on graph construction: (1) most methods rely on the Gaussian weighting graph, which is largely affected by the scale of analysis; (2) the data noise and outliers make the obtained graph inaccurate; (3) the data manifold is always neglected in the graph construction procedure. Some methods [12, 13] have been presented to solve the first two problems, but the third one is still unsolved. In addition, the similarity between the dissimilar points should be zero intuitively, which indicates the sparsity of obtained graph. However, only a few papers [14, 15] emphasize this aspect.

In this research, we propose a Robust Adaptive Sparse Learning (RASL) method to reduce the above problems. The data graph is constructed with the sparse learning formulation, and the $\ell_{2,1}$ norm is incorporated to mitigate the impact of data noise and outliers. And the manifold structure is exploited according to the local connectivity between data points. Quantitative experimental results show that the proposed RASL achieves better performance compared with the state-of-the-arts.

2. METHODOLOGY

In this section, the sparse representation method is introduced as the preliminary. Then we formulate the proposed Robust Adaptive Sparse Learning (RASL) objective function, and introduce the corresponding optimization method.

2.1. Preliminary

Suppose the data matrix is $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, and d and n are the data dimension and the number of samples respectively. According to [14], the sparse representation method learns a transformation vector $\beta \in \mathbb{R}^{n\times 1}$ by solving the following problem

$$\min_{\beta} ||\mathbf{X}\beta - \mathbf{y}||_2^2 + \lambda_0||\beta||_1, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{d\times 1}$ is the desired new representation, and λ_0 is a regularization parameter. The first term ensures that $\mathbf{y} \approx \mathbf{X}\beta$, and the second term guarantees the sparsity of the transformation. Sparse representation method reduces the computational complexity significantly, and removes the irrelevant factors. So it has been widely-used in many areas, such as face recognition [16] and hyperspectral image processing [17].

Recently, a Simplex Sparse Representation (SSR) model [14] has been proposed to utilize the sparse representation theory to learn the data similarity. The objective function of SSR is written as

$$\min_{\alpha_i\geq 0, \alpha_i^T\mathbf{1}=1} ||\mathbf{X}_{-i}\alpha_i - \mathbf{x}_i||_2^2, \tag{2}$$

where $\mathbf{1} \in \mathbb{R}^{(n-1)\times 1}$ is a vector with all the elements equal to 1, and $\mathbf{X}_{-i}$ is the data matrix without the i-th column. And $\alpha_i \in \mathbb{R}^{(n-1)\times 1}$ indicates the similarities between point i and all the others. The above problem assumes that a data point can be approximated by the linear combination of the others, and the constraint $\alpha_i^T\mathbf{1} = 1$ has the similar effect to the sparsity constraint (second term) in Eq. (1). Thus, SSR is able to learn a sparse similarity graph, and doesn't need to determine the scale of analysis. However, since the error for each data point is squared in the objective function, SSR is prone to be affected by the outliers. Moreover, the local relationship between the points is also neglected.

2.2. Robust Adaptive Sparse Learning Method

Here we propose the Robust Adaptive Sparse Learning (RASL) method. Instead of optimizing the similarity vector of each point separately, we propose to learn the data graph $\mathbf{S} \in \mathbb{R}^{n\times n}$ directly. So we have the following objective function

$$\min_{\mathbf{S}\geq 0, \sum_j \mathbf{S}_{ij}=1} ||\mathbf{X}\mathbf{S} - \mathbf{X}||_{2,1}. \tag{3}$$

Compared to the ℓ_2 norm in Eq. (2), the $\ell_{2,1}$ norm in the above problem mainly has two advantages: (1) it is more robust to outliers and data noise than ℓ_2 norm (2) it is invariance to the feature rotation [18].

In order to perceive the data manifold, it is essential to capture the local relationship between data points. According to the common sense, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close to each other, their similarity $\mathbf{S}_{ij}$ should be large. Then the objective function is rewritten as

$$\min_{\mathbf{S}\geq 0, \sum_j \mathbf{S}_{ij}=1} ||\mathbf{X}\mathbf{S} - \mathbf{X}||_{2,1} + \lambda_1 \sum_{ij} ||\mathbf{x}_i - \mathbf{x}_j||_2^2\mathbf{S}_{ij}, \tag{4}$$

where λ_1 is a parameter. Thus, the proposed RASL is able to exploit the manifold structure, and learn the sparse data graph with more robustness. After the graph $\mathbf{S}$ is obtained, we compute the spectral vectors with Laplacian matrix, and then accomplish the clustering by k-means. In the following part, an efficient algorithm is developed to solve problem (4).

2.3. Optimization Algorithm

Since both the terms in problem (4) depend on $\mathbf{S}$, we solve it with the Augmented Lagrangian Multiplier (ALM) [19], and introduce two auxiliary variables $\mathbf{W} = \mathbf{S}$ and $\mathbf{E} = \mathbf{X}\mathbf{W} - \mathbf{X}$. Then problem (4) is rewritten as the following ALM problem:

$$\begin{aligned} \min_{\mathbf{E},\mathbf{W},\mathbf{S}} & ||\mathbf{E}||_{2,1} + \lambda_1 \sum_{ij} ||\mathbf{x}_i - \mathbf{x}_j||_2^2\mathbf{S}_{ij} \\ & + \tfrac{\mu}{2}||\mathbf{E} - \mathbf{X}\mathbf{W} + \mathbf{X} + \tfrac{\Lambda_1}{\mu}||_F^2 \\ & + \tfrac{\mu}{2}||\mathbf{W} - \mathbf{S} + \tfrac{\Lambda_2}{\mu}||_F^2 \\ s.t. & S \geq 0, \textstyle\sum_j \mathbf{S}_{ij} = 1 \end{aligned} \tag{5}$$

where $\mu \in \mathbb{R}^{1\times 1}$ is a regularization parameter, $\Lambda_1 \in \mathbb{R}^{d\times n}$ and $\Lambda_2 \in \mathbb{R}^{n\times n}$ are ALM multipliers to enforce the auxiliary variable to be close to the target variable. $\mathbf{S}$ is first initialized with an efficient method [12]. Then we can optimize one variable while keeping the others fixed.

Update E:

When updating $\mathbf{E}$, we fix $\mathbf{W}$ and $\mathbf{S}$, and the objective function becomes

$$\min_{\mathbf{E}} ||\mathbf{E}||_{2,1} + \frac{\mu}{2}||\mathbf{E} - \mathbf{X}\mathbf{S} + \mathbf{X} + \frac{\Lambda_1}{\mu}||_F^2. \tag{6}$$

Denoting $\mathbf{M} = \mathbf{X}\mathbf{S} - \mathbf{X} - \frac{\Lambda_1}{\mu}$, then the problem is simplified into

$$\min_{\mathbf{E}} ||\mathbf{E}||_{2,1} + \frac{\mu}{2}||\mathbf{E} - \mathbf{M}||_F^2. \tag{7}$$

According to [18], the optimal $\mathbf{E}$ can be computed as

$$\mathbf{E}_{:,i} = \begin{cases} (1 - \frac{1}{\mu||\mathbf{M}_{:,i}||_2})\mathbf{M}_{:,i}, & \text{if } ||\mathbf{M}_{:,i}||_2 \geq \frac{1}{\mu} \\ 0, & \text{else} \end{cases}, \tag{8}$$

where $\mathbf{m}_i$ and $\mathbf{e}_i$ are the i-th column of $\mathbf{M}$ and $\mathbf{E}$ respectively.

Update W:

Optimizing $\mathbf{W}$ is equivalent to solving the following problem:

$$\min_{\mathbf{W}} ||\mathbf{E} - \mathbf{X}\mathbf{W} + \mathbf{X} + \tfrac{\Lambda_1}{\mu}||_F^2 + ||\mathbf{W} - \mathbf{S} + \tfrac{\Lambda_2}{\mu}||_F^2. \tag{9}$$

Denoting $\mathbf{A} = \mathbf{E} + \mathbf{X} + \frac{\Lambda_1}{\mu}$ and $\mathbf{B} = \mathbf{S} - \frac{\Lambda_2}{\mu}$, the above problem is reduced to

$$\min_{\mathbf{W}} ||\mathbf{X}\mathbf{W} - \mathbf{A}||_F^2 + ||\mathbf{W} - \mathbf{B}||_F^2. \tag{10}$$

Taking the derivation of Eq. (10) with respect to $\mathbf{W}$ and setting it to zero, we can get the optimal solution as

$$\mathbf{W} = (\mathbf{X}^T\mathbf{X} + \mathbf{I})^{-1}(\mathbf{X}^T\mathbf{A} + \mathbf{B}). \tag{11}$$

Update S:

With $\mathbf{E}$ and $\mathbf{W}$ fixed, problem (4) becomes

$$\begin{aligned} &\min_{\mathbf{S}} \lambda_1 \sum_{ij} ||\mathbf{x}_i - \mathbf{x}_j||_2^2 \mathbf{S}_{ij} + \tfrac{\mu}{2}||\mathbf{W} - \mathbf{S} + \tfrac{\Lambda_2}{\mu}||_F^2 \\ &s.t. \mathbf{S} \geq 0, \textstyle\sum_j \mathbf{S}_{ij} = 1 \end{aligned} \tag{12}$$

Denoting $\mathbf{C} = \mathbf{W} + \frac{\Lambda_2}{\mu}$, the above problem can be solved by optimizing each row separately:

$$\min_{\mathbf{S}} \lambda_1 \sum_j ||\mathbf{x}_i - \mathbf{x}_j||_2^2 \mathbf{S}_{ij} + \tfrac{\mu}{2} \sum_j ||\mathbf{C}_{ij} - \mathbf{S}_{ij}||_2^2 \tag{13}$$

Denoting $\mathbf{p}_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||_2^2$ and denoting a column vector $\mathbf{p}_i$ with the j-th element equal to $\mathbf{p}_{ij}$ (and denoting $\mathbf{s}_i$ and $\mathbf{c}_i$ similarly), the above problem can be written as a closed-form:

$$\min_{\mathbf{s}_i \geq 0, \mathbf{s}_i^T 1 = 1} ||\mathbf{s}_i - \mathbf{c}_i + \frac{\lambda_1}{2\mu}\mathbf{p}_i||_2^2, \tag{14}$$

which can be solved with an efficient iterative algorithm [14].

Update ALM Parameters:

The ALM parameters Λ_1, Λ_2 and μ are updated in each iteration according to [19]:

$$\begin{aligned} \Lambda_1 &= \Lambda_1 + \mu(\mathbf{E} - \mathbf{X}\mathbf{W} + \mathbf{X}), \\ \Lambda_2 &= \Lambda_2 + \mu(\mathbf{W} - \mathbf{S}), \\ \mu &= \rho\mu, \end{aligned} \tag{15}$$

where ρ is a parameter (set as 1.5 in the experiments).

During the optimization of each variable, a closed-form solution is obtained. So the objective value of ALM problem (5) decreases in each iteration. Moreover, problem (5) will converge to the original problem (4) as μ increases exponentially. So the proposed problem will converges to a local minimum value finally. And the details of the proposed RASL is summarized in Algorithm 1.

Algorithm 1 Robust Adaptive Sparse Learning

Input: Data matrix $\mathbf{X}$, cluster number c, parameter λ_1, ρ.

1: Initialize data graph $\mathbf{S}$.
2: **repeat**
3: Update $\mathbf{E}$ with Eq. (8).
4: Update $\mathbf{W}$ with Eq. (11).
5: Update $\mathbf{S}$ by solving problem (14).
6: Update Λ_1, Λ_2 and μ.
7: **until** Converge
8: Perform k-means with the spectral vectors of $\mathbf{S}$.

Output: Clustering results.

3. EXPERIMENTS

3.1. Performance on Toy Dataset

In this part, a toy dataset is built to validate the capability on exploiting local data structure.

Dataset: As shown in Fig. 1 (a), the toy dataset consists of 200 randomly generated points from two clusters. And the points from each cluster are distributed in the moon shape.

Competitors: The Self-Tuning Spectral Clustering (ST - SC) [20] and Simplex Sparse Representation (SSR) [14] are taken as competitors. STSC constructs the Gaussian graph with the self-tuned scaling parameter, while SSR learns the graph with sparse representation.

Performance: Fig. 1 (b)-(d) shows the graphs constructed by STSC, SSR and the proposed RASL. The points are connected with green lines if their similarities are larger than 0.01. It can be seen that the graphs of STSC and SSR incorrectly connect the points from different classes. And for RASL, there is no line between the two clusters, and all the points within the same cluster are connected together. STSC and SSR fail to capture the local relationship between points, so they cannot learn the graph with clear structure. The proposed RASL captures the local data relationship with the graph regularization term, so it is able to exploit the manifold structure. In addition, compared with STSC and SSR, RASL connects the points with less lines, which demonstrates that the graph learned by RASL is very sparse.

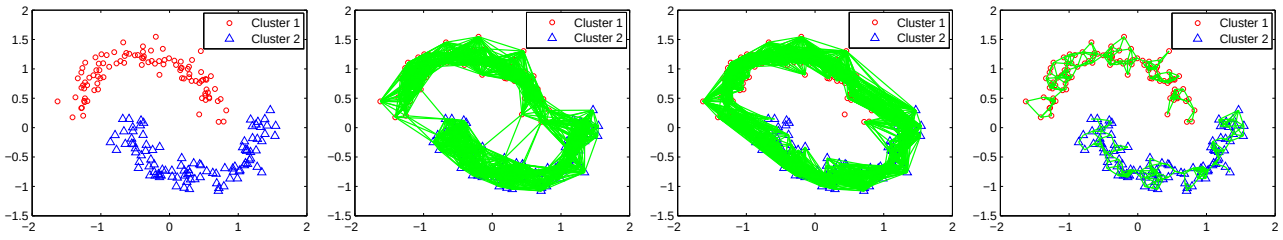
3.2. Performance on Real-World Datasets

Eight real-world datasets are used to evaluate the clustering performance of the proposed RASL.

Datasets: The experiments are conducted on eight datasets, including two handwritten digit dataset, i.e., Binary Alphabet (BA) [21] and subset of Mnist [22], one face dataset, i.e. Yale [23], one biology dataset, i.e., Carcinom [24], and four datasets from UCI Machine Learning Repository [25], i.e., Control, Dermatology, Movement and Iris.

Competitors: The proposed RASL is compared with eight state-of-the-art methods , including k-means, Ratio Cut (RCut) [6] Normalized Cut (NCut) [8], NMF [9], Constrained Laplacian Rank L1-norm (CLR_ L1) [12], Constrained Laplacian Rank L2-norm (CLR_ L2) [12], Clustering with Adaptive Neighbors (CAN) [13] and SSR [14]. For k-means, RCut and NCut, the affinity graph is constructed with the self-tune Gaussian method [20]. For CLR_ L1, CLR_ L2 and CAN, the graph is initialized by the method in [12], and the neighborhood size is set as 5. In addition, we repeat k-means, RCut, NCut and the proposed RASL for 200 times and report the average results, since all of them involve the k-means processing. And the parameter λ_1 of RASL is set as 0.1 empirically.

Performance: The quantitative comparison of different methods is shown in Table 1. Except for the lower ACC



(a) Data (b) STSC (c) SSR (d) RASL

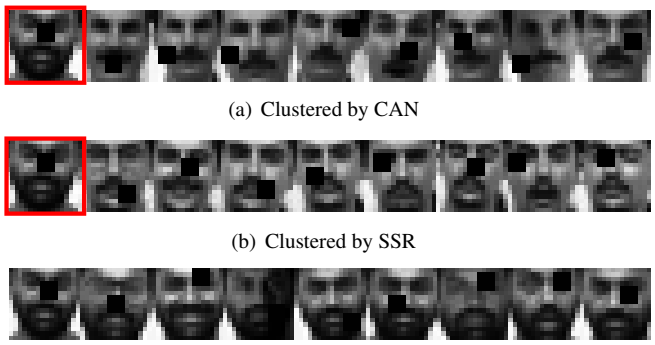
Fig. 1. Performance on toy dataset. Green lines connect the points whose similarities are larger than 0.01.

Table 1. The averaged ACC/NMI of different methods. The best results are in bold face.

Datasets	k-means	RCut	NCut	NMF	CLR_L1	CLR_L2	CAN	SSR	RASL
BA	0.41/0.57	0.15/0.24	0.42/0.57	0.18/0.30	0.28/0.44	0.13/0.19	0.31/0.40	0.11/0.26	**0.46/0.62**
Mnist	0.43/0.35	0.13/0.10	0.37/0.32	0.42/0.33	0.14/0.05	0.13/0.02	0.25/0.17	0.36/0.42	**0.45/0.46**
Yale	0.44/0.51	0.25/0.30	0.51/0.56	0.34/0.39	0.50/0.59	0.43/0.48	0.50/0.58	0.58/0.59	**0.59/0.61**
Carcinom	0.55/0.56	0.30/0.23	0.65/0.66	0.48/0.48	0.67/0.65	0.50/0.52	0.51/0.47	0.72/0.75	**0.73 /0.76**
Control	0.69/0.73	0.57/0.72	0.68/0.72	0.50/0.41	0.58/0.73	0.58/0.73	0.57/0.74	0.53/0.58	**0.71 /0.75**
Dermatology	0.70/0.80	0.66/0.69	0.92/0.88	0.70/0.64	0.84/0.87	0.83/0.86	0.83/0.85	**0.95**/0.91	0.94/**0.92**
Movement	0.44/0.57	0.40/0.52	0.46/0.59	0.36/0.42	0.43/0.61	0.48/0.63	0.48/0.63	0.26/0.30	**0.50 /0.64**
Iris	0.78/0.63	0.54/0.58	0.54/0.58	0.49/0.38	0.58/0.59	0.58/0.57	0.46/0.27	0.69/0.58	**0.83/0.64**

on Dermatology, the proposed RASL obtains the best performance under all circumstances. Especially on Iris, RASL outperforms the second best one a lot. k-means, RCut, NCut and NMF fail because they neglect the graph sparsity. CLR_L1, CLR_L2 and CAN learn a sparse graph by constraint the rank of the Laplacian matrix, while SSR borrows the sparse representation framework. But these methods can only capture the flat sparsity and ignore the local correlations between points. The proposed RASL is robust to noise with the $\ell_{2,1}$ norm objective function, and exploit the local data relationship to learn the affinity graph. So it achieves the best clustering performance.

Furthermore, to evaluate the robustness of RASL, we randomly occlude each face image in the Yale dataset with a 3×3 black area. Some representative clustering results of CAN, SSR and RASL are shown in Fig. 2. Both of CAN and SSR cluster the first image into the wrong class, while RASL performs well. Since the ℓ_2 norm objective functions of CAN and SSR square the residue error of each sample, they are sensitive to noise. The proposed RASL learns the graph with the $\ell_{2,1}$ norm, so it is robust to the data noise and achieves good performance on the occluded face data.



(a) Clustered by CAN

(b) Clustered by SSR

(c) Clustered by RASL

Fig. 2. Clustering results of CAN, SSR and RASL on occluded Yale dataset. Red square boxes show the incorrect results.

4. CONCLUSIONS

In this paper, a Robust Adaptive Sparse Learning method is proposed for graph clustering. The proposed method is able to learn a sparse data graph, and it is robust to data noise with the $\ell_{2,1}$ norm objective function. In addition, by capturing points' relationship adaptively, our method exploits the local data manifold during the graph learning procedure. So it is suitable to handle data with complex structure. In the optimization, the RASL objective function is decomposed into three closed-form sub-problems, which ensures its convergence. Extensive experiments on various real-world datasets show that RASL outperforms the state-of-the-art methods.

References

[1] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[2] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *AAAI*, 2017, pp. 4147–4153.

[3] C. Deng, X. Liu, C. Li, and D. Tao, "Active multi-kernel domain adaptation for hyperspectral image classification," *Pattern Recognition*, vol. 77, pp. 306–315, 2018.

[4] X. Yao, J. Han, D. Zhang, and F. Nie, "Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering," *IEEE TIP*, vol. 26, no. 7, pp. 3196–3209, 2017.

[5] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *NIPS*, 2001, pp. 849–856.

[6] P. Chan, M. Schlag, and J. Zien, "Spectral k-way ratio-cut partitioning and clustering," *IEEE Transactions on CAD ICS*, vol. 13, no. 9, pp. 1088–1096, 1994.

[7] Q. Wang, M. Chen, and X. Li, "Quantifying and detecting collective motion by manifold learning," in *AAAI*, 2017, pp. 4292–4298.

[8] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE TPAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[9] D. Li and H. Seung, "Algorithms for nonnegative matrix factorization," *NIPS*, vol. 13, no. 6, pp. 556–562, 2000.

[10] A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik, "Support vector clustering," *JMLR*, vol. 2, pp. 125–137, 2001.

[11] R. Mall and J. Suykens, "Kernel spectral document clustering using unsupervised precision-recall metrics," in *IJCNN*, 2015, pp. 1–7.

[12] F. Nie, X. Wang, M. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering," in *AAAI*, 2016, pp. 1969–1976.

[13] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *KDD*, 2014, pp. 977–986.

[14] J. Huang, F. Nie, and H. Huang, "A new simplex sparse learning model to measure data similarity for clustering," in *IJCAI*, 2015, pp. 3569–3575.

[15] C. Li, C. You, and R. Vidal, "Structured sparse subspace clustering: A joint affinity learning and subspace clustering framework," *IEEE TIP*, vol. 26, no. 6, pp. 2988–3001, 2017.

[16] Q. Liu, J. Deng, and D. Tao, "Dual sparse constrained cascade regression for robust face alignment," *IEEE TIP*, vol. 25, no. 2, pp. 700–712, 2016.

[17] Y. Yuan, G. Zhu, and Q. Wang, "Hyperspectral band selection by multitask sparsity pursuit," *IEEE TGRS*, vol. 53, no. 2, pp. 631–644, 2015.

[18] J. Huang, F. Nie, H. Huang, and C. Ding, "Robust manifold nonnegative matrix factorization," *KDD*, vol. 8, no. 3, pp. 11:1–11:21, 2013.

[19] F. Nie, H. Wang, H. Huang, and C. Ding, "Joint schatten p -norm and ℓ_p -norm robust matrix completion for missing value recovery," *KIS*, vol. 42, no. 3, pp. 525–544, 2015.

[20] L. Manor and P. Perona, "Self-tuning spectral clustering," in *NIPS*, 2004, pp. 1601–1608.

[21] P. Belhumeur, J. Hespanha, and D Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE TPAMI*, vol. 19, no. 7, pp. 711–720, 1997.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[23] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *IEEE TPAMI*, vol. 27, no. 3, pp. 328–340, 2005.

[24] J. Li, K. Cheng, S. Wang, F. Morstatter, T. Robert, J. Tang, and H. Liu, "Feature selection: A data perspective," *arXiv:1601.07996*, 2016.

[25] M. Lichman, "UCI machine learning repository," 2013.