SKETCHED SPARSE SUBSPACE CLUSTERING FOR LARGE-SCALE HYPERSPECTRAL IMAGES

Shaoguang Huang[1], *Hongyan Zhang*[2] *and Aleksandra Pižurica*[1]

[1]Department of Telecommunications and Information Processing, TELIN-GAIM, Ghent University, Belgium
[2]The State Key Lab. of Inform. Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China

ABSTRACT

Sparse subspace clustering (SSC) has achieved the state-of-the-art performance in clustering of hyperspectral images. However, the computational complexity of SSC-based methods is prohibitive for large-scale problems. We propose a large-scale SSC-based method, which processes efficiently large-scale HSIs without sacrificing the clustering accuracy. The proposed approach incorporates sketching of the self-representation dictionary reducing thereby largely the number of optimization variables. In addition, we employ a total variation (TV) regularization of the sparse matrix, resulting in a robust sparse representation. We derive a solver based on the alternating direction method of multipliers (ADMM) for the resulting optimization problem. Experimental results on real data show improvements over the traditional SSC-based methods in terms of accuracy and running time.

Index Terms— Sparse subspace clustering, sketching, hyperspectral image, large-scale data

1. INTRODUCTION

Hyperspectral images (HSIs) contain hundreds of spectral bands in the visible and infrared parts of the spectrum (400-2500 nm). As different materials usually show different spectral reflectance patterns, HSIs have become a powerful and valuable tool in remote sensing. HSI clustering aims at grouping similar pixels into different clusters without using any labelled samples. Two most widely used clustering methods are fuzzy c-means (FCM) [1] and k-means [2] due to their simplicity. However, their performance is sensitive to the initialization and to noise.

Sparse subspace clustering (SSC) approach [3] provides the current state-of-the-art performance in the clustering of HSIs [4–9]. SSC models a high-dimensional data space as a union of low-dimensional spaces. The key idea is that a sparse representation of a data point from this union of subspaces selects a few points from the same subspace [3]. Technically, the sparse coefficients are identified by solving the sparse coding problem under the self-representation dictionary. The coefficient matrix lends itself readily to the construction of the similarity matrix, which is further applied within the standard spectral clustering framework [10].

As the SSC model calculates the sparse coefficients independently for each pixel, its performance is sensitive to noise. Recent methods alleviate this problem by introducing different spatial constraints [4–9] on the coefficient matrix, promoting its smoothness, resulting in the improved clustering accuracy.

However, the application of any of these SSC-based methods on large-scale HSIs is practically infeasible due to the high computational complexity, resulting from iterative optimization. The time complexity in each iteration excedes $\mathcal{O}(n^3)$, where n is the number of pixels in the HSI, which renders SSC inappropriate for large-scale data. In recent years, some general large-scale methods based on SSC have been proposed for the clustering tasks in computer vision [11, 12] but were not reported yet on HSIs. In [11], SSC was first applied on a small amount of selected samples, and the rest were grouped by sparse representation classification with a dictionary constructed by the selected samples. The sketched SSC model of [12] uses a clever random projection technique to compact the self-representation dictionary. Despite the scalability of those methods, their clustering performances on HSIs turns out to be poor. This is attributed to the complex spatial structure of HSIs, spectral noise and spectral variability.

To address these problems, we propose a sketched SSC model with total variation (TV) spatial regularization, termed Sketch-SSC-TV. The employed sketching technique compresses the large self-representation dictionary to a much smaller one, which significantly reduces the number of optimization variables, and enables the application on large-scale HSIs. In addition, a spatial constraint in the form of TV-norm on the coefficient matrix is incorporated to promote the connectivity of neighbouring pixels and piecewise smoothness of

This work is supported in part by the Fonds voor Wetenschappelijk Onderzoek (FWO) project: G.OA26.17N and in part by the National Natural Science Foundation of China under grants 61871298 and 41711530709.

clustering maps. In order to solve the resulting optimization problem, we derive an iterative solver based on the alternating direction method of multipliers (ADMM) [13]. Experimental results on the small HSI and large-scale HSI demonstrate the superior performance of our method over the traditional SSC-based methods and the related state-of-the-art large-scale clustering methods.

The rest of this paper is organized as follows, Section 2 briefly introduces the SSC model for HSI clustering. Section 3 describes the proposed Sketch-SSC-TV model and the optimization problem. Section 4 presents the experimental results on real HSI data and Section 5 concludes the paper.

2. THE SSC MODEL OF HSI CLUSTERING

We denote by $\mathbf{Y} \in \mathbb{R}^{B\times MN}$ the flattened 2-D matrix from the original 3-D HSI data cube with a size of $M \times N \times B$, where M and N represent the height and the width of the HSI, respectively, and B denotes the number of bands. Each vector $\mathbf{y}_i \in \mathbb{R}^B$ represents the spectral signature of each pixel in HSI. We assume that there are c classes in the data, and each class is seen as a cluster. With the assumption of SSC that each pixel in a union of subspaces can be represented as a linear or affine combination of others from the same subspace with sparsity constraint, the sparse matrix $\mathbf{C} \in \mathbb{R}^{MN\times MN}$ can be calculated by:

$$\begin{aligned} &\arg\min_{\mathbf{C}} \|\mathbf{C}\|_1 + \frac{\lambda}{2}\|\mathbf{Y} - \mathbf{Y}\mathbf{C}\|_F^2 \\ &s.t.\ diag(\mathbf{C}) = \mathbf{0},\ \mathbf{1}^T\mathbf{C} = \mathbf{1}^T, \end{aligned} \tag{1}$$

where $\|\mathbf{C}\|_1 = \sum_i \sum_j |C_{ij}|$; $\mathbf{1}$ is an all-one vector; $diag(\mathbf{C})$ is a diagonal matrix whose entries outside the main diagonal are zero and λ is a parameter, which controls the balance between the data fidelity and the sparsity of the coefficient matrix. The first constraint is introduced to avoid the trivial solution of representing a sample by itself and the second constraint indicates the case of affine subspace.

The model in (1) can be solved by the ADMM algorithm [13]. As matrix $\mathbf{C}$ interprets the relationship between pixels according to the subspace preserving property of SSC, i.e. a non-zero entry C_{ij} indicates that the samples $\mathbf{y}_i$ and $\mathbf{y}_j$ are in the same class. Thus, the similarity matrix $\mathbf{W} \in \mathbb{R}^{MN\times MN}$ is constructed by $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}|^T$. Clustering results are obtained by employing the similarity matrix $\mathbf{W}$ within the spectral clustering [10]. Specifically, the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{MN\times MN}$ is first formed by

$$\mathbf{L} := \mathbf{D} - \mathbf{W} \tag{2}$$

where $\mathbf{D} \in \mathbb{R}^{MN\times MN}$ is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$ [14]. Afterwards, the c eigenvectors $\{\mathbf{v}_k\}_{k=1}^c$ of $\mathbf{L}$ corresponding to the c smallest eigenvalues of $\mathbf{L}$ are calculated via singular-value decomposition (SVD). Finally, the clustering result is obtained by applying the matrix $\mathbf{V} = [\mathbf{v_1}, ..., \mathbf{v_c}] \in \mathbb{R}^{MN\times c}$ to the k-means clustering method.

3. PROPOSED METHOD

A Sketch-SSC-TV model is proposed for the clustering of large-scale HSIs. Instead of the independent sparse coding in SSC, we incorporate a TV spatial constraint to promote the connectivity between neighbouring pixels, which guarantees the robustness to noise and spectral variability. Compared with the traditional SSC-based methods, a sketching technique is employed in the clustering model, resulting in a scalable clustering method.

3.1. Sketch-SSC-TV model

The optimization problem in (1) actually cannot be efficiently solved in practice due to its prohibitively high computational complexity. The traditional SSC-based methods [4–9] also suffer from the same problem. Their key obstacle is the need to calculate and save the inverse of the entire large matrix $(\mathbf{Y}^T\mathbf{Y} + \mu\mathbf{I}) \in \mathbb{R}^{MN\times MN}$ in memory, whose computation complexity reaches to $\mathcal{O}((MN)^3)$, which is infeasible for the large-scale data sets. Inspired by the sketching technique in [12], we replace the self-representation dictionary $\mathbf{Y}$ with a compact dictionary $\mathbf{D} \in \mathbb{R}^{B\times n} := \mathbf{Y}\mathbf{R}$ with a random matrix $\mathbf{R} \in \mathbb{R}^{MN\times n}$, which significantly reduces the number of optimization variables in the sparse matrix. In addition, we exploit a TV-norm based spatial regularization on the sparse matrix to strengthen the dependency between the pixels in the local regions, avoiding independent sparse coding of SSC. This yields a more robust clustering performance. The proposed Sketch-SSC-TV model with respect to sparse matrix $\mathbf{A} \in \mathbb{R}^{n\times MN}$ is given by

$$\arg\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_1 + \lambda_{tv}\|\mathbf{A}\|_{TV}, \tag{3}$$

where λ and λ_{tv} are the penalty parameters for the sparsity level and spatial smoothness, respectively, and the TV norm is defined by:

$$\|\mathbf{A}\|_{TV} = \|\mathbf{H}_x\mathbf{A}^T\|_1 + \|\mathbf{H}_y\mathbf{A}^T\|_1, \tag{4}$$

where $\mathbf{H}_x$ and $\mathbf{H}_y$ are the forward finite-difference operators in the horizontal and vertical directions, respectively, with periodic boundary conditions. The constraint $diag(\mathbf{C}) = \mathbf{0}$ in (1) is removed as identity matrix is not a trivial solution in (3).

Note that the sketched dictionary size n usually is far smaller than MN, and empirically can be set to the value less than 100. Thus the number of optimization variables in sparse matrix is significantly reduced from $(MN)^2$ to MNn. The resulting model can be efficiently solved by ADMM which is introduced next.

The sparse matrix $\mathbf{A}$, cannot be applied directly in the same way as in the traditional SSC-based methods for constructing the similarity matrix $\mathbf{W}$, because now $\mathbf{A}$ is not a square matrix. We construct $\mathbf{W}$ via k nearest neighbours (KNN) with $\mathbf{A}$. In order to reduce the computation burden, approximate nearest neighbours (ANN) [15] is applied. Finally, the similarity matrix $\mathbf{W}$ is plugged in the spectral clustering framework to obtain the clustering results.

3.2. Optimization

To solve the model in (3), we first introduce three auxiliary variables $\mathbf{B}, \mathbf{Z} \in \mathbb{R}^{n\times MN}$ and $\mathbf{U} \in \mathbb{R}^{2MN\times n}$ and derive the equivalent formulation:

$$\begin{aligned}\arg\min_{\mathbf{B},\mathbf{A},\mathbf{Z},\mathbf{U}} &\frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{B}\|_F^2+\lambda\|\mathbf{Z}\|_1+\lambda_{tv}\|\mathbf{U}\|_1 \\ &s.t.\ \ \mathbf{A}=\mathbf{B}, \mathbf{A}=\mathbf{Z}, \mathbf{H}\mathbf{A}^T=\mathbf{U} \end{aligned} \tag{5}$$

where $\mathbf{H}=[\mathbf{H}_x;\mathbf{H}_y]$ is the TV operator in spatial direction of HSIs.

We can solve the optimization problem (5) by minimizing the resulting augmented Lagrangian function as follows:

$$\begin{aligned}\frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{B}\|_F^2+\lambda\|\mathbf{Z}\|_1+\lambda_{tv}\|\mathbf{U}\|_1+\frac{\mu}{2}\|\mathbf{A}-\mathbf{B}+\frac{\mathbf{Y}_1}{\mu}\|_F^2 \\ +\frac{\mu}{2}\|\mathbf{A}-\mathbf{Z}+\frac{\mathbf{Y}_2}{\mu}\|_F^2+\frac{\mu}{2}\|\mathbf{H}\mathbf{A}^T-\mathbf{U}+\frac{\mathbf{Y}_3}{\mu}\|_F^2\end{aligned} \tag{6}$$

where $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{R}^{n\times MN}$ and $\mathbf{Y}_3 \in \mathbb{R}^{2MN\times n}$ are the Lagrange multipliers, and μ is a weighting parameter. To this end, we can solve the following subproblems iteratively based on ADMM.

1) Update $\mathbf{B}$: The subproblem for $\mathbf{B}$ is given by:

$$\arg\min_{\mathbf{B}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{B}\|_F^2+\frac{\mu}{2}\|\mathbf{A}^k-\mathbf{B}^k+\frac{\mathbf{Y}_1^k}{\mu}\|_F^2. \tag{7}$$

A closed form solution can be easily obtained by setting the first-order derivative to zero:

$$\mathbf{B}^{k+1}=(\mathbf{D}^T\mathbf{D}+\mu\mathbf{I})^{-1}(\mathbf{D}^T\mathbf{Y}+\mu\mathbf{A}^k+\mathbf{Y}_1^k). \tag{8}$$

2) Update $\mathbf{A}$: The subproblem for $\mathbf{A}$ is given by:

$$\begin{aligned}\arg\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{A}-\mathbf{B}^{k+1}+\frac{\mathbf{Y}_1^k}{\mu}\|_F^2+\frac{1}{2}\|\mathbf{A}-\mathbf{Z}^k+\frac{\mathbf{Y}_2^k}{\mu}\|_F^2 \\ +\frac{1}{2}\|\mathbf{H}\mathbf{A}^T-\mathbf{U}^k+\frac{\mathbf{Y}_3^k}{\mu}\|_F^2\end{aligned} \tag{9}$$

By setting the first-order derivative to zero, we can obtain:

$$\mathbf{A}^{k+1}=\mathcal{F}^{-1}\left[\frac{\mathbf{G}}{2+(\mathcal{F}(\mathbf{H}_x))^2+(\mathcal{F}(\mathbf{H}_y))^2}\right] \tag{10}$$

where $\mathbf{G}=\mathcal{F}(\mathbf{Z}^k+\mathbf{B}^{k+1}-\mathbf{Y}_1^k/\mu-\mathbf{Y}_2^k/\mu+(\mathbf{U}^{k^T}-\mathbf{Y}_3^{k^T}/\mu)\mathbf{H})$, and $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the FFT and the inverse FFT, respectively.

3) Update $\mathbf{Z}$: The subproblem for $\mathbf{Z}$ is given by:

$$\mathbf{Z}^{k+1}=\arg\min_{\mathbf{Z}} \lambda\|\mathbf{Z}\|_1+\frac{\mu}{2}\|\mathbf{A}^{k+1}-\mathbf{Z}+\frac{\mathbf{Y}_2^k}{\mu}\|_F^2, \tag{11}$$

which can be solved by a thresholding algorithm [13].

4) Update $\mathbf{U}$: The subproblem for $\mathbf{U}$ is given by:

$$\arg\min_{\mathbf{U}} \lambda_{tv}\|\mathbf{U}\|_1+\frac{\mu}{2}\|\mathbf{H}\mathbf{A}^{(k+1)^T}-\mathbf{U}+\frac{\mathbf{Y}_3^k}{\mu}\|_F^2, \tag{12}$$

which can be solved in the same way as $\mathbf{Z}$.

5) Update other parameters:

$$\begin{aligned}\mathbf{Y}_1^{k+1}&=\mathbf{Y}_1^k+\mu(\mathbf{A}^{k+1}-\mathbf{B}^{k+1}) \\ \mathbf{Y}_2^{k+1}&=\mathbf{Y}_2^k+\mu(\mathbf{A}^{k+1}-\mathbf{Z}^{k+1}) \\ \mathbf{Y}_3^{k+1}&=\mathbf{Y}_3^k+\mu(\mathbf{H}\mathbf{A}^{(k+1)^T}-\mathbf{U}^{k+1}).\end{aligned} \tag{13}$$

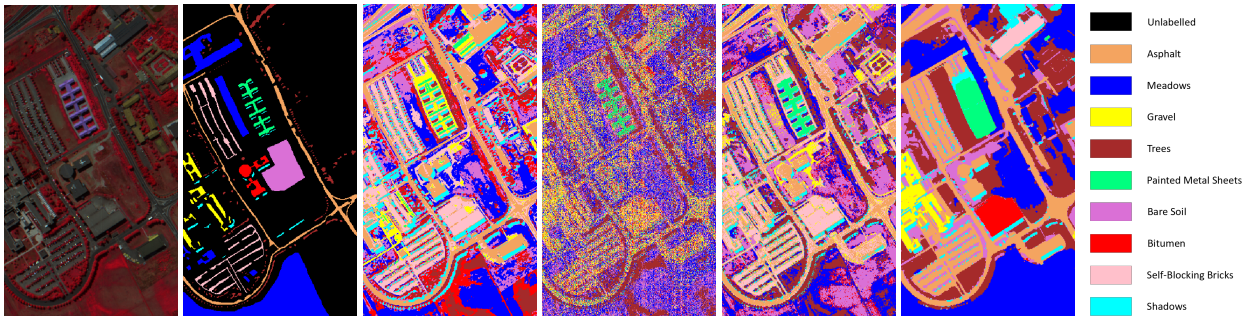
The above 5 steps are iteratively updated until the stopping criterion is satisfied.

4. EXPERIMENTAL RESULTS

We conduct experiments on two benchmark data sets: *Indian Pines* and *PaviaU*. *Indian Pines* has a spatial size of 85×70 and contains 200 spectral bands, including 4 classes such as corn-notill, grass-trees, soybean-notill and soybean-mintill. The *PaviaU* image size is $610 \times 340 \times 203$, including nine classes as shown in Fig. 1 (g). The clustering methods for comparison include a classical method k-means [2], the original SSC method [3], the SSC-based extensions L2-SSC [6] and the state-of-the-art large-scale clustering methods SSSC [11] and Ske-SSC [12]. Three performance measures: overall accuracy (OA), Kappa coefficient (κ) and running time (t) are used for quantitative assessment of the clustering results. All the methods are implemented in MATLAB on a computer with an Intel© core-i7 3930K CPU with 64 GB of RAM. The results of SSSC, Ske-SSC and the proposed method are reported in average of 5 runs. We set empirically $n = 70$ and $k = 30$ for our method in both data sets.

4.1. Experiments on real data

The first experiment is conducted on a small data set *Indian Pines*, and we report the results in Table 1. The optimal parameters of our method are $\lambda = 10^{-3}, \lambda_{tv} = 10^{-2}$. The results reported in Table 1 indicate that the proposed method yields the best performance in terms of OA and κ, achieving 23.35 % and 20.68% accuracy improvement over the traditional SSC-based methods. Compared with the large-scale SSSC and Ske-SSC methods, our method obtains the significant accuracy improvements with comparable running time.



(a) (b) (c) OA=53.41 (d) OA=33.78 (e) OA=50.63 (f) OA=61.15 (g)

Fig. 1. *PaviaU* image. (a) False color image, (b) Ground truth, and clustering maps of (c) k-means, (d) SSSC, (e) Ske-SSC and (f) Sketch-SSC-TV.

Table 1. Clustering results for *Indian Pines*.

No. of class	k-means	SSC	L2-SSC	SSSC	Ske-SSC	Proposed
1	69.85	60.00	61.09	53.31	62.19	61.41
2	53.84	98.36	99.32	89.73	**100**	**100**
3	0	76.91	79.37	49.13	68.80	**100**
4	57.59	50.68	54.89	63.85	58.87	**93.81**
OA(%)	50.17	65.11	67.78	63.28	68.12	**88.46**
κ	0.2833	0.5296	0.5629	0.4772	0.5628	**0.8342**
t (sec)	3	543	624	**2**	3	6

We conduct the second experiment on a much larger data set *PaviaU* consisting of 207400 pixels. The optimal parameters are $\lambda = 5 \times 10^{-2}, \lambda_{tv} = 5 \times 10^{-1}$. Due to the high memory requirement of SSC and L2-SSC, they cannot be run on our computer. The results in Table 2 demonstrate the effectiveness of our method on the large-scale data sets. Compared with SSSC, the accuracy improvement of our method is significant. Compared with Ske-SSC, our method also achieves much higher accuracy and comparable running time, which mainly benefits from the exploitation of the TV-norm based spatial regularization. Regarding the running time, k-means is the fastest clustering method. The clustering maps in Fig. 1 indicate that the proposed method suffers from less impulse noise than others, especially compared to the SSSC method.

Moreover, we analyse the effect of two important parameters λ and λ_{tv} in our model. The results in Fig. 2 indicate that the clustering accuracy is more stable with respect to λ than λ_{tv}. According to the results, we recommend to set $\lambda = 10^{-3}$ for all the data. For λ_{tv}, the value may be different for different data sets, but the clustering accuracy is stable and superior over other methods in a wide range.

5. CONCLUSION

In this paper, we proposed a scalable SSC-based clustering method, which incorporates the sketching technique by a random matrix and a TV-based spatial regularization. In order to solve the resulting model, we derived an efficient solver based on the ADMM algorithm. The experimental results conducted on both small HSI and large HSI verify that our method is much more effective than the traditional SSC-based methods and the related large-scale clustering methods.

Table 2. Clustering results for *PaviaU*.

No. of class	k-means	SSC	L2-SSC	SSSC	Ske-SSC	Proposed
1	90.51	-	-	35.60	64.88	**99.78**
2	43.83	-	-	28.47	42.55	**57.25**
3	0.10	-	-	11.92	**20.14**	19.43
4	63.67	-	-	61.29	**91.17**	75.05
5	48.25	-	-	62.05	99.79	**100**
6	32.89	-	-	18.56	27.94	**60.93**
7	0	-	-	**5.86**	0.38	0
8	**94.24**	-	-	31.48	65.11	0.15
9	**100**	-	-	8.91	75.73	73.75
OA	53.41	-	-	30.13	49.84	**58.71**
κ	0.4337	-	-	0.1794	0.3957	**0.4858**
t(sec)	**17**	-	-	30	838	974

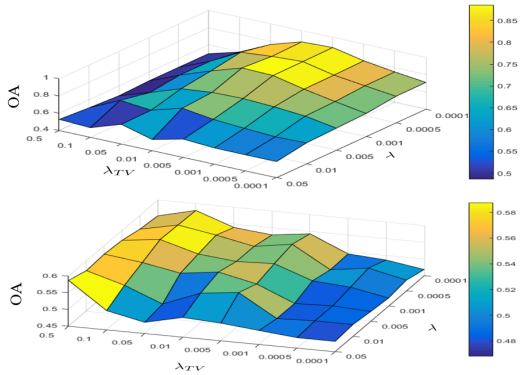


Fig. 2. Analysis for parameters λ and λ_{tv}. Top: *Indian Pines*; Bottom: *PaviaU*

6. REFERENCES

[1] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms," 1981.

[2] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[3] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[4] H. Zhang, H. Zhai, L. Zhang, and P. Li, "Spectral–spatial sparse subspace clustering for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3672–3684, 2016.

[5] H. Zhai, H. Zhang, X. Xu, L. Zhang, and P. Li, "Kernel sparse subspace clustering with a spatial max pooling operation for hyperspectral remote sensing data interpretation," *Remote Sensing*, vol. 9, no. 4, pp. 335, 2017.

[6] H. Zhai, H. Zhang, L. Zhang, P. Li, and A. Plaza, "A new sparse subspace clustering algorithm for hyperspectral remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 43–47, 2017.

[7] S. Huang, H. Zhang, and A. Pižurica, "Joint sparsity based sparse subspace clustering for hyperspectral images," in *Proc. IEEE ICIP*, 2018, pp. 3878–3882.

[8] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Total variation regularized collaborative representation clustering with a locally adaptive dictionary for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 166–180, 2018.

[9] S. Huang, H. Zhang, and A. Pižurica, "Semisupervised sparse subspace clustering method with a joint sparsity constraint for hyperspectral remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 12, no. 3, pp. 989–999, 2019.

[10] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[11] X. Peng, L. Zhang, and Z. Yi, "Scalable sparse subspace clustering," in *Proc. IEEE CVPR*, 2013, pp. 430–437.

[12] P. A. Traganitis and G. B. Giannakis, "Sketched subspace clustering," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1663–1675, 2018.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[14] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[15] Piotr Indyk and Rajeev Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM, 1998, pp. 604–613.