

A Structural Subspace Clustering Approach for Hyperspectral Band Selection

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

Abstract—Band selection, which removes irrelevant bands from hyperspectral images (HSIs) and keeps essential spectral information contained in a relatively few bands, allows huge savings in data storage, computation time and imaging hardware. In this paper, we propose a novel structural subspace clustering (STSC) method for hyperspectral band selection, which leverages the self-representation property of data and structural prior information to learn the cluster structure of bands. Particularly, we propose a general clustering model where the coarse coefficients matrix derived from a self-representation model is decomposed as a combination of a desirable coefficients matrix and a sparse matrix. This strategy adaptively adjusts the coarse coefficients matrix to learn the intrinsic data structure in low-dimensional subspaces. To guide this learning process, we introduce a structural regularization approach which makes use of the prior information about local and global properties of spectral bands. Moreover, we incorporate also prior knowledge about the dictionary, which demonstrates to yield a better clustering performance. We develop an adaptive method to estimate the number of selected bands by analysing eigenvalue gaps of Laplacian matrix. To solve the resulting model, an efficient algorithm based on alternating direction method of multipliers (ADMM) is developed. Extensive experiments on benchmark HSIs show that our method outperforms the state-of-the-art.

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

I. INTRODUCTION

HYPERSPECTRAL images (HSIs) capture spectral signatures of observed objects in hundreds of spectral bands, enabling thereby to differentiate between materials that are often indistinguishable in visible and conventional multispectral images. Hence, HSIs enjoy a great success in numerous emerging domains such as food safety, environmental monitoring and art investigation. However, their processing is challenging due to high dimensionality (often hundreds of bands) and redundant spectral information [1]. Problems associated with analysis of such high-dimensional data are referred to as the curse of dimensionality [2].

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, TELIN-GAIM, 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).

Feature extraction and band selection are two common approaches to reduce the dimensionality of HSIs. Both aim at extracting relevant information from the data for the subsequent data analysis, reducing thereby the computational complexity with the minimum loss of information. Representative feature extraction methods for HSIs, such as principle component analysis (PCA) [3, 4], linear discriminant analysis (LDA) [5, 6] and locality preserving projections (LPP) [7, 8], transform the initial spectral bands to a new set of features by a learned projection matrix while keeping relevant information as much as possible. The resulting features with the projection operation are derived from a linear combination of the initial spectral bands. However, this results in an explicit spectral distortion, and thereby makes the physical interpretation difficult and obstacles applications with the demand on physical spectral measures [9].

Band selection removes less relevant (often redundant) spectral bands and selects a subset of informative ones from an HSI, preserving well the original spectral information from sensors. Thus, band selection facilitates data storage, and enables reduction of the algorithmic complexity as well as simplifying imaging hardware in data acquisition. Band selection can be carried out in unsupervised [10–13], semi-supervised [14–18] and supervised [19–22] manner depending on how labeled samples and unlabeled samples are involved in the model (see a recent review paper [23] for more details). Since data annotation is expensive in practice, resulting in scarce labeled data, in this paper we mainly focus on unsupervised band selection method. Unsupervised band selection generally can be categorized into four groups: ranking-based, clustering-based, searching-based and hybrid methods. Ranking-based methods select the top-ranked bands according to a given criterion, such as energy maximization [24] and band correlation minimization [25, 26]. Clustering-based approaches first use clustering algorithms to decorrelate spectral bands [9, 13, 27–30]. All spectral bands are consequently assigned into different groups based on their similarities. Since the spectral bands in each group typically show high similarity, by selecting the bands that are closest to the centroids in each cluster, the most representative bands in HSIs are obtained. Searching based methods either update all the selected bands iteratively by using evolutionary searching approach [31, 32] or sequentially select/remove bands based on previously selected bands until the desired number of bands is reached [33]. Hybrid methods combine different types of band selection approaches, such as clustering-based and ranking-based methods [34, 35], to make the best use of each of them.

Recently, self-representation based subspace clustering (SSC) models have achieved remarkable results in unsupervised band selection [30, 36]. In general, SSC assumes that each data point can be represented by a linear combination of a few others from the same subspace [37]. By using the input data as a dictionary, SSC aims at solving a sparse coding problem to reveal the affinity structure of data, which is further applied in spectral clustering to obtain clustering results [38]. Zhu et al. develop a low-rank representation (LRR) based band selection model by regularizing the coefficients matrix to be low rank [12]. The utilized low-rank constraint captures effectively global data property, which enables a desirable classification performance with the selected bands. Sun et al. propose a collaborative representation based band selection method by using a ℓ_2 norm for the coefficients matrix, which results in a closed-form solution and thereby a fast band selection method [36]. Zhai et al. take into account the correlations of neighbouring bands in the LRR model by adopting a Laplacian-based regularization, leading to an improved performance than LRR [29]. A recent latent low-rank subspace clustering method [30] is proposed to solve the problem of insufficient band samples in LRR model. The learning model uncovers hidden features in HSIs and achieves the state-of-the-art performance in the field.

However, there is still an important gap in terms of exploiting local correlations among the bands, which limits the current performance. Typically, the existing methods treat each spectral band independently in the optimization process like [36] (resulting in sensitive performance to sparse noise) or adopt a global low-rank constraint as [12, 30] (which does not capture local correlations among neighbouring bands). The method in [29] considers local correlations of neighbouring bands by using a Laplacian-based regularization. This constraint imposed on the coefficients matrix cannot ensure the subspace representations of neighbouring bands from the same subspace to be similar. Moreover, none of the existing works exploited the inherent local and global information in the dictionary itself. We believe this should help to learn better the intrinsic data cluster structure. However, this aspect was not explored yet in the literature.

In this paper, we propose a structural subspace clustering method (STSC) for hyperspectral band selection. Compared with the aforementioned SSC models, which impose regularizations on the coefficients matrix derived from self-representation model, we propose a more general framework where self-representation and regularizations are formulated separately with respect to different variables. Specifically, we decompose the coarse coefficient matrix from the self-representation as a sum of a desirable coefficients matrix and a sparse error matrix. The desirable coefficients matrix is learned under the guidance of a novel structural regularization. Taking into account the prior information of bands in terms of local and global property, we formulate this structural regularization comprehensively on three aspects. Firstly, we propose a $\ell_{2,1}$ norm based joint total variation to capture the local similarity of adjacent bands in the input data. Secondly, as the input data is employed as a dictionary, the neighbouring atoms in the dictionary also show strong correlations. We explore such

structural prior information with a 1-D total variation, which ensures a consistent response of the neighbouring atoms. Thirdly, we employ a spectral nuclear norm regularization to capture the global low-rank property, which admits the fact that the bands from the same subspaces are distributed globally. With the proposed structural regularization, our model learns more comprehensively the cluster structure in the low-dimensional subspaces, facilitating thereby the performance of band selection. We also develop an efficient algorithm to solve the resulting optimization problem. Moreover, we propose an adaptive method to estimate the number of selected bands by analysing eigenvalue gaps of Laplacian matrix. Experimental results on real HSIs demonstrate the superior performance of our band selection method over the state-of-the-art.

The rest of this paper is organized as follows. Section II reviews briefly the subspace clustering based band selection methods for HSIs. Section III presents the proposed structural subspace clustering method for band selection and develops an efficient algorithm to solve the resulting optimization problem. Experiments and analysis on benchmark datasets are conducted in Section IV. Section V concludes the paper.

II. PRIOR WORKS

In this section, we briefly review the existing unsupervised band selection methods based on subspace clustering. Let $\mathbf{X} \in \mathbb{R}^{N \times B}$ represent the input matrix of a HSI, where N is the total amount of pixels and B is the number of spectral bands. A general subspace clustering problem is formulated as follows:

$$\arg \min_{\mathbf{C}} \Phi(\mathbf{X} - \mathbf{X}\mathbf{C}) + \lambda\Gamma(\mathbf{C}), \quad s.t. \text{diag}(\mathbf{C}) = \mathbf{0} \quad (1)$$

where $\mathbf{C} \in \mathbb{R}^{B \times B}$ is a learned coefficients matrix. $\Phi(\cdot)$ is a data fidelity function, typically the Frobenius norm $\|\mathbf{X}\|_F^2 = \sum_{i,j} X_{ij}^2$ or the $\ell_{2,1}$ norm $\|\mathbf{X}\|_{2,1} = \sum_j \sqrt{\sum_i X_{ij}^2}$. $\Gamma(\mathbf{C})$ is a regularization term, which encodes prior information about \mathbf{C} . Common regularization functions include $\|\mathbf{C}\|_F^2$ [36], $\|\mathbf{C}\|_{1,2}$ [39, 40] and the nuclear norm [12, 29, 30] $\|\mathbf{C}\|_* = \sum_i \sigma_i$ with σ_i being the singular values of \mathbf{C} . λ is a regularization parameter. Observe that the clustering model (1) is formulated as a self-representation model where the input data \mathbf{X} is utilized as a dictionary. The constraint $\text{diag}(\mathbf{C}) = \mathbf{0}$ and the regularization term $\Gamma(\mathbf{C})$ prevent trivial solutions.

These regularizations promote different characteristics of the learned coefficients matrix \mathbf{C} . For instance, the ℓ_2 norm based regularization $\|\mathbf{C}\|_F^2$ promotes a representation associated with all atoms in the subspace representation. The $\ell_{1,2}$ norm $\|\mathbf{C}\|_{1,2}$ facilitates a joint representation of \mathbf{X} by a few common spectral bands, resulting in a row-sparsity pattern of \mathbf{C} . The nuclear norm regularization $\|\mathbf{C}\|_*$ favors a low rank coefficients matrix \mathbf{C} .

The learned coefficients matrix \mathbf{C} is utilized to construct a similarity matrix $\mathbf{W} \in \mathbb{R}^{B \times B}$, typically as $\mathbf{W} = (|\mathbf{C}| + |\mathbf{C}^T|)/2$, where $|\mathbf{C}|$ takes the absolute values of \mathbf{C} and \mathbf{C}^T is the transpose of \mathbf{C} . Other methods to build \mathbf{W} include the weighted-square strategy [29] and the correntropy-based approach [30]. By applying \mathbf{W} within the standard spectral clustering, we obtain the clustering results. Specifically, the c eigenvectors $\{\mathbf{v}_k\}_{k=1}^c$ of the Laplacian matrix $\mathbf{L}_\mathbf{W} =$

$\text{Diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}$ corresponding to its c smallest eigenvalues are first calculated and then the clustering results are obtained by applying the k-means algorithm to the matrix $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_c]$, where c is the number of selected bands. Finally, the bands that are closest to the centroid bands within each cluster are selected.

III. STRUCTURAL SUBSPACE CLUSTERING FOR BAND SELECTION

In this section, we propose a novel structural subspace clustering method for hyperspectral band selