Heterogeneous Regularizations Based Tensor Subspace Clustering for Hyperspectral Band Selection

Shaoguang Huang, *Member, IEEE,* Hongyan Zhang, *Senior Member, IEEE,* Jize Xue, *Student Member, IEEE,* and Aleksandra Pižurica, *Senior Member, IEEE*

***Abstract*—Band selection (BS) reduces effectively the spectral dimension of a hyperspectral image (HSI) by selecting relatively few representative bands, which allows efficient processing in subsequent tasks. Existing unsupervised BS methods based on subspace clustering are built on matrix-based models, where each band is reshaped as a vector. They encode the correlation of data only in spectral-mode (dimension) and neglect strong correlations between different modes, i.e., spatial modes and spectral mode. Another issue is that the subspace representation of bands is performed in the raw data space, where the dimension is often excessively high, resulting in a less efficient and less robust performance. To address these issues, in this paper we propose a tensor based subspace clustering model for hyperspectral band selection. Our model is developed on the well-known Tucker decomposition. The three factor matrices and a core tensor in our model encodes jointly the multi-mode correlations of HSI, avoiding effectively to destroy the tensor structure and information loss. In addition, we propose well-motivated heterogeneous regularizations on the factor matrices by taking into account the important local and global property of HSI along three dimensions, which facilitates the learning of the intrinsic cluster structure of bands in the low-dimensional subspaces. Instead of learning the correlations of bands in the original domain, a common way for the matrix-based models, our model learns naturally the band correlations in a low-dimensional latent feature space, which is derived by the projections of two factor matrices associated with spatial dimensions, leading to a computationally efficient model. More importantly, the latent feature space is learned in a unified framework. We also develop an efficient algorithm to solve the resulting model. Experimental results on benchmark data sets demonstrate that our model yields improved performance compared to the state-of-the-art.**

***Index Terms*—Band selection, hyperspectral image, remote sensing, tensor, subspace clustering.**

I. Introduction

RECENT advances on hyperspectral sensors significantly increase the spectral resolution of hyperspectral images (HSIs) [1]. While providing richer spectral information than multi-spectral data, which enables improved discriminations between different materials, HSIs raise in the meantime the challenges of dealing with high-dimensional data [2]. On the one hand, high-dimensional data surges the burdens in the data acquisition, storage and transmission. On the other hand, it leads to the curse of dimensionality problem, which deteriorates the performance of the related post-processing algorithms [3, 4]. Both issues limit the real applications of HSIs.

Band selection (BS), as an effective dimension reduction method, selects the most relevant spectral bands from HSIs for post-processing such as classification, which reduces effectively the dimension of HSIs while obtaining comparable or even better performance in the related task [5, 6]. Depending on how supervised information is involved in the band selection, band selection methods can be categorized into supervised methods [7–9], semi-supervised methods [10–12] and unsupervised methods [13–17]. A recent overview for BS can be found in [18]. We here mainly focus on the unsupervised BS method given the fact that data labelling is expensive, leading to often a scarce of labeled data in practice [19, 20].

Common unsupervised BS methods include ranking-based [14, 21–26], searching-based [5, 27–31] and clustering-based methods [6, 16, 17, 32–38]. Ranking-based methods select the top-ranked bands as representatives according to the scores of bands measured by a given criterion such as band variation [21], mutual information [23] and band correlation [22, 24]. Searching-based methods select desired bands by greedy algorithm or evolutionary algorithms, including immune clone [30], firefly algorithms [31] and particle swarm [27]. Clustering-based methods cluster the bands of HSIs into different groups by well-developed clustering algorithms, including spectral clustering [39], hierarchical clustering [6], subspace clustering [37, 40] and probabilistic clustering [41]. As each of the resulting clusters contains spectral bands of high similarity, one can select a representative band from each cluster to obtain the optimal subset of bands.

In general, spectral clustering deals well with non-spherical cluster structure of bands by using graph spectral analysis. However, the performance of spectral clustering is sensitive to the neighborhood size and similarity measurement between

This work was supported in part by the Flanders AI Research Programme under Grant 174B09119, in part by the Bijzonder Onderzoeksfonds (BOF) under Grant BOF.24Y.2021.0049.01 and in part by the National Natural Science Foundation of China under Grant 61871298 and Grant 42071322. *(Corresponding author: Hongyan Zhang.)*

S. Huang and A. Pižurica are with the Department of Telecommunications and Information Processing, Ghent University, 9000 Ghent, Belgium (e-mail: shaoguang.huang@ugent.be; Aleksandra.Pizurica@ugent.be).

H. Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: zhanghongyan@whu.edu.cn).

J. Xue is with the Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China and also with the Image Processing and Interpretation, IMEC Research Group, Ghent University, Ghent, Belgium(e-mail: xuejize900507@mail.nwpu.edu.cn)

two bands to build a similarity matrix [42]. Hierarchical clustering learns a tree cluster structure of bands, which avoids repetitive clustering process of bands from scratch when a different number of bands is required to adapt to the complexity of subsequent tasks. However, hierarchical clustering is sensitive to noise, outliers and initial conditions. Moreover, wrong clustering results in previous stage cannot be reversed, resulting in error propagation to the following clustering stages. Probabilistic clustering based methods allow a more flexible interpretation of band clustering than others by providing soft clustering results, which deals well with smooth transition between clusters (because of smoothing spectrum). However, this type of methods often assume specific probability distributions for clusters, which might not be correct in practice, resulting in thereby unstable performance of band selection.

Recently, subspace clustering methods yield remarkable performances in unsupervised hyperspectral band selection [37]. The success mainly benefits from the accurate learning of cluster structure of bands in low-dimensional subspaces within a representation-based framework, which makes them robust to noise and outliers and efficient for clustering of high-dimensional data. The superior clustering performance facilitates a better band selection result. Subspace clustering typically consists of two steps: subspace representation and spectral clustering, among which the former is more important to the BS performance and received increasing attention [18, 36, 37]. Generally, subspace representation seeks for a lower-dimensional representation of bands given a dictionary, which is typically the input data itself [40]. The resulting representation in the low dimensional subspace reveals the correlations of bands and is utilized to construct a similarity matrix, which is further fed into spectral clustering to obtain clustering results [15, 18, 36, 37, 40]. Despite their excellent band selection performance, these methods show limitations on addressing the following two critical problems. The first problem is that current subspace clustering based BS methods are mostly built on matrix-based models where each 2-D spectral band in HSIs is reshaped to a vector being as a column/row of the input matrix. The vectorization operation on the bands inevitably destroys the spatial structure of HSIs, resulting in a loss of the internal structural information of data in terms of the local and global data property [2]. For example, the spatially local correlations between neighbouring pixels can be hardly preserved in the matricized data. Such loss of important spatial information might lead to a suboptimal learning of data representation in the low-dimensional subspace. The second problem comes from the fact that most of the existing approaches study the band correlations in the original domain. In the matrix-based subspace clustering models, as each data point is the concatenation of all the pixels in one band, the input data is often of very high dimensionality. The dimension of the vectorized bands is often several hundreds of thousands. It is obvious that analysing the cluster structure of these bands in the original domain is less efficient because of the high dimensionality. This can be even worse for large-scale data, which is often the case in the tasks of earth observations and environment monitoring. Moreover, band selection in the original domain also implies the analysis of band correlations using the raw data. However, due to the present of noise in HSI, this might result in a degraded performance.

Instead of matrix-based model, it is more favorable to process HSIs by using a higher-order model, e.g., a tensor-based model, where the loss of spatial information can be avoided. An Nth-order tensor is a multi-dimensional array with N indices. Hyperspectral image can be seen as a third-order tensor whose three indices correspond to the width and height in the 2-D spatial domain and the bands in the spectral domain, respectively. Recently, tensor-based methods have attracted extensive interests [43] and received a great success in various HSIs related tasks, including super-resolution [44], anomaly detection [45], unmixing [46], tensor completion [47, 48], denoising [49–51], compression [52, 53], feature extraction [54] and classification [55]. However, analogous research in the task of band selection is not explored in the literature.

Tucker decomposition (TD), as a representative tensor-based model, decomposes a tensor into a core tensor multiplied by a factor matrix along each mode (dimension) [56]. In this paper, we propose a heterogeneous regularizations based tensor subspace clustering (TSC) model with the decomposition form of TD for hyperspectral band selection, which addresses well the aforementioned problems. Unlike existing approaches, our model learns the correlations of bands in a tensor-based model, which processes the 3-D HSI directly and efficiently. To the best of our knowledge, this is the first attempt to tensor-based clustering model for band selection. Although there exist some tensor-based subspace clustering models, e.g., [57–61], they are designed for multi-view clustering tasks in the fileds of computer vision, including clustering of faces, objects and scenes, where traditional gray-scale or color images are used and specially they assume that different views of data are accessible. In the task of band selection, HSI often contains hundreds of channels, which results in a special data property of HSI in spectral domain. Existing tensor-based models in computer vision tasks fail to address this aspect. Moreover, the lack of multi-view data of HSIs makes the application of the tensor-based clustering models [57–61] in band selection infeasible.

The main contributions are summarized as follows:

1) We propose a novel tensor-based subspace clustering model for band selection for the first time. Unlike existing matrix-based models, we treat the 3-D HSI as a whole in a tensor-based model by using three factor matrices and a core tensor, which encodes jointly the multi-mode correlations of HSI, avoiding to destroy the inherent spatial structure of HSIs. Our model is more flexible and generalized in comparison with the matrix-based methods and in a certain condition, it can be reduced to the matrix-based model.
2) In order to capture well the complex 3-D structure of HSI, we take into account the prior information regarding both spatial dependencies of pixels and the spectral correlations of bands in TSC. By differentiating spatial modes from spectral mode, we propose accordingly heterogeneous regularizations (HRs) on the factor matrices

with each considering different aspects of the data property in terms of local and global geometric structure. The developed HRs consist of local and global constraints. The heterogeneous local constraints, imposed in the gradient domain of factor matrices in different modes, facilitate a robust and accurate representation of HSIs and improved discrimination between bands in the feature space. The heterogeneous global constraints on the factor matrices promote a low-rank tensor recovery and result in an efficient and interpretable learning algorithm. The developed HRs yield better band clustering and improve thereby subsequent band selection procedure.

3) Unlike existing approaches, which formulate subspace representation in the raw data space, we learn the cluster structure of bands in a projected data space, which is derived from the explicit projection by two factor matrices corresponding to the spatial modes of HSIs. The resulting dimension and redundancy in the projected data space are much lower than that in the original domain, facilitating a robust and efficient learning model. Moreover, the projections are learned jointly in a unified framework.
4) We devise an efficient algorithm based on alternating direction method of multipliers (ADMM) to optimize the proposed model. We provide theoretical analysis for the convergence property and time complexity of our algorithm. Experimental results on real data sets demonstrate that our model outperforms the state-of-the-art in the field.

The rest of this paper is organized as follows. Section II introduces the notations and basic concepts behind tensor decomposition, and reviews briefly matrix-based subspace clustering methods for band selection. Section III presents the proposed TSC model, develops an efficient algorithm to optimize the resulting model and provides a theoretical analysis on the convergence property and time complexity of the algorithm. Experiments and analysis on benchmark data sets are conducted in Section IV. Section V concludes the paper.

II. Preliminaries and Related Works

A. Notations and Preliminaries

We denote scalars by lowercase letters, e.g., x, vectors by boldface lowercase letters, e.g., $\mathbf{x}$, matrices by boldface capital letters, e.g., $\mathbf{X}$, and tensors by capital calligraphic letters, e.g., $\mathcal{X}$, throughout this paper. For a Nth-order (mode) tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, $x_{i_1 i_2 \cdots i_N}$ denotes the $(i_1, i_2, \cdots, i_N)$-th entry of $\mathcal{X}$. Mode-n fiber of $\mathcal{X}$ is denoted by $\mathbf{x}_{i_1, i_2, \cdots, i_{n-1}, :, i_{n+1}, \cdots, i_N}$, which consists of the entries by varying one index while fixing others. By arranging every mode-n fiber to be columns of a matrix, we obtain the mode-n unfolding of the tensor $\mathcal{X}$, which is denoted by $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times I_1 I_2, \cdots, I_{n-1} I_{n+1}, \cdots, I_N}$. The F-norm of a tensor is defined as $\|\mathcal{X}\|_F = (\sum_{i_1, i_2, \cdots, i_N} x_{i_1 i_2 \cdots i_N}^2)^{1/2}$. The Kronecker product of matrices $\mathbf{A} \in \mathbb{R}^{I \times J}$ and $\mathbf{B} \in \mathbb{R}^{K \times L}$ is denoted by $\mathbf{A} \otimes \mathbf{B}$, whose matrix size is $IK \times JL$.

Definition 1 (Tensor n-mode product): The n-mode product of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ with a matrix $\mathbf{U} \in \mathbb{R}^{J \times I_n}$ is denoted by

$$\mathcal{Y} = \mathcal{X} \times_n \mathbf{U}, \tag{1}$$

where $\mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N}$ and its elements are

$$\mathcal{Y}_{i_1 \cdots i_{n-1} j i_{n+1} \cdots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \cdots i_N} u_{j i_n}. \tag{2}$$

The n-mode product of a tensor with a matrix can be also represented in the unfolded matrix form:

$$\mathbf{Y}_{(n)} = \mathbf{U}\mathbf{X}_{(n)}. \tag{3}$$

Definition 2 (Tucker decomposition (TD)): The Tucker decomposition of a tensor $\mathcal{X}$ is defined as [56]:

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \cdots \times_N \mathbf{U}_N, \quad s.t.\ \mathbf{U}_n^T \mathbf{U}_n = \mathbf{I} \tag{4}$$

where $\mathbf{U}_n \in \mathbb{R}^{I_n \times R_n}$ $(R_n \leq I_n)$ are the column-orthogonal factor matrices with constraints $\mathbf{U}_n^T \mathbf{U}_n = \mathbf{I}$, and $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times \cdots \times R_N}$ is called the core tensor whose entries represent the interactions between the columns in different factor matrices. Elementwise, we have

$$x_{i_1 i_2 \cdots i_N} = \sum_{k_1, k_2, \cdots, k_N} g_{k_1 k_2 \cdots k_N} u_{1_{i_1 k_1}} u_{2_{i_2 k_2}} \cdots u_{N_{i_N k_N}}, \tag{5}$$

where $u_{j_{i_n k_n}}$ is the (i_n, k_n)-th entry of $\mathbf{U}_j$.

The Tucker decomposition (4) can be represented in the mode-n unfolding matrix form:

$$\mathbf{X}_{(n)} \approx \mathbf{U}_n \mathbf{G}_{(n)} (\mathbf{U}_N \otimes \cdots \otimes \mathbf{U}_{n+1} \otimes \mathbf{U}_{n-1} \otimes \cdots \otimes \mathbf{U}_1)^T. \tag{6}$$

Moreover, the Tucker decomposition can be reformulated in the vector form:

$$\text{vec}(\mathcal{X}) \approx (\mathbf{U}_N \otimes \cdots \otimes \mathbf{U}_{n+1} \otimes \mathbf{U}_{n-1} \otimes \cdots \otimes \mathbf{U}_1) \text{vec}(\mathcal{G}), \tag{7}$$

where $\text{vec}(\mathcal{X})$ denotes an operation to vectorize the tensor $\mathcal{X}$.

B. Subspace Clustering based Band Selection Models

Clustering based methods for band selection first employ clustering algorithms to assign bands into different clusters and then one representative band from each cluster is selected. Subspace clustering based methods also follow this procedure. Let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ be a hyperspectral image, where I_1 and I_2 are the spatial sizes and I_3 is the number of bands. Existing subspace clustering methods for band selection mostly unfold the tensor $\mathcal{X}$ as a mode-3 matrix, i.e., $\mathbf{X}_{(3)} \in \mathbb{R}^{I_3 \times I_1 I_2}$, and assume each band (each row of $\mathbf{X}_{(3)}$) can be represented by a linear combination of others in the same subspace. Given a self-representation dictionary $\mathbf{X}_{(3)}^T$, the resulting coefficients matrix $\mathbf{C} \in \mathbb{R}^{I_3 \times I_3}$ for such representation can be derived by solving the general optimization problem as follows:

$$\arg\min_{\mathbf{C}} \mathcal{L}(\mathbf{X}_{(3)}^T - \mathbf{X}_{(3)}^T \mathbf{C}) + \lambda \mathcal{R}(\mathbf{C}), \ s.t.\ \text{diag}(\mathbf{C}) = \mathbf{0} \tag{8}$$

where $\mathcal{L}(\cdot)$ is a given loss function which accounts for approximation error between $\mathbf{X}_{(3)}^T\mathbf{C}$ and the input data $\mathbf{X}_{(3)}^T$, $\mathcal{R}(\cdot)$ is a regularization term and λ is a parameter to balance the trade-off between data-fidelity and regularization terms. The constraint $\text{diag}(\mathbf{C}) = \mathbf{0}$ excludes a trivial solution of $\mathbf{C}$ to be an identity matrix. Commonly adopted loss function $\mathcal{L}(\cdot)$ is the Frobenius norm $\|\cdot\|_F$. According to the prior information of data, representative methods impose regularizations on coefficients matrix $\mathbf{C}$ such as sparsity term $\|\mathbf{C}\|_1 = \sum_i \sum_j |C_{ij}|$ [42], ℓ_2-norm based collaborative term $\|\mathbf{C}\|_F^2 = \sum_i \sum_j C_{ij}^2$ [40] and low-rank term $\|\mathbf{C}\|_* = \sum_i \sigma_i$ [15, 36, 37], where σ_i are the singular values of $\mathbf{C}$, aiming to capture better the underlying cluster structure of data.

The learned coefficients matrix $\mathbf{C}$ in (8) is used to construct the similarity matrix by $\mathbf{W} = (|\mathbf{C}|+|\mathbf{C}^T|)/2$, which is further applied into the standard spectral clustering [62] to obtain the band clustering results. By selecting the closest band to the centroid in each cluster, the representative bands are obtained.

III. The Tensor-based Subspace Clustering for Hyperspectral Band Selection

A. Problem Formulation

The matrix-based model in (8) requires to flatten each 2-D spectral band to a vector, which inevitably destroys the spatial structure of HSIs. Moreover, the learning of latent representations of bands in the original domain is less efficient, especially for large-scale HSIs due to the high dimensionality of the vectorized bands. In order to solve these problems, we propose a tensor-based subspace clustering mode for band selection. We borrow the decomposition form of (4) and represent a 3-D HSI $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ as follows:

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3, \tag{9}$$

where $\mathbf{U}_1 \in \mathbb{R}^{I_1 \times R_1}$ and $\mathbf{U}_2 \in \mathbb{R}^{I_2 \times R_2}$ are the factor matrices associated with the spatial dimensions; $\mathbf{U}_3 \in \mathbb{R}^{I_3 \times R_3}$ is the factor matrix corresponding to the spectral dimension and $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ is a core tensor. Compared to the matrix-based model, this tensor representation encodes the multimode correlation of data simultaneously. Specifically, according to (6) we derive the single-mode representations of $\mathcal{X}$ in each dimension as follows:

$$\mathbf{X}_{(1)} \approx \mathbf{U}_1\mathbf{G}_{(1)}(\mathbf{U}_3 \otimes \mathbf{U}_2)^T, \tag{10}$$

$$\mathbf{X}_{(2)} \approx \mathbf{U}_2\mathbf{G}_{(2)}(\mathbf{U}_3 \otimes \mathbf{U}_1)^T, \tag{11}$$

$$\mathbf{X}_{(3)} \approx \mathbf{U}_3\mathbf{G}_{(3)}(\mathbf{U}_2 \otimes \mathbf{U}_1)^T, \tag{12}$$

where $\mathbf{X}_{(1)} \in \mathbb{R}^{I_1 \times I_2 I_3}$, $\mathbf{X}_{(2)} \in \mathbb{R}^{I_2 \times I_1 I_3}$, $\mathbf{X}_{(3)} \in \mathbb{R}^{I_3 \times I_1 I_2}$, $\mathbf{G}_{(1)} \in \mathbb{R}^{R_1 \times R_2 R_3}$, $\mathbf{G}_{(2)} \in \mathbb{R}^{R_2 \times R_1 R_3}$ and $\mathbf{G}_{(3)} \in \mathbb{R}^{R_3 \times R_1 R_2}$.

It is observed that the three single-mode representations are mutually correlated by the factor matrices $\{\mathbf{U}_i\}_{i=1}^3$ and the core tensor $\mathcal{G}$. Thus, a faithful learning of $\{\mathbf{U}_i\}_{i=1}^3$ and $\mathcal{G}$ is essential. The solution to (9) in the least square sense involves solving an ill-posed optimization problem. This can be solved by introducing appropriate regularization on factor matrices

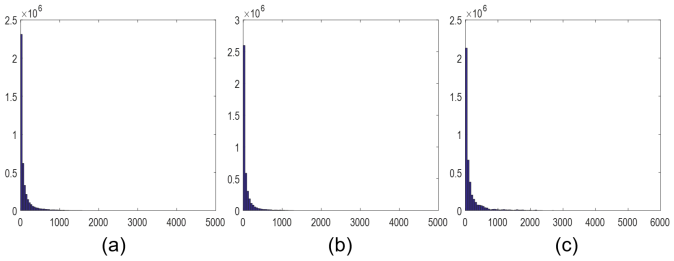

Fig. 1. The histograms of difference matrices (a) $\nabla_{y_1}\mathbf{X}_{(1)}$, (b) $\nabla_{y_2}\mathbf{X}_{(2)}$ and (c) $\nabla_{y_3}\mathbf{X}_{(3)}$ on the data set *Indian Pines*. The results show that most of the values in difference matrices are zero or near zero, indicating thereby the strong local correlations of HSI in both spatial and spectral dimensions.

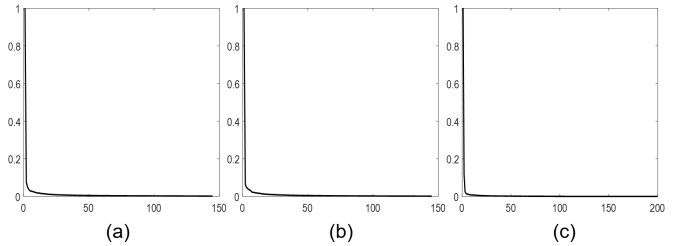

Fig. 2. The eigenvalues in three unfolding matrices: (a) $\mathbf{X}_{(1)}$, (b) $\mathbf{X}_{(2)}$ and $\mathbf{X}_{(3)}$ (the largest eigenvalue is normalized to one). Most of the eigenvalues of unfolding matrices are close to zero, showing the important low-rank properties of HSI along each mode.

according to the prior information of data. To this end, we formulate our basic TSC model as follows:

$$\underset{\mathcal{G},\mathbf{U}_i}{\arg\min}\ \frac{1}{2}\|\mathcal{X} - [\![\mathcal{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]\!]\|_F^2 + \sum_{i=1}^{3} \Phi_i(\mathbf{U}_i), \tag{13}$$

where $[\![\mathcal{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]\!] = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$ and $\Phi_i(\mathbf{U}_i)$ are heterogeneous constraints on factor matrices $\mathbf{U}_i$, which will be detailed later.

Taking the transpose of both sides of (12), we have:

$$\mathbf{X}_{(3)}^T \approx (\mathbf{U}_2 \otimes \mathbf{U}_1)\mathbf{G}_{(3)}^T\mathbf{U}_3^T. \tag{14}$$

As each column of $\mathbf{X}_{(3)}^T$ is a spectral band, we can view the matrix $(\mathbf{U}_2 \otimes \mathbf{U}_1)\mathbf{G}_{(3)}^T$ as the dictionary and $\mathbf{U}_3^T$ as the corresponding coefficients matrix, which represents the latent features of bands in the TSC model. While solving the problem (13), we utilize the factor matrix $\mathbf{U}_3$ to build a K-nearest neighbours (KNN) graph for spectral clustering. Then, we select the closest band to the centroid in each cluster as representatives of HSI.

B. Heterogeneous Regularizations Based TSC Model

The well-know Tucker decomposition [56, 63] leverages a tensor rank denoted by $\text{rank}_t(\mathcal{X})$, which is defined as $\text{rank}_t(\mathcal{X}) := (\text{rank}(\mathbf{X}_{(1)}), \text{rank}(\mathbf{X}_{(2)}), \text{rank}(\mathbf{X}_{(3)}))$, and imposes homogeneous constraints $\mathbf{U}_i^T\mathbf{U}_i = \mathbf{I}$ for the learning of a low-rank tensor representation. Although this tensor rank works effectively for TD to capture the global low-rank data structure, in the case of HSI the important local structure of data can be hardly unveiled, which leads to a partially learned data property in the low-dimensional data representations.

Here, we consider specially the local and global data structure of HSIs along three dimensions at the same time, which motives accordingly different constraints on the three factor matrices.

1) Local constraints: The local and global property of HSIs can be observed both in the spatial and spectral dimensions. We start with the local property of HSIs and elaborate here the local constraints in TSC. In the spatial dimensions, neighbouring pixels often belong to the same class and thereby their spectral signatures are very similar, which is commonly referred to spatial information. In the spectral dimension, due to the high spectral resolution of HSI, the neighbouring bands often have strong correlations, which means that most of the neighbouring bands have similar spectral intensities. Let $\nabla_{y_i}\mathbf{X}_{(i)}$ be the difference matrices of $\mathbf{X}_{(i)}$, where ∇_{y_i} is the corresponding difference operator for $\mathbf{X}_{(i)}$ in the vertical direction. We show the statistical distributions of $\nabla_{y_i}\mathbf{X}_{(i)}$ by their histograms in Fig. 1 on a benchmark HSI *Indian Pines*. The histograms show that most of the difference matrices are around zero, which demonstrates the local homogeneity of HSIs in all three dimensions.

The aforementioned spatial information of HSIs indicates the important intra-signal dependencies of $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$, which means that each column of them is smoothing. As shown in (10) and (11), the factor matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ can be viewed as representation basis in the mode-1 and mode-2 unfolding matrix form. Taking into account the intra-signal dependencies in mode-1 and mode-2, we focus on learning smoothing basis of $\mathbf{U}_1$ and $\mathbf{U}_2$ in TSC by introducing the local regularizations:

$$\Phi_i(\mathbf{U}_i) = \|\nabla_{y_i}\mathbf{U}_i\|_1,\ (i = 1, 2). \tag{15}$$

The ℓ_1 norm of the two regularizations promotes the difference matrices $\{\nabla_{y_i}\mathbf{U}_i\}_{i=1}^2$ to be sparse and induces thereby the smoothness of $\mathbf{U}_1$ and $\mathbf{U}_2$. The smoothing basis imply the smoothness of any data point represented over the basis [64], which allows our model preserve well the inherent local property of HSI in the spatial domain as shown in Fig. 1 (a) and (b). As evidenced in the matrix-based model [65], the smoothness of basis will also favor a robust representation to non-smooth noise that present in the input data.

We treat $\mathbf{U}_3$ differently with $\mathbf{U}_1$ and $\mathbf{U}_2$ regarding the local data property. In TSC model, we aim to produce consistent low-dimensional representations of bands such that the bands belonging to the same subspace have similar latent features while the bands belong to different subspaces obtain substantially different features. This helps improve the intra-cluster coherence and inter-clusters discriminability, which is beneficial for a better clustering performance. Aforementioned analysis from Fig. 1 (c) verifies the strong correlations of neighbouring bands, and thus it is reasonable to assume that they often belong to the same cluster and have similar coefficients vectors, i.e., the neighbouring columns of $\mathbf{U}_3^T$. Let $\nabla_x\mathbf{U}_3^T$ be the difference matrix of $\mathbf{U}_3^T$ in the horizontal direction. Therefore, $\nabla_x\mathbf{U}_3^T$ ideally should be a column-wise sparse matrix. Equivalently, $\nabla_{y3}\mathbf{U}_3$ is row-wise sparse in the ideal case. Thus, we utilize an $\ell_{1,2}$ norm based local regularization $\|\nabla_{y3}\mathbf{U}_3\|_{1,2}$ on the factor matrix $\mathbf{U}_3$. It is known that $\ell_{1,2}$ norm promotes effectively row-sparsity of a matrix, which improves thereby the consistency of subspace representation. This regularization treats all the bands equally with the same weight of 1. However, it is more favorable to discriminate the bands with different weights according to the prior information of data. The idea is that the neighbouring bands with a smaller distance are more likely to have similar coefficients vectors. Thus, we define a diagonal weight matrix $\mathbf{W}$ as follows:

$$W_{ii} = e^{-\frac{\sum_j D_{i,j}^2}{2\sigma^2}}, \tag{16}$$

where $\mathbf{D} = \nabla_{y_3}\mathbf{X}_{(3)}$ is the difference matrix of $\mathbf{X}_{(3)}$ and σ is a parameter. Then we propose a weighted local regularization:

$$\Phi_3(\mathbf{U}_3) = \|\mathbf{W}\nabla_{y_3}\mathbf{U}_3\|_{1,2}. \tag{17}$$

It is observed that the neighbouring bands with smaller distance yield a larger weight in (16), which regularizes the associated row of $\nabla_{y_3}\mathbf{U}_3$ in (17) to be zeros more likely than others.

2) Global constraints: The low rankness of HSIs is often considered as an important global data structure in the processing of HSIs. Such global property can be observed along three dimensions because of the connectivities of pixels and correlated bands. In other words, $\{\mathbf{X}_{(i)}\}_{i=1}^3$ are low rank. We show an example of the eigenvalues of $\{\mathbf{X}_{(i)}\}_{i=1}^3$ in Fig. 2 on the data set *Indian Pines*, where we can see most of the eigenvalues are around zero, which verifies the low-rankness of real HSIs along three dimensions. Existing matrix-based models can only capture the global structure along a single dimension, while tensor-based model enjoys the benefit of a joint learning of the global data property in three dimensions. To explore the global property of HSIs, we introduce the orthogonal constraints $\mathbf{U}_i^T\mathbf{U}_i = \mathbf{I}\ (i = 1, 2)$ on factor matrices $\{\mathbf{U}_i\}_{i=1}^2$ and utilize a nuclear norm $\|\mathbf{U}_3\|_*$ for $\mathbf{U}_3$. Thus we obtain:

$$\Phi_3(\mathbf{U}_3) = \|\mathbf{W}\nabla_{y_3}\mathbf{U}_3\|_{1,2} + \beta\|\mathbf{U}_3\|_*, \tag{18}$$

where β is a parameter.

The orthogonal constraints $\mathbf{U}_i^T\mathbf{U}_i = \mathbf{I}$ are important here. On the one hand, it guarantees the low-rank property of $\mathcal{X}$ in the first two dimensions, i.e., $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$ [56]. On the other hand, they enables a learning of bands correlation in a projected data space. To show it explicitly, we reformulate (9) equivalently to the equation:

$$\mathcal{X} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \approx \mathcal{G} \times_3 \mathbf{U}_3, \tag{19}$$

which leverages the orthogonal property of $\mathbf{U}_1$ and $\mathbf{U}_2$. Let $\mathcal{X}_p = \mathcal{X} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \in \mathbb{R}^{R_1\times R_2\times I_3}$ be the projected data of $\mathcal{X}$, and we represent (19) in the mode-3 unfolding form:

$$\mathbf{X}_{p_{(3)}}^T \approx \mathbf{G}_{(3)}^T\mathbf{U}_3^T. \tag{20}$$

This indicates that the factor matrix $\mathbf{U}_3$ is actually learned from the projected data $\mathbf{X}_{p_{(3)}}$. Due to the low-rank property of $\mathcal{X}$, R_1 and R_2 are often much smaller than the original spatial size of data, I_1 and I_2, respectively, leading to a more computationally efficient model. Compared with the matrix-based models, we here learn the representation of bands in

the projected data space instead of using the raw data, which can be less affected by the noise.

Considering the proposed heterogeneous regularizations, we formulate our TSC model formally by:

$$\begin{aligned}\underset{\mathcal{G},\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3}{\arg\min}\ & \frac{1}{2}\|\mathcal{X}-[\![\mathcal{G};\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3]\!]\|_F^2+\sum_{i=1}^{2}\alpha_i\|\nabla_{y_i}\mathbf{U}_i\|_1\\ &+\alpha_3\|\mathbf{W}\nabla_{y_3}\mathbf{U}_3\|_{1,2}+\beta\|\mathbf{U}_3\|_*+\lambda\|\mathcal{G}\|_F^2\\ &s.t.\ \mathbf{U}_1^T\mathbf{U}_1=\mathbf{I},\ \mathbf{U}_2^T\mathbf{U}_2=\mathbf{I},\end{aligned}\tag{21}$$

where λ, α_i and β are the parameters to control the trade-off between different constraints and $\|\mathcal{G}\|_F^2$ is adopt to avoid overfitting problem.

Our TSC model is more generic than the matrix-based models (8). When $\mathbf{U}_1=\mathbf{I}$, $\mathbf{U}_2=\mathbf{I}$ and $\mathcal{G}=\mathcal{X}$, our model reduces to the traditional matrix-based approach (8). In other words, the matrix-based model can be seen as a special case of our tensor-based model. Thus, our model is more flexible. Also the learning of the projection matrices and band representation in the projected data space are performed jointly in a unified model, avoiding effectively the exploitation of external dimensionality reduction approach. Our TSC model takes into account the prior information of HSIs in terms of local and global data property along three dimensions, which is expected to capture better the complex 3-D structure of HSI than the matrix-based models.

C. Optimization

To solve the problem in (21), we introduce auxiliary variables $\{\mathbf{U}_i=\mathbf{V}_i\}_{i=1}^3$, $\{\nabla_{y_i}\mathbf{U}_i=\mathbf{B}_i\}_{i=1}^3$ and $\mathbf{U}_3=\mathbf{Z}$, and reformulate equivalently (21) to the optimization problem:

$$\begin{aligned}\min\ & \frac{1}{2}\|\mathcal{X}-\mathcal{G}\times_1\mathbf{V}_1\times_2\mathbf{V}_2\times_3\mathbf{V}_3\|_F^2+\sum_{i=1}^{2}\alpha_i\|\mathbf{B}_i\|_1\\ &+\alpha_3\|\mathbf{W}\mathbf{B}_3\|_{1,2}+\beta\|\mathbf{Z}\|_*+\lambda\|\mathcal{G}\|_F^2\\ &s.t.\ \mathbf{U}_1^T\mathbf{U}_1=\mathbf{I},\ \mathbf{U}_2^T\mathbf{U}_2=\mathbf{I},\ \{\mathbf{U}_i=\mathbf{V}_i\}_{i=1}^3\\ &\{\nabla_{y_i}\mathbf{U}_i=\mathbf{B}_i\}_{i=1}^3,\ \mathbf{U}_3=\mathbf{Z}\end{aligned}\tag{22}$$

We obtain the augmented Lagrangian function of (22) by:

$$\begin{aligned}&L(\mathcal{G},\mathbf{U}_i,\mathbf{V}_i,\mathbf{B}_i,\mathbf{Z})\\ &=\frac{1}{2}\|\mathcal{X}-\mathcal{G}\times_1\mathbf{V}_1\times_2\mathbf{V}_2\times_3\mathbf{V}_3\|_F^2+\sum_{i=1}^{2}\alpha_i\|\mathbf{B}_i\|_1\\ &+\alpha_3\|\mathbf{W}\mathbf{B}_3\|_{1,2}+\beta\|\mathbf{Z}\|_*+\lambda\|\mathcal{G}\|_F^2+\sum_{i=1}^{3}(\langle\mathbf{Y}_i^v,\mathbf{U}_i-\mathbf{V}_i\rangle\\ &+\frac{\mu}{2}\|\mathbf{U}_i-\mathbf{V}_i\|_F^2+\langle\mathbf{Y}_i^b,\nabla_{y_i}\mathbf{U}_i-\mathbf{B}_i\rangle+\frac{\mu}{2}\|\nabla_{y_i}\mathbf{U}_i-\mathbf{B}_i\|_F^2)\\ &+\langle\mathbf{Y}^z,\mathbf{U}_3-\mathbf{Z}\rangle+\frac{\mu}{2}\|\mathbf{U}_3-\mathbf{Z}\|_F^2,\end{aligned}\tag{23}$$

where $\mathbf{Y}_i^v$, $\mathbf{Y}_i^b$ and $\mathbf{Y}^z$ are the Lagrange multipliers and μ is a penalty parameter. We then optimize the problem based on the ADMM algorithm by updating one variable while fixing others.

Algorithm 1 Cayley transform based algorithm to solve (28)

Input: $\mathbf{V}_i$, $\mathbf{B}_i$, ∇_{y_i}, $\mathbf{Y}_i^v$, $\mathbf{Y}_i^b$ and μ
Initialization: $\mathbf{U}_i^0$ is a random matrix with ${\mathbf{U}_i^0}^T\mathbf{U}_i^0=\mathbf{I}$, $k=0$, $\varepsilon=10^{-5}$ and $k_{max}=1000$.
while *true* **do**
  Step 1: Compute the derivative of the objective function (28) at $\mathbf{U}_i^k$ by
  $\mathbf{H}^k=\mathbf{U}_i^k-\mathbf{V}_i+\frac{\mathbf{Y}_i^v}{\mu}+\nabla_{y_i}^T(\nabla_{y_i}\mathbf{U}_i^k-\mathbf{B}_i+\frac{\mathbf{Y}_i^b}{\mu})$
  Step 2: Obtain $\mathbf{E}^k=[\mathbf{H}^k,\mathbf{U}_i^k]$; $\mathbf{F}^k=[\mathbf{U}_i^k,-\mathbf{H}^k]$; $\mathbf{R}^k=\mathbf{E}^k{\mathbf{F}^k}^T$
  Step 3: Compute the optimal step size τ_k to update $\mathbf{U}_i$ by linear search algorithm [66].
  Step 4: Update $\mathbf{U}_i$ by
  $\mathbf{U}_i^{k+1}=\mathbf{U}_i^k-\tau_k\mathbf{E}^k(\mathbf{I}+\frac{\tau_k}{2}{\mathbf{F}^k}^T\mathbf{E}^k)^{-1}{\mathbf{F}^k}^T\mathbf{U}_i^k$
  Step 5: Stopping check: If $\|\mathbf{R}^k\mathbf{U}_i^k\|<\varepsilon$ or $k>k_{max}$, then **STOP**; Otherwise, $k:=k+1$ and continue.
end while
Output: $\mathbf{U}_i$

1) Update $\{\mathbf{V}_i\}_{i=1}^3$: The objective function with respect to $\mathbf{V}_1$ is shown as

$$\begin{aligned}\underset{\mathbf{V}_1}{\arg\min}\ & \frac{1}{2}\|\mathcal{X}-\mathcal{G}\times_1\mathbf{V}_1\times_2\mathbf{V}_2\times_3\mathbf{V}_3\|_F^2\\ &+\frac{\mu}{2}\|\mathbf{U}_1-\mathbf{V}_1+\frac{\mathbf{Y}_1^v}{\mu}\|_F^2.\end{aligned}\tag{24}$$

Based on (6), we reformulate the (24) in the mode-1 unfolding matrix form:

$$\begin{aligned}\underset{\mathbf{V}_1}{\arg\min}\ & \frac{1}{2}\|\mathbf{X}_{(1)}-\mathbf{V}_1\mathbf{G}_{(1)}(\mathbf{V}_3\otimes\mathbf{V}_2)^T\|_F^2\\ &+\frac{\mu}{2}\|\mathbf{U}_1-\mathbf{V}_1+\frac{\mathbf{Y}_1^v}{\mu}\|_F^2.\end{aligned}\tag{25}$$

By setting the derivative of (25) with respect to $\mathbf{V}_1$ to zero, we update $\mathbf{V}_1$ as follows:

$$\mathbf{V}_1=(\mathbf{X}_{(1)}\mathbf{P}_1^T+\mu\mathbf{U}_1+\mathbf{Y}_1^v)(\mathbf{P}_1\mathbf{P}_1^T+\mu\mathbf{I})^{-1},\tag{26}$$

where $\mathbf{P}_1=\mathbf{G}_{(1)}(\mathbf{V}_3\otimes\mathbf{V}_2)^T$. We can update $\mathbf{V}_2$ and $\mathbf{V}_3$ similarly to $\mathbf{V}_1$ by:

$$\mathbf{V}_i=(\mathbf{X}_{(i)}\mathbf{P}_i^T+\mu\mathbf{U}_i+\mathbf{Y}_i^v)(\mathbf{P}_i\mathbf{P}_i^T+\mu\mathbf{I})^{-1}\tag{27}$$

where $\mathbf{P}_2=\mathbf{G}_{(2)}(\mathbf{V}_3\otimes\mathbf{V}_1)^T$ and $\mathbf{P}_3=\mathbf{G}_{(3)}(\mathbf{V}_2\otimes\mathbf{V}_1)^T$.

2) Update $\{\mathbf{U}_i\}_{i=1}^2$: We optimize the following optimization problem to update $\mathbf{U}_i(i=1,2)$:

$$\begin{aligned}\underset{\mathbf{U}_i}{\arg\min}\ & \frac{1}{2}\|\mathbf{U}_i-\mathbf{V}_i+\frac{\mathbf{Y}_i^v}{\mu}\|_F^2+\frac{1}{2}\|\nabla_{y_i}\mathbf{U}_i-\mathbf{B}_i+\frac{\mathbf{Y}_i^b}{\mu}\|_F^2\\ &s.t.\ \mathbf{U}_i^T\mathbf{U}_i=\mathbf{I}\end{aligned}\tag{28}$$

To solve this problem, we adopt a Cayley transform based algorithm [66], which deals effectively with the optimizations with orthogonality constraints. A detailed explanation to the algorithm in [66] is included in the supplementary material of this paper. We show the algorithm to solve (28) in Algorithm 1.

3) Update $\mathbf{U}_3$: We update $\mathbf{U}_3$ by optimizing:

$$\arg\min_{\mathbf{U}_3} \frac{1}{2}\|\mathbf{U}_3 - \mathbf{V}_3 + \frac{\mathbf{Y}_3^v}{\mu}\|_F^2 + \frac{1}{2}\|\nabla_{y_3}\mathbf{U}_3 - \mathbf{B}_3 + \frac{\mathbf{Y}_3^b}{\mu}\|_F^2 + \frac{1}{2}\|\mathbf{U}_3 - \mathbf{Z} + \frac{\mathbf{Y}^z}{\mu}\|_F^2 \quad (29)$$

We set the first-order derivative of (29) with respect to $\mathbf{U}_3$ to zero and update $\mathbf{U}_3$ by

$$\mathbf{U}_3 = (\nabla_{y_3}^T\nabla_{y_3} + 2\mathbf{I})^{-1} (\mathbf{V}_3 - \frac{\mathbf{Y}_3^v}{\mu} + \nabla_{y_3}^T(\mathbf{B}_3 - \frac{\mathbf{Y}_3^b}{\mu}) + \mathbf{Z} - \frac{\mathbf{Y}^z}{\mu}) \quad (30)$$

4) Update $\{\mathbf{B}_i\}_{i=1}^2$: The optimization function with respect to $\mathbf{B}_i$ $(i = 1, 2)$ is

$$\arg\min_{\mathbf{B}_i} \alpha_i\|\mathbf{B}_i\|_1 + \frac{\mu}{2}\|\nabla_{y_i}\mathbf{U}_i - \mathbf{B}_i + \frac{\mathbf{Y}_i^b}{\mu}\|_F^2 \quad (31)$$

We introduce the following soft-thresholding operator:

$$\mathcal{R}_\triangle(x) = \begin{cases} sgn(x)(|x| - \triangle) & |x| \geq \triangle \\ 0 & otherwise, \end{cases} \quad (32)$$

then we derive the solution of (31) by

$$\mathbf{B}_i = \mathcal{R}_{\frac{\alpha_i}{\mu}}(\nabla_{y_i}\mathbf{U}_i + \frac{\mathbf{Y}_i^b}{\mu}). \quad (33)$$

5) Update $\mathbf{B}_3$: The objective function with respect to $\mathbf{B}_3$ is

$$\arg\min_{\mathbf{B}_3} \alpha_3\|\mathbf{W}\mathbf{B}_3\|_{1,2} + \frac{\mu}{2}\|\nabla_{y_3}\mathbf{U}_3 - \mathbf{B}_3 + \frac{\mathbf{Y}_3^b}{\mu}\|_F^2 \quad (34)$$

As $\mathbf{W}$ is a diagonal matrix, the problem (34) can be solved in a row-wise manner. Denote by $\mathbf{b}_3^i$ the i-th row of $\mathbf{B}_3$ and $\mathbf{t}^i$ the i-th row of $\nabla_{y_3}\mathbf{U}_3 + \frac{\mathbf{Y}_3^b}{\mu}$, we obtain the optimization problem with respect to $\mathbf{b}_3^i$ as follows:

$$\arg\min_{\mathbf{b}_3^i} \alpha_3 W_{ii}\|\mathbf{b}_3^i\|_2 + \frac{\mu}{2}\|\mathbf{t}^i - \mathbf{b}_3^i\|_2^2. \quad (35)$$

The solution of $\mathbf{b}_3^i$ is derived by

$$\mathbf{b}_3^i = (1 - \alpha_3 W_{ii}/\mu/\|\mathbf{t}^i\|_2)_+\mathbf{t}^i, \quad (36)$$

where $(x)_+$ is defined by $(x)_+ = \max(x, 0)$.

6) Update $\mathbf{Z}$: We update $\mathbf{Z}$ by solving the optimization problem as follows:

$$\arg\min_{\mathbf{Z}} \beta\|\mathbf{Z}\|_* + \frac{\mu}{2}\|\mathbf{U}_3 - \mathbf{Z} + \frac{\mathbf{Y}^z}{\mu}\|_F^2 \quad (37)$$

To solve this problem, we introduce the following thresholding operator:

$$\mathcal{D}_\delta(\mathbf{X}) = \mathbf{U}\mathcal{D}_\delta(\mathbf{\Sigma})\mathbf{V}^T, \quad (38)$$

where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the singular value decomposition (SVD) of the matrix $\mathbf{X}$; $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors, respectively; $\mathbf{\Sigma}$ is a diagonal matrix with diagonal elements being the singular values σ_i and $\mathcal{D}_\delta(\mathbf{\Sigma}) = \text{diag}(\max((\sigma_i - \delta), 0))$. Then, we can update $\mathbf{Z}$ by

$$\mathbf{Z} = \mathcal{D}_{\frac{\beta}{\mu}}(\mathbf{U}_3 + \frac{\mathbf{Y}^z}{\mu}). \quad (39)$$

Algorithm 2 The proposed TSC model

Input: $\mathcal{X}$, α_i, β, λ
Initialize $\mathbf{U}_i$ and $\mathcal{G}$ by the Tucker decomposition (4), $\mathbf{V}_i = \mathbf{U}_i$, $\mathbf{Y}_i^v = \mathbf{0}$, $\mathbf{Y}_i^b = \mathbf{0}$, $\mathbf{Y}^z = \mathbf{0}$, $\mu > 0$, $\rho > 1$, $r = 0$, $\varepsilon = 1e-5$, $R = 100$
while $||\mathbf{U}_i - \mathbf{V}_i||_\infty > \varepsilon$ or $||\nabla_{y_i}\mathbf{U}_i - \mathbf{B}_i||_\infty > \varepsilon$ or $||\mathbf{U}_3 - \mathbf{Z}||_\infty > \varepsilon$ and $r < R$ **do**
  Update $\{\mathbf{V}_i\}_{i=1}^3$ via (27)
  Update $\{\mathbf{U}_i\}_{i=1}^2$ via Algorithm 1
  Update $\mathbf{U}_3$ by (30)
  Update $\{\mathbf{B}_i\}_{i=1}^2$ via (33)
  Update $\mathbf{B}_3$ by (36)
  Update $\mathbf{Z}$ by (39)
  Update $\mathcal{G}$ by (42)
  Update multipliers $\{\mathbf{Y}_i^v\}_{i=1}^3$, $\{\mathbf{Y}_i^b\}_{i=1}^3$ and $\mathbf{Y}^z$ by (45)
  $\mu = \rho\mu$, $r := r + 1$
end while
Output: Factor matrices $\mathbf{U}_i$ and core tensor $\mathcal{G}$

7) Update $\mathcal{G}$: The objective function with respect to $\mathcal{G}$ is formulated by

$$\arg\min_{\mathcal{G}} \frac{1}{2}\|\mathcal{X} - \mathcal{G} \times_1 \mathbf{V}_1 \times_2 \mathbf{V}_2 \times_3 \mathbf{V}_3\|_F^2 + \lambda\|\mathcal{G}\|_F^2 \quad (40)$$

According to (7), this problem can be equivalently reformulated to the following form:

$$\arg\min_{\text{vec}(\mathcal{G})} \frac{1}{2}\|\text{vec}(\mathcal{X}) - \mathbf{S}\text{vec}(\mathcal{G})\|_2^2 + \lambda\|\text{vec}(\mathcal{G}\|_2^2, \quad (41)$$

where $\mathbf{S} = \mathbf{V}_3 \otimes \mathbf{V}_2 \otimes \mathbf{V}_1$. By setting the derivative of (41) with respect to $\text{vec}(\mathcal{G})$ to zero, we obtain the close-form solution:

$$\mathcal{G} = \text{fold}((\mathbf{S}^T\mathbf{S} + 2\lambda\mathbf{I})^{-1}(\mathbf{S}^T\text{vec}(\mathcal{X}))), \quad (42)$$

where $\text{fold}(\cdot)$ is the inverse operation of $\text{vec}(\cdot)$, which folds a vector as a tensor. Directly computing $\mathcal{G}$ by (42) might be difficult due to the computationally expensive inverse calculation of the large matrix $\mathbf{S}^T\mathbf{S} + 2\lambda\mathbf{I}$. Here, we adopt a more efficient way [44] to update $\mathcal{G}$.

Let $\mathbf{Q}_i\mathbf{\Sigma}_i\mathbf{Q}_i^T$ be the SVD of $\mathbf{V}_i^T\mathbf{V}_i$, we can calculate $(\mathbf{S}^T\mathbf{S} + 2\lambda\mathbf{I})^{-1}$ by:

$$(\mathbf{S}^T\mathbf{S} + 2\lambda\mathbf{I})^{-1} = (\mathbf{Q}_3 \otimes \mathbf{Q}_2 \otimes \mathbf{Q}_1)(\mathbf{\Sigma}_3 \otimes \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1 + 2\lambda\mathbf{I})^{-1} (\mathbf{Q}_3^T \otimes \mathbf{Q}_2^T \otimes \mathbf{Q}_1^T). \quad (43)$$

As $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$ are diagonal matrices, $\mathbf{\Sigma}_3 \otimes \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ is diagonal as well. Thus, $(\mathbf{\Sigma}_3 \otimes \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1 + 2\lambda\mathbf{I})^{-1}$ can be computed much efficiently. To compute $\mathbf{S}^T\text{vec}(\mathcal{X})$ in (42), we use the following equation:

$$\mathbf{S}^T\text{vec}(\mathcal{X}) = \text{vec}(\mathcal{X} \times_1 \mathbf{V}_1^T \times_2 \mathbf{V}_2^T \times_3 \mathbf{V}_3^T), \quad (44)$$

which can be seen as a vector that is flatten from a tensor obtained by a series of tensor products, i.e., $\mathcal{X} \times_1 \mathbf{V}_1^T \times_2 \mathbf{V}_2^T \times_3 \mathbf{V}_3^T$. This avoids the computationally expensive calculation of $\mathbf{S} = \mathbf{V}_3 \otimes \mathbf{V}_2 \otimes \mathbf{V}_1$.

8) Update others: We update other variables by:

$$\begin{cases} \mathbf{Y}_i^v &= \mathbf{Y}_i^v + \mu(\mathbf{U}_i - \mathbf{V}_i) \\ \mathbf{Y}_i^b &= \mathbf{Y}_i^b + \mu(\nabla_{y_i}\mathbf{U}_i - \mathbf{B}_i) \\ \mathbf{Y}^z &= \mathbf{Y}^z + \mu(\mathbf{U}_3 - \mathbf{Z}) \end{cases} \tag{45}$$

We summarize the aforementioned optimization algorithm in Algorithm 2. According to (20), columns of $\mathbf{U}_3^T$ can be viewed as the latent features of the projected bands. Once $\mathbf{U}_3$ is obtained by Algorithm 2, we build a KNN graph based on $\mathbf{U}_3^T$ and apply the KNN graph into spectral clustering, which derives final clustering results of bands. In each cluster, the band, which is closest to the corresponding centroid, is selected.

D. Computational Complexity

We here analyse the time complexity of Algorithm 2. The major cost comes from updating $\{\mathbf{V}_i\}_{i=1}^3$, which have time complexity of $\mathcal{O}(R_1I_2I_3(I_1+I_3))$ for updating $\mathbf{V}_1$, $\mathcal{O}(I_1R_2I_3(I_2+I_3))$ for updating $\mathbf{V}_2$ and $\mathcal{O}(I_1I_2I_3^2)$ for updating $\mathbf{V}_3$. The updates of $\{\mathbf{U}_i\}_{i=1}^3$ have time complexity of $\mathcal{O}(I_1R_1^2)$, $\mathcal{O}(I_2R_2^2)$ and $\mathcal{O}(I_3^3)$, respectively. The complexity of updating $\mathbf{B}_1$ and $\mathbf{B}_2$ is neglectable, and the updates of $\mathbf{B}_3$, $\mathbf{Z}$ and $\mathcal{G}$ have time complexity of $\mathcal{O}(I_3R_3)$, $\mathcal{O}(I_3^3)$ and $\mathcal{O}(R_1I_1I_2I_3)$, respectively. Considering the fact that $I_1, I_2, I_3 \gg R_1$ and R_2, the overall time complexity of Algorithm 2 is $\mathcal{O}(RI_1I_2I_3^2)$, where R is the number of iterations.

E. Convergence Analysis

The convergence property of ADMM algorithm has been theoretically proved in [67] for solving problems with a convex and nonsmooth objective function with two blocks of variables. However, it is difficult to guarantee the convergence of ADMM for solving nonconvex multiblock problems. We show a week convergence property of our algorithm by proving that the solution obtained by Algorithm 2 converges to a Karush-Kuhn-Tucker (KKT) point under some mild conditions. The convergence property is stated in Theorem 1. A detailed proof of Theorem 1 is provided in the supplementary material.

Theorem 1. *Let* $\{\Gamma^r = (\{\mathbf{U}_i^r\}_{i=1}^3, \{\mathbf{V}_i^r\}_{i=1}^3, \{\mathbf{B}_i^r\}_{i=1}^3, \mathbf{Z}^r, \mathcal{G}^r, \{\mathbf{Y}_i^{v^r}\}_{i=1}^3), \{\mathbf{Y}_i^{b^r}\}_{i=1}^3, \mathbf{Y}^{z^r})\}_{r=1}^\infty$ *be the sequence that is derived from Algorithm 2. If* $\lim_{r\to\infty}\mu^r(\mathbf{U}_i^{r+1} - \mathbf{V}_i^{r+1}) = \mathbf{0}$ *and* $\lim_{r\to\infty}\mu^r(\mathbf{U}_i^{r+1} - \mathbf{U}_i^r) = \mathbf{0}$ *($i = 1, 2, 3$), the sequence* $\{\Gamma^r\}_{r=1}^\infty$ *is bounded, and its accumulation point* $\Gamma^* = (\{\mathbf{U}_i^*\}_{i=1}^3, \{\mathbf{V}_i^*\}_{i=1}^3, \{\mathbf{B}_i^*\}_{i=1}^3, \mathbf{Z}^*, \mathcal{G}^*, \{\mathbf{Y}_i^{v^*}\}_{i=1}^3), \{\mathbf{Y}_i^{b^*}\}_{i=1}^3, \mathbf{Y}^{z^*})$ *satisfies the KKT conditions. The sequence of* $\{\Gamma^r\}_{r=1}^\infty$ *converges to a KKT point.*

IV. Experiments

By selecting representative bands, band selection algorithms are able to effectively reduce the spectral dimension of HSI. The performance of band selection is often validated in the pixel-wise classification task of HSI [6, 16, 33, 34], where the selected subset of bands are used as features of pixels for classifier to discriminate between different land covers. The same training data and test data are used when evaluating different band selection algorithms. In the experiments, we employ the classifiers support vector machine (SVM) and KNN for performance analysis of band selection. The overall accuracy (OA), average accuracy (AA) and Kappa coefficient (κ) are utilized as the quantitative evaluation metrics for the classifiers. The classifiers SVM with radial basis function (RBF) kernel and KNN are implemented by using the optimization toolbox in MATLAB 2015b. The number of neighbours in the classifier KNN is set to three as in [6, 68].

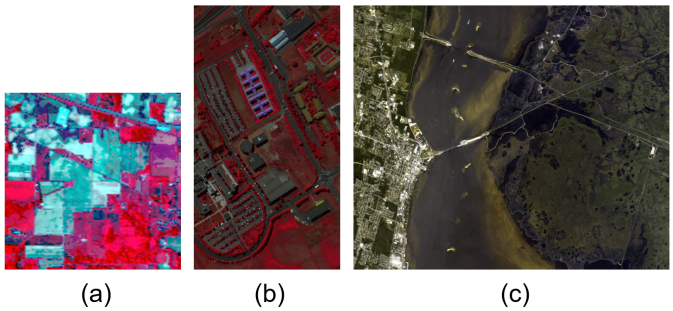

Fig. 3. The false color images of: (a) *IndianP*, (b) *PaviaU* and (c) *KSC*.

A. Data Sets

1) Indian Pines (IndianP): The data set *IndianP* was captured by the Airborne/Visible Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines region in North-western Indiana on June 12, 1992. The spatial size of IndianP is 145×145 and there are 220 spectral bands. The false color image is shown in Fig. 3 (a). In the experiment, we remove 20 spectral bands in 104-108, 150-163 and 200 due to water absorption. The remaining 200 bands are utilized for test data. *IndianP* contains 16 classes in total.

2) Pavia University (PaviaU): The second data set was acquired by the Reflected Optics System Imaging Spectrometer (ROSIS) during a flight campaign over Pavia, Northern Italy. The image size is $512 \times 217 \times 103$ and there are nine classes. The false color image is shown in Fig. 3 (b).

3) Kennedy Space Center (KSC): *KSC* was collected by the AVIRIS sensor over the Kennedy Space Center, Florida, which contains 13 classes. The data size is $512 \times 614 \times 176$ after removing water absorption and low SNR bands. We show the false color image of *KSC* in Fig. 3 (c). All three data sets are accessible by https://rslab.ut.ac.ir/data.

B. Compared Methods

1) Maximum-variance principal component analysis (MVPCA) [21]: a ranking based band selection method in the criterion of band variation.
2) E-FDPC [26]: a ranking based method which makes use of clustering algorithm.
3) Uniform band selection (UBS) [22]: a simple but efficient strategy to select bands uniformly.
4) Optimal clustering framework (OCF) [35]: a clustering based method which assumes that bands in the same cluster have contiguous wavelengths.

TABLE I
CLASSIFICATION RESULTS WHEN TEN BANDS ARE SELECTED

Datasets	Evaluations	MVPCA [21]	E-FDPC [26]	UBS [22]	OCF [35]	LRR [69]	ISSC [40]	L2-LRSC [36]	TLRR [70]	ETLMSC [59]	TSC
IndianP	SVM (OA)	46.44	56.02	61.88	69.30	70.95	70.62	71.05	68.36	71.01	**71.44**
	SVM (AA)	41.26	50.51	55.02	66.17	62.07	65.43	67.14	62.07	67.09	**67.27**
	SVM (κ)	0.3852	0.4974	0.5633	0.6488	0.6679	0.6640	0.6689	0.6384	0.6685	**0.6734**
	KNN (OA)	45.07	55.83	58.13	62.87	65.34	64.21	65.28	63.64	65.84	**65.97**
	KNN (AA)	37.94	45.39	50.17	56.51	58.55	56.30	57.96	55.00	59.33	**59.54**
	KNN (κ)	0.3704	0.4927	0.5187	0.5744	0.6028	0.5894	0.6022	0.5831	0.6084	**0.6101**
PaviaU	SVM (OA)	48.36	78.03	80.48	79.77	81.18	81.91	82.79	82.81	83.24	**83.90**
	SVM (AA)	38.21	73.62	76.74	76.34	78.42	78.66	79.82	79.91	79.97	**80.83**
	SVM (κ)	0.2958	0.7052	0.7385	0.7284	0.7477	0.7574	0.7696	0.7697	0.7753	**0.7846**
	KNN (OA)	45.96	77.23	78.21	78.94	79.91	80.83	81.44	81.06	80.65	**82.28**
	KNN (AA)	32.53	69.54	72.40	74.25	75.48	76.11	77.18	76.75	76.19	**77.34**
	KNN (κ)	0.2440	0.6907	0.7033	0.7143	0.7280	0.7400	0.7483	0.7435	0.7378	**0.7599**
KSC	SVM (OA)	46.85	84.16	81.67	84.28	82.84	84.97	83.97	83.35	84.02	**85.91**
	SVM (AA)	33.49	76.54	73.06	76.20	74.86	77.15	76.03	75.42	76.30	**78.49**
	SVM (κ)	0.4023	0.8234	0.7955	0.8248	0.8087	0.8324	0.8213	0.8143	0.8218	**0.8428**
	KNN (OA)	40.58	82.62	80.79	83.95	81.96	83.66	83.90	82.76	83.37	**84.77**
	KNN (AA)	30.03	74.14	71.78	75.39	73.67	75.04	75.93	74.77	75.43	**76.81**
	KNN (κ)	0.3354	0.8061	0.7855	0.8210	0.7987	0.8176	0.8204	0.8076	0.8145	**0.8301**

TABLE II
SELECTED TEN BANDS OF DIFFERENT METHODS

Methods	*IndianP*	*PaviaU*	*KSC*
MVPCA [21]	23 24 25 26 27 28 29 30 32 42	87 88 89 90 91 92 93 94 95 96	1 133 160 168 170 171 173 174 175 176
E-FDPC [26]	118 125 128 146 157 163 173 181 189 192	19 29 33 41 48 53 56 61 88 92	6 10 22 42 50 55 58 67 69 74
UBS [22]	1 23 45 67 89 111 133 155 177 200	1 12 23 35 46 57 69 80 91 103	1 20 39 59 78 98 117 137 156 176
OCF [35]	8 28 43 50 67 107 118 128 141 173	4 15 19 33 36 48 53 61 80 88	2 10 22 29 34 39 42 50 69 77
LRR [69]	10 25 47 61 72 99 132 143 161 184	5 13 26 37 44 50 59 72 80 96	3 4 13 35 56 65 80 116 130 152
ISSC [40]	31 41 74 96 115 125 137 163 186 199	6 13 22 31 41 49 62 71 79 99	8 15 22 42 52 73 83 134 136 149
L2-LRSC [36]	14 27 35 51 62 88 98 132 143 174	7 19 29 42 58 73 80 84 93 99	15 27 33 35 36 39 44 47 72 121
TLRR [70]	4 14 31 47 83 93 133 143 155 179	5 17 29 40 52 64 73 87 90 99	4 10 24 33 57 77 94 119 147 158
ETLMSC [59]	10 25 43 53 71 89 99 132 152 174	4 12 20 33 46 58 73 76 88 98	2 16 34 45 58 95 97 114 152 155
TSC	15 25 51 72 98 106 123 137 157 167	12 21 29 38 48 58 66 83 92 101	1 7 17 26 38 51 59 77 97 153

5) Low-rank representation (LRR) [69]: a matrix-based subspace clustering method with the low-rank constraint.
6) Improved sparse subspace clustering (ISSC) [40]: a computationally efficient matrix-based subspace clustering method with a collaborative representation.
7) L2-LRSC [36]: a more effective matrix-based subspace clustering method with a local Laplacian constraint.
8) Tensor low-rank representation (TLRR) [70]: a recent tensor-based clustering method in computer vision task using a tensor nuclear norm in the discrete frequency domain.
9) Essential tensor learning for multi-view spectral clustering (ETLMSC) [59]: a tensor-based model for multi-view clustering where a tensor singular value decomposition (t-SVD)-based tensor nuclear norm is imposed to capture principle information from multiple views. To generate "multiple views" of HSI for band selection, we divide evenly a HSI $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ into two portions with each one of size $\lfloor \frac{I_1}{2} \rfloor \times I_2 \times I_3$ and consider each portion as a "view", where $\lfloor x \rfloor$ takes the maximum integer that is not larger than x.

C. Experimental Results

We report the classification performances of different methods in Table I when ten bands are selected. In the following experiments, we select five percent of samples in each class to train the classifiers, and the rest are used for testing. We conduct all the experiments ten times and report the results in average. The parameters of all methods are tuned to yield the optimal results in terms of OA. In our TSC method, in order to reduce the burden of searching optimal parameters, we simply set $\alpha_1 = \alpha_2 = \alpha$. We set empirically $\alpha = 10, \alpha_3 = 0.25, \beta = 0.1$ for *IndianP*, $\alpha = 0.01, \alpha_3 = 0.75, \beta = 0.01$ for *PaviaU* and $\alpha = 0.01, \alpha_3 = 0.5, \beta = 1$ for *KSC*. We set $\lambda = 0.01, R_1 = R_2 = 20$ in all the data sets. The parameter study is provided in the subsequent section.

The results in Table I show that TSC yields the best classification performance in both classifiers, which demonstrates that our selected bands offer more discriminative features than other methods. The ranking-based method MVPCA yields poor results on three data sets due to the highly redundant information in the selected bands. Another ranking-based method E-FDPC performs much better than MVPCA because of the decorrelation of bands by using clustering algorithm. It is surprised that the simplest method UBS outperforms MVPCA and E-FDPC on the data sets *IndianP* and *PaviaU*, which confirms the effectiveness of uniform band selection. Compared to E-FDPC, OCF achieves improved performance in terms of OA and κ on the three data sets, which benefits from the assumption that the bands belonging to the same cluster are adjacent. It is noticed that most of representation-based methods, including LRR, ISSC, L2-LRSC, TLRR, ETLMSC and TSC, show often better classification performance than others, demonstrating their superiority in capturing complex data structure in low-dimensional subspaces. L2-LRSC yields improved performance on three data sets compared to LRR,

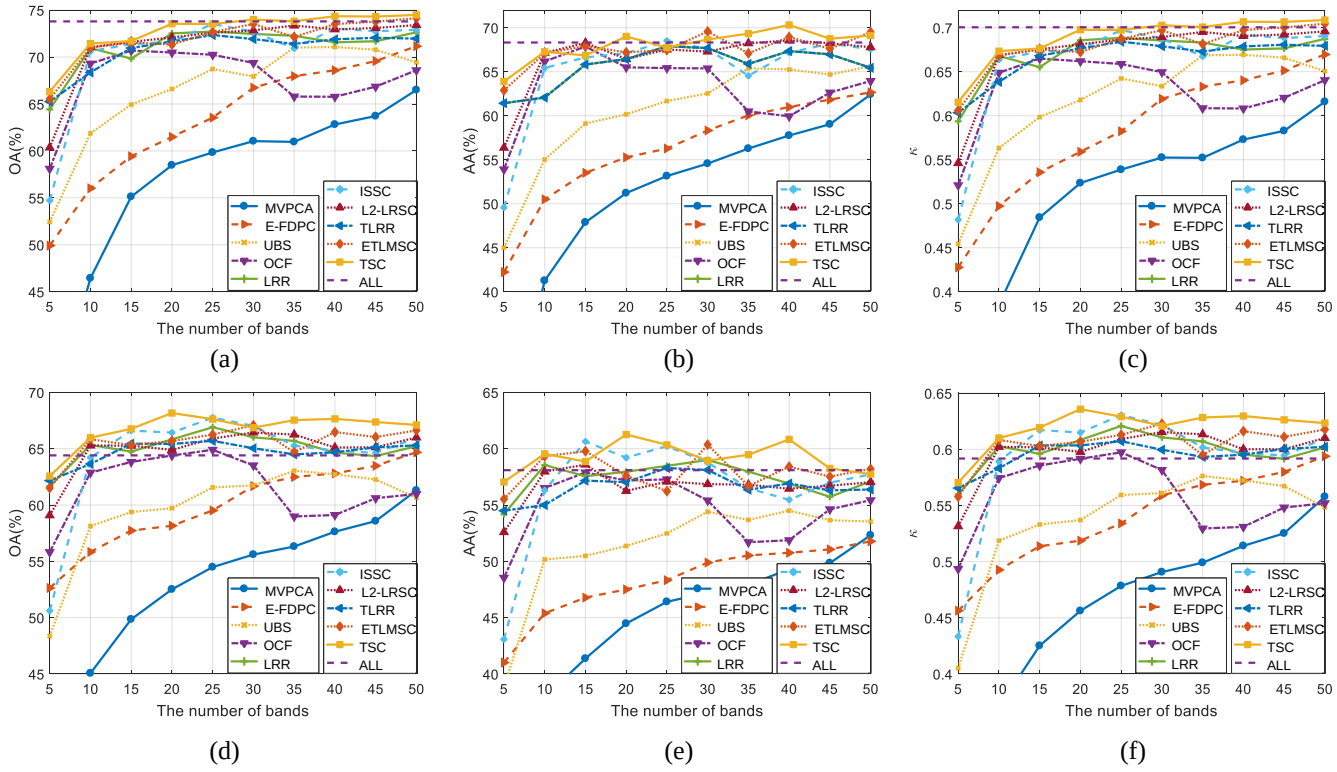

Fig. 4. The influence of the number of selected bands on the classification performance in classifiers SVM and KNN on the dataset *IndianP*. (a) OA in SVM, (b) AA in SVM, (c) κ in SVM, (d) OA in KNN, (e) AA in KNN and (f) κ in KNN.

which can be mainly attributed to the exploitation of local Laplacian constraint. The tensor-based method TLRR shows improved performance compared to the matrix-based method LRR on the data sets *PaviaU* and *KSC*, which validates the effectiveness of high-order model in dealing with tensor data. Although TLRR utilizes an elaborate nuclear norm in the discrete frequency domain, which avoids destroying the multi-mode structure of HSI, the important spatial information of HSI and local correlation of bands are not taken into account, resulting in inferior performance to TSC on the three data sets as shown in Table I. ETLMSC incorporates information from "multi-view" of HSI to learn the cluster structure of bands and yields better performance than TLRR but it performs worse than TSC. The selected ten bands of different methods are shown in Table II.

D. Performance with Varying Number of Bands

We show the influence of the number of selected bands, n, on the classification performance of SVM and KNN in Figs. 4-6. We here include the classification performance when using all the bands, which is referred as "ALL". Generally, Almost all the methods obtain improved classification performance with increasing number of bands. MVPCA does not perform well when only a few bands ($n < 30$) are selected. But it shows sharp performance improvements in the ranges of $n \in [5, 20]$ in *IndianP*, $n \in [5, 30]$ in *PaviaU* and $n \in [5, 50]$ in *KSC*. When $n = 50$, MVPCA achieves comparable or even better performance than some of the reference methods such as E-FDPC in Fig. 5. It is observed that the proposed TSC method outperforms all the compared methods ("ALL" is excluded) in most cases on the three data sets. Compared with the results using all the bands, most of representation-based methods show comparable or even better classification performances with only a few bands, especially on the data set *KSC*. This indicates the high redundancy of spectral bands in HSI and confirms the significance of band selection in the classification task. UBS shows poor classification performance when $n = 5$, but its performance gets improved significantly in the range of $10 \leq n \leq 25$. The saturated accuracy of UBS in Fig. 6 shows that the simplest method UBS is able to yield a better classification result than that using all the bands. The performance of OCF is less stable on the data set *IndianP* in Fig. 4, where the significant fluctuations of performance around $n = 35$ can be observed.

The representation-based methods show a similar evolution curve as shown in Figs. 4-6 and often perform better than the others. For the sake of discussion, we report the average results of Figs. 4-6 in Table III, where each value represents the mean of the classification results with varying n. Table III shows that TSC yields mostly the best classification performance except for the classifier KNN on the data set *KSC*. However, the accuracy and κ of KNN for our method in the *KSC* are comparable to the best result yielded by OCF. In addition, Fig. 6 (c) and (d) show that our method performs much better than OCF when only a few bands are utilized such as

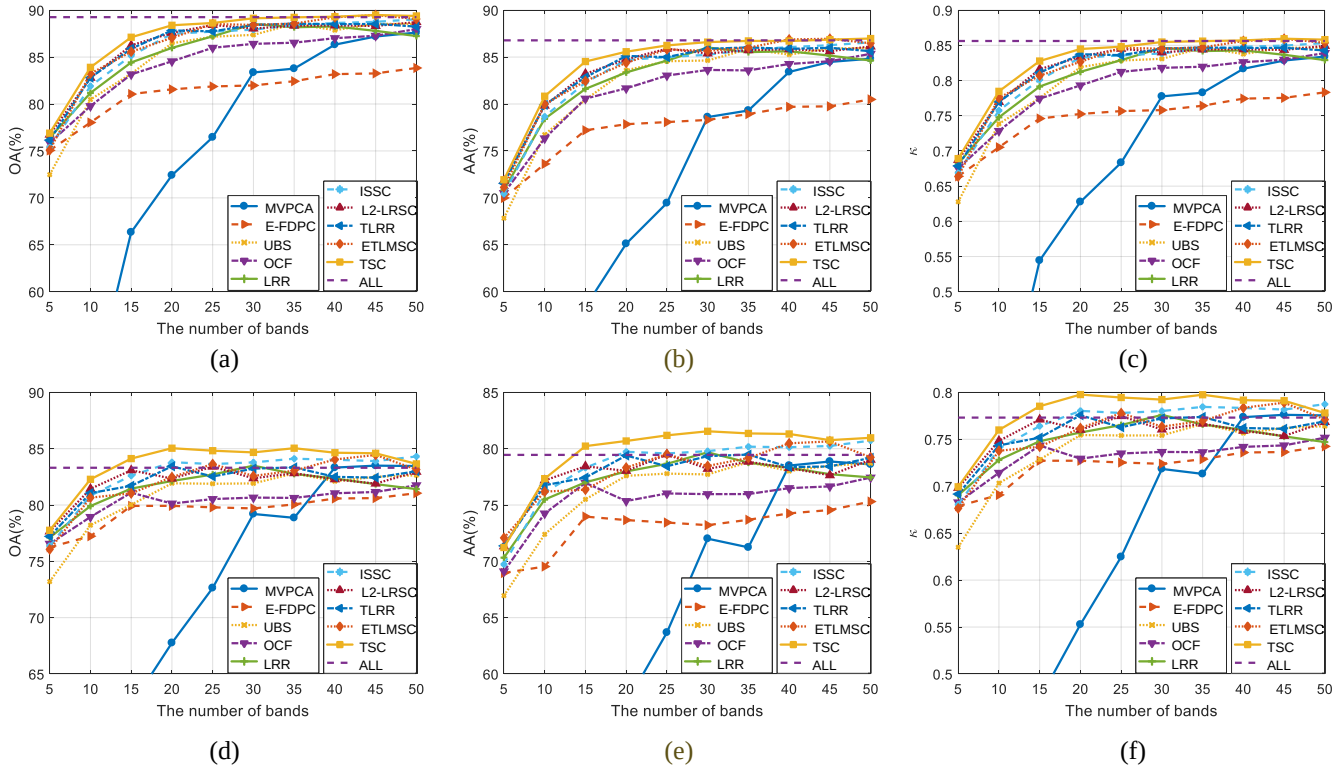

Fig. 5. The influence of the number of selected bands on the classification performance in classifiers SVM and KNN on the dataset *PaviaU*. (a) OA in SVM, (b) AA in SVM, (c) κ in SVM, (d) OA in KNN, (e) AA in KNN and (f) κ in KNN.

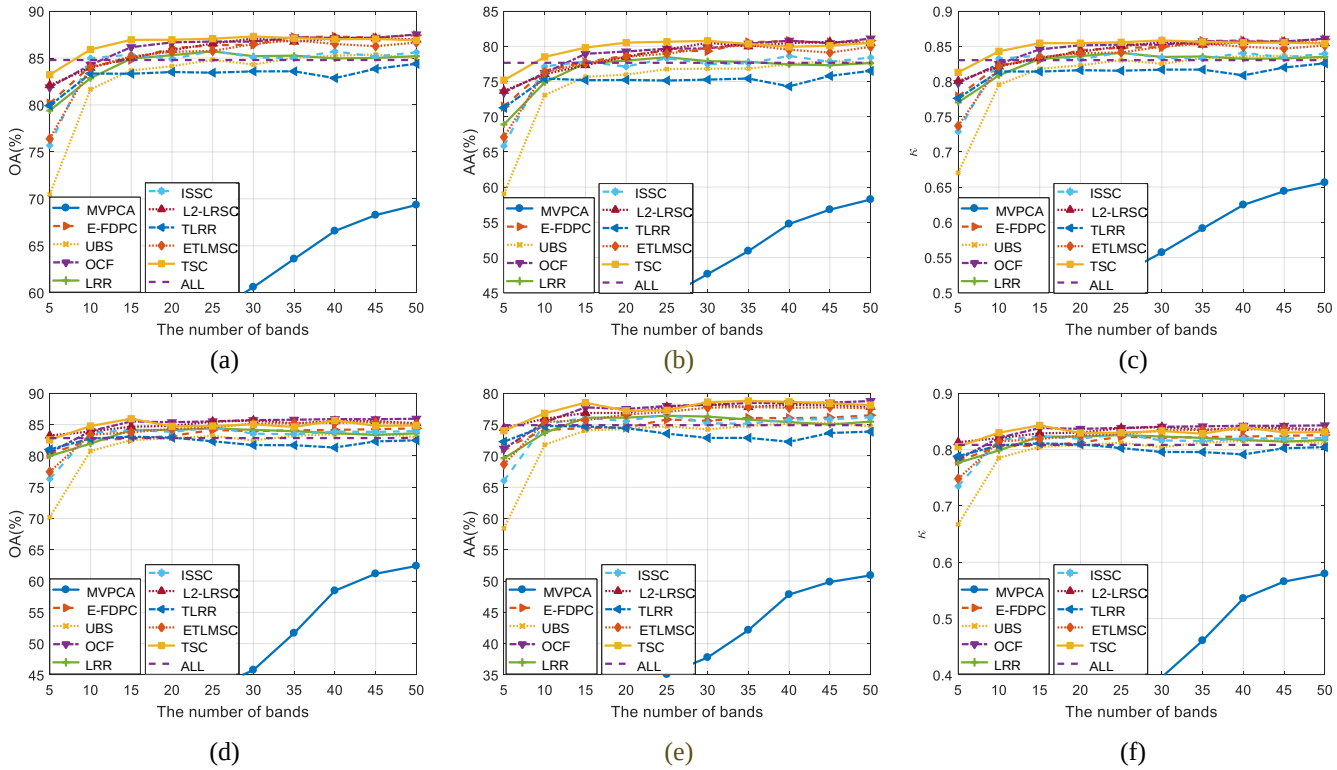

Fig. 6. The influence of the number of selected bands on the classification performance in classifiers SVM and KNN on the dataset *KSC*. (a) OA in SVM, (b) AA in SVM, (c) κ in SVM, (d) OA in KNN, (e) AA in KNN and (f) κ in KNN.

TABLE III
THE AVERAGE CLASSIFICATION RESULTS WITH DIFFERENT BAND SUBSETS

Datasets	Evaluations	MVPCA [21]	E-FDPC [26]	UBS [22]	OCF [35]	LRR [69]	ISSC [40]	L2-LRSC [36]	TLRR [70]	ETLMSC [59]	TSC
IndianP	SVM (OA)	56.60	63.44	66.48	67.52	71.11	70.42	71.34	70.78	71.92	**72.75**
	SVM (AA)	50.56	56.16	60.45	63.12	65.70	65.23	66.58	65.70	67.53	**68.09**
	SVM (κ)	0.5018	0.5816	0.6164	0.6283	0.6699	0.6618	0.6722	0.6661	0.6790	**0.6886**
	KNN (OA)	52.42	59.90	59.76	61.50	65.04	64.33	64.93	64.70	65.55	**66.76**
	KNN (AA)	43.96	48.30	51.31	54.64	57.34	56.49	56.63	56.60	57.96	**59.22**
	KNN (κ)	0.4547	0.5389	0.5379	0.5585	0.5994	0.5910	0.5981	0.5954	0.6051	**0.6193**
PaviaU	SVM (OA)	72.88	81.23	85.06	84.45	85.52	86.03	86.35	86.27	86.42	**87.14**
	SVM (AA)	66.71	77.39	82.08	81.35	82.58	83.10	83.42	83.38	83.57	**84.29**
	SVM (κ)	0.6337	0.7479	0.7997	0.7916	0.8060	0.8130	0.8173	0.8160	0.8180	**0.8279**
	KNN (OA)	69.48	79.51	80.72	80.25	81.52	82.74	82.02	82.04	82.11	**83.66**
	KNN (AA)	60.53	73.06	76.19	75.42	77.14	78.45	77.64	77.82	78.05	**79.66**
	KNN (κ)	0.5792	0.7217	0.7378	0.7315	0.7491	0.7660	0.7560	0.7564	0.7573	**0.7787**
KSC	SVM (OA)	57.09	85.71	82.97	86.15	84.38	84.32	85.85	83.18	84.98	**86.54**
	SVM (AA)	44.69	78.58	74.71	78.99	76.58	76.61	78.71	74.95	77.69	**79.63**
	SVM (κ)	0.5177	0.8407	0.8100	0.8456	0.8258	0.8251	0.8423	0.8124	0.8325	**0.8499**
	KNN (OA)	47.64	83.38	81.52	**85.00**	83.33	83.12	84.93	82.17	83.96	84.78
	KNN (AA)	38.12	75.05	72.75	77.20	74.99	74.75	77.21	73.53	76.20	**77.64**
	KNN (κ)	0.4154	0.8146	0.7937	**0.8327**	0.8140	0.8116	0.8319	0.8009	0.8210	0.8301

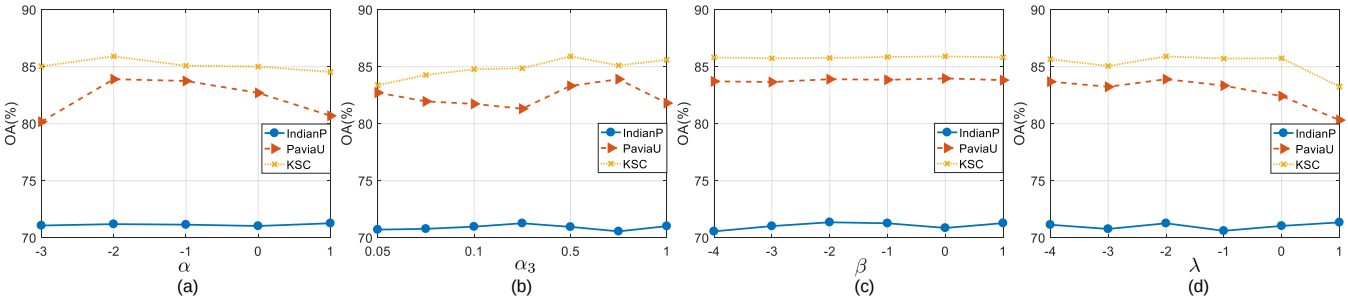

Fig. 7. The classification accuracies of SVM with respect to parameters (a) α, (b) α_3, (c) β and (d) λ. The x-axes of (a), (c) and (d) are in log_{10}.

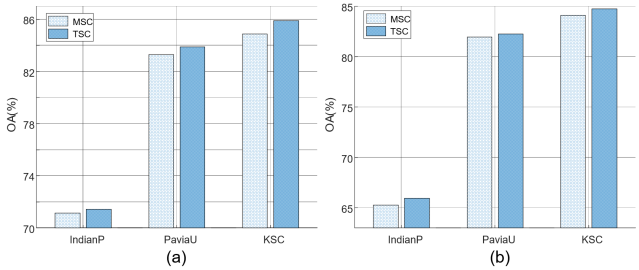

Fig. 8. The comparison between MSC and TSC, where MSC is the reduced version of TSC. (a) OA in SVM and (b) OA in KNN.

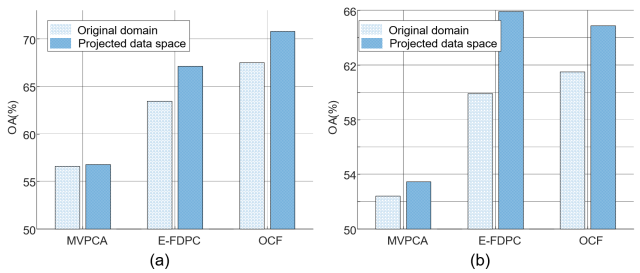

Fig. 9. The influence of projected data on band selection. (a) OA in SVM and (b) OA in KNN.

$n = 5$ and $n = 10$. L2-LRSC outperforms other matrix-based methods LRR and ISSC in most cases, which demonstrates the importance of local band correlation. Compared to the matrix-based methods, our tensor-based approach TSC often achieves improved performance in both classifiers, showing the efficiency of multi-way representation for HSI. In comparison with other tensor-based methods TLRR and ETLMSC, we obtain consistently better performance on all three data sets.

E. Discussion and Analysis

1) The analysis of parameters: We show in Fig. 7 the influence of parameters α, α_3, β and λ on the band selection performance on the three data sets. Basically, the parameters α and α_3 are associated with the local constraints and β is related to the global low-rank constraint. Ten bands are selected by TSC and then fed into the classifier SVM. Fig. 7 shows that the performance of TSC is rather stable in a certain range with respect to each parameter. Specifically, when $\alpha \in [0.01, 0.1]$, $\alpha_3 \in [0.25, 0.75]$, $\beta \in [0.01, 1]$ and $\lambda \in [0.01, 0.1]$, our method yields relatively stable results and is able to achieve near optimal performance on the three data sets. It is observed that in Fig. 7 (a) the results on the data set *KSC* outperform "ALL" (see Fig. 6 (a)) in most cases. Compared to β, the performance of TSC is more sensitive to α and α_3, especially on the data set *PaviaU*, demonstrating the importance of local constraints. In the data set *IndianP*, our method yields stable performance with respect to α and it is noted that the accuracies of our model in *IndianP* with

TABLE IV
TIME COMPLEXITY AND RUNNING TIME (IN SECONDS) OF DIFFERENT BAND SELECTION METHODS

Methods		MVPCA	E-FDPC	OCF	LRR	ISSC	L2-LRSC	TLRR	ETLMSC	TSC
Time complexity		$\mathcal{O}(I_1I_2I_3)$	$\mathcal{O}(I_1I_2I_3^2)$	$\mathcal{O}(I_1I_2I_3^2)$	$\mathcal{O}(RrI_1I_2I_3)$	$\mathcal{O}(I_1I_2I_3^2)$	$\mathcal{O}(RI_3^3+I_1I_2I_3^2)$	$\mathcal{O}(Rr_{\mathcal{X}}I_1(I_2I_3+(I_2+I_3)log(I_1)))$	$\mathcal{O}(I_3^3+2RI_3^2(2+log(I_3)))$	$\mathcal{O}(RI_1I_2I_3^2)$
Running time	*IndianP*	0.01	0.07	0.99	18.68	0.07	1.08	223.60	2.19	81.15
	PaviaU	0.04	0.23	0.65	80.23	0.06	0.79	441.46	1.37	130.52
	KSC	0.14	0.73	1.85	230.68	0.90	1.49	2041.90	2.95	320.88

$\alpha \in [10^{-3}, 10]$ are higher than that of all compared methods in Table I.

2) Comparison with the corresponding matrix-based model: We here compare our TSC model with its corresponding matrix-based model by setting $\mathbf{U}_1 = \mathbf{I}$, $\mathbf{U}_2 = \mathbf{I}$ and $\mathcal{G} = \mathcal{X}$ in (21). We refer to the matrix-based model as MSC, which is a special case of TSC. The number of bands is set to ten. The results shown in Fig. 8 reveal that TSC yields better classification performance in the classifiers SVM and KNN on the three data sets, which demonstrates that tensor-based model can deal with high-order data more effectively than matrix-based model. Moreover, it is observed that compared to the results in Table I, the reduced model MSC yields mostly better results in terms of OA and κ.

3) Band selection in the projected data space: An interesting study here will be whether the existing band selection methods in the projected data space, which is derived by the projections of factor matrices $\mathbf{U}_1^T$ and $\mathbf{U}_2^T$, can obtain improved results compared to band selection in the original domain. We conduct experiments on *IndianP* using three methods MVPCA, E-FDPC and OCF for this analysis as they are parameter-free approaches except the parameter n, i.e., the number of selected bands. Specifically, we vary n in the same way as in Fig. 4 and feed the original data $\mathcal{X}$ and the projected data $\mathcal{X}_p = \mathcal{X} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T$ respectively into the band selection methods, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are obtained by Algorithm 2. The averaged classification results of SVM and KNN with varying n are shown in Fig. 9. The results show that the selected bands using MVPCA, E-FDPC and OCF in the projected data space yield consistently better classification performance in SVM and KNN than the results in the original domain. In the classifier KNN, E-FCPC achieves a significant OA improvement of 6%. This confirms the superiority of band selection in the projected data space. Moreover, as $\mathcal{X}_p$ often has a much smaller data size than $\mathcal{X}$, it is more efficient to extract representative bands in the projected data space.

4) Computational complexity and running time: The time complexity and running time of different band selection methods are shown in Table IV. All methods are implemented in MATLAB on a laptop with Intel core-i7 6700HQ CPU with 16GB of RAM. We set the number of selected bands to ten for all the methods when running time is reported. In Table IV, I_1 and I_2 are spatial dimensions of HSI; I_3 represents spectral dimension of HSI; R is the number of iterations in optimization algorithms; r is the rank of $\mathbf{X}_{(3)}$ and $r_{\mathcal{X}}$ is the tensor tubal rank of $\mathcal{X}$ (see Definition 3 of [70]). The complexity of UBS is not shown in Table IV but it is the fastest method with complexity $\mathcal{O}(1)$ because UBS selects bands uniformly without using any learning algorithms. The results in Table IV show that MVPCA obtains a lower time complexity of $\mathcal{O}(I_1I_2I_3)$ than others. E-FDPC, OCF and ISSC yield the same time complexity of $\mathcal{O}(I_1I_2I_3^2)$. For the time complexities of other methods, we observe that L2-LRSC<ETLMSC<LRR<TSC<TLRR. Compared with matrix-based models ISSC, LRR and L2-LRSC, tensor-based models TSC and TLRR have higher time complexities. Among tensor-based models, the time complexity of ETLMSC is lowest due to the adopted efficient strategy of tensor rotation. TSC yields a time complexity of $\mathcal{O}(RI_1I_2I_3^2)$. In general, the running time of different methods is consistent with the time complexity on three data sets. MVPCA is the most efficient algorithm and E-FDPC, OCF and ISSC yield comparable running time. Among representation-based models, ISSC is the fastest approach. Tensor-based models TLRR and TSC use more running time than matrix-based models ISSC, LRR and L2-LRSC. Our TSC algorithm is much faster than TLRR but slower than ETLMSC.

V. CONCLUSION

In this paper, we propose a tensor-based subspace clustering model for hyperspectral band selection. To the best of our knowledge, this is the first attempt of using a tensor-based model in the task of band selection. Compared to the existing matrix-based models, our TSC model treats the 3-D HSI as a whole by the product of three factor matrices and a core tensor without using any data flattening operator, preserving thus effectively the spatial structure of HSI. This renders TSC more flexible than matrix-based approaches. Under certain conditions, our TSC model reduces to the traditional matrix-based models. By taking into account the prior information of HSI in terms of the local and global properties of data along three dimensions, we propose well-motivated heterogeneous constraints on the factor matrices, which improves the encoding of multi-mode correlations of data. Our TSC model naturally learns the cluster structure of bands in a low-dimensional latent feature space derived by the factor matrices in the spatial dimension, which is more efficient than the learning in the original domain. We propose an efficient algorithm based on ADMM to solve the resulting optimization problem. Experimental results on benchmark data sets show that our method outperforms the state-of-the-art in this domain.

REFERENCES

[1] B. Lu, P. D. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sens.*, vol. 12, no. 16, 2020.

[2] Y.-J. Deng, H.-C. Li, K. Fu, Q. Du, and W. J. Emery, "Tensor low-rank discriminant embedding for hyperspectral image dimensionality reduction," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7183–7194, 2018.

[3] H. Huang and M. Yang, "Dimensionality reduction of hyperspectral images with sparse discriminant embedding," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 5160–5169, 2015.

[4] L. Zhang, L. Zhang, B. Du, J. You, and D. Tao, "Hyperspectral image unsupervised classification by robust manifold matrix factorization," *Inf. Sci.*, vol. 485, pp. 154–169, 2019.

[5] X. Geng, K. Sun, L. Ji, and Y. Zhao, "A fast volume-gradient-based band selection method for hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7111–7119, 2014.

[6] A. MartÍnez-UsÓMartinez-Uso, F. Pla, J. M. Sotoca, and P. García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158–4171, 2007.

[7] H. Yang, Q. Du, H. Su, and Y. Sheng, "An efficient method for supervised hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 138–142, 2010.

[8] X. Cao, T. Xiong, and L. Jiao, "Supervised band selection using local spatial information for hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 329–333, 2016.

[9] M. Habermann, V. Fremont, and E. H. Shiguemori, "Supervised band selection in hyperspectral images using single-layer neural networks," *Int. J. Remote Sens.*, vol. 40, no. 10, pp. 3900–3926, 2019.

[10] J. Bai, S. Xiang, and C. Pan, "Classification oriented semi-supervised band selection for hyperspectral images," in *Proc. Int. Conf. Pattern Recognit.*, 2012, pp. 1888–1891.

[11] X. Bai, Z. Guo, Y. Wang, Z. Zhang, and J. Zhou, "Semisupervised hyperspectral band selection via spectral–spatial hypergraph model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2774–2783, 2015.

[12] A. Sellami and I. R. Farah, "A spatial hypergraph based semi-supervised band selection method for hyperspectral imagery semantic interpretation," *Int. J. Comput. Inf. Eng.*, vol. 10, no. 10, pp. 1839–1846, 2016.

[13] G. Zhu, Y. Huang, J. Lei, Z. Bi, and F. Xu, "Unsupervised hyperspectral band selection by dominant set extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 227–239, 2015.

[14] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1279–1289, 2016.

[15] G. Zhu, Y. Huang, S. Li, J. Tang, and D. Liang, "Hyperspectral band selection via rank minimization," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2320–2324, 2017.

[16] Q. Wang, F. Zhang, and X. Li, "Hyperspectral band selection via optimal neighborhood reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8465–8476, 2020.

[17] Q. Wang, Q. Li, and X. Li, "A fast neighborhood grouping method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5028–5039, 2021.

[18] W. Sun and Q. Du, "Hyperspectral band selection: A review," *IEEE Geosci.Remote Sens. Mag.*, vol. 7, no. 2, pp. 118–139, 2019.

[19] S. Huang, H. Zhang, and A. Pižurica, "Hybrid-hypergraph regularized multiview subspace clustering for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[20] S. Huang, H. Zhang and A. Pižurica, "Subspace clustering for hyperspectral images via dictionary learning with adaptive regularization," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[21] C.-I. Chang, Q. Du, T.-L. Sun, and M. L. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2631–2641, 1999.

[22] C.-I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, 2006.

[23] B. Guo, S. R. Gunn, R. I. Damper, and J. D. Nelson, "Band selection for hyperspectral image classification using mutual information," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 4, pp. 522–526, 2006.

[24] Q. Du and H. Yang, "Similarity-based unsupervised band selection for hyperspectral image analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 564–568, 2008.

[25] A. Datta, S. Ghosh, and A. Ghosh, "Combination of clustering and ranking techniques for unsupervised band selection of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2814–2823, 2015.

[26] S. Jia, G. Tang, J. Zhu, and Q. Li, "A novel ranking-based clustering approach for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 88–102, 2015.

[27] H. Su, Q. Du, G. Chen, and P. Du, "Optimized hyperspectral band selection using particle swarm optimization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2659–2670, 2014.

[28] J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang, "Mutual-information-based semi-supervised hyperspectral band selection with high discrimination, high information, and low redundancy," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2956–2969, 2014.

[29] M. Gong, M. Zhang, and Y. Yuan, "Unsupervised band selection based on evolutionary multiobjective optimization for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 544–557, 2015.

[30] J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang, "Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images," *Pattern Recognit.*, vol. 51, pp. 295–309, 2016.

[31] H. Su, Y. Cai, and Q. Du, "Firefly-algorithm-inspired framework with band selection and extreme learning machine for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 1, pp. 309–320, 2016.

[32] Y.-Q. Zhao, L. Zhang, and S. G. Kong, "Band-subset-based clustering and fusion for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 747–756, 2010.

[33] S. Jia, Z. Ji, Y. Qian, and L. Shen, "Unsupervised band selection for hyperspectral imagery classification without manual band removal," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 531–543, 2012.

[34] V. Kumar, J. Hahn, and A. M. Zoubir, "Band selection for hyperspectral images based on self-tuning spectral clustering," in *Proc. 21st European Signal Process. Conf.*, 2013, pp. 1–5.

[35] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5910–5922, 2018.

[36] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Laplacian-regularized low-rank subspace clustering for hyperspectral image band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1723–1740, 2019.

[37] W. Sun, J. Peng, G. Yang, and Q. Du, "Fast and latent low-rank subspace clustering for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3906–1915, 2020.

[38] S. Huang, H. Zhang, and A. Pižurica, "A structural subspace clustering approach for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[39] W. Sun, J. Peng, G. Yang, and Q. Du, "Correntropy-based sparse spectral clustering for hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 3, pp. 484–488, 2020.

[40] W. Sun, L. Zhang, B. Du, W. Li, and Y. M. Lai, "Band selection using improved sparse subspace clustering for hyperspectral

imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2784–2797, 2015.

[41] M. Bevilacqua and Y. Berthoumieu, "Multiple-feature kernel-based probabilistic clustering for unsupervised band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6675–6689, 2019.

[42] 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.

[43] L. Zhang, L. Song, B. Du, and Y. Zhang, "Nonlocal low-rank tensor completion for visual data," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 673–685, 2021.

[44] R. Dian, S. Li, L. Fang, T. Lu, and J. M. Bioucas-Dias, "Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion," *IEEE Trans. Cybern.*, 2019.

[45] X. Zhang, G. Wen, and W. Dai, "A tensor decomposition-based anomaly detection algorithm for hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5801–5820, 2016.

[46] T. Imbiriba, R. A. Borsoi, and J. C. M. Bermudez, "Low-rank tensor modeling for hyperspectral unmixing accounting for spectral variability," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1833–1842, 2019.

[47] M. K.-P. Ng, Q. Yuan, L. Yan, and J. Sun, "An adaptive weighted tensor completion method for the recovery of remote sensing images with missing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3367–3381, 2017.

[48] J. Xue, Y. Zhao, W. Liao, J. C.-W. Chan, and S. G. Kong, "Enhanced sparsity prior model for low-rank tensor completion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4567–4581, 2020.

[49] J. Xue, Y. Zhao, W. Liao, and S. G. Kong, "Joint spatial and spectral low-rank regularization for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 1940–1958, 2017.

[50] J. Xue, Y. Zhao, W. Liao, and J. C.-W. Chan, "Nonlocal low-rank regularized tensor decomposition for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5174–5189, 2019.

[51] H. Zhang, L. Liu, W. He, and L. Zhang, "Hyperspectral image denoising with total variation regularization and nonlocal low-rank tensor decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3071–3084, 2019.

[52] M. A. Veganzones, J. E. Cohen, R. C. Farias, J. Chanussot, and P. Comon, "Nonnegative tensor CP decomposition of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2577–2588, 2015.

[53] S. Yang, M. Wang, P. Li, L. Jin, B. Wu, and L. Jiao, "Compressive hyperspectral imaging via sparse tensor and nonlinear compressed sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 5943–5957, 2015.

[54] Y. Gao, X. Wang, Y. Cheng, and Z. J. Wang, "Dimensionality reduction for hyperspectral data based on class-aware tensor neighborhood graph and patch alignment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1582–1593, 2014.

[55] X. Guo, X. Huang, L. Zhang, L. Zhang, A. Plaza, and J. A. Benediktsson, "Support tensor machines for classification of hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3248–3264, 2016.

[56] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[57] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *Proc. IEEE Int. Conf. Comput.*, 2015, pp. 1582–1590.

[58] Y. Xie, D. Tao, W. Zhang, Y. Liu, L. Zhang, and Y. Qu, "On unifying multi-view self-representations for clustering by tensor multi-rank minimization," *Int. J. Comput. Vis.*, vol. 126, no. 11, pp. 1157–1179, 2018.

[59] J. Wu, Z. Lin, and H. Zha, "Essential tensor learning for multi-view spectral clustering," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5910–5922, 2019.

[60] C. Zhang, H. Fu, J. Wang, W. Li, X. Cao, and Q. Hu, "Tensorized multi-view subspace representation learning," *Int. J. Comput. Vis.*, vol. 128, no. 8, pp. 2344–2361, 2020.

[61] Y. Chen, X. Xiao, and Y. Zhou, "Multi-view subspace clustering via simultaneously learning the representation tensor and affinity matrix," *Pattern Recognit.*, vol. 106, 2020.

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

[63] Y. Chen, W. He, N. Yokoya, and T.-Z. Huang, "Hyperspectral image restoration using weighted group sparsity-regularized low-rank tensor decomposition," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3556–3570, 2020.

[64] Y. Yankelevsky and M. Elad, "Dual graph regularized dictionary learning," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 4, pp. 611–624, 2016.

[65] T. Yokota, R. Zdunek, A. Cichocki, and Y. Yamashita, "Smooth nonnegative matrix and tensor factorizations for robust multi-way data analysis," *Signal Process.*, vol. 113, pp. 234–249, 2015.

[66] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Mathematical Programming*, vol. 142, no. 1-2, pp. 397–434, 2013.

[67] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[68] Q. Wang, Q. Li, and X. Li, "Hyperspectral band selection via adaptive subspace partition strategy," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 4940–4950, 2019.

[69] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, 2013.

[70] P. Zhou, C. Lu, J. Feng, Z. Lin, and S. Yan, "Tensor low-rank representation for data recovery and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1718–1732, 2021.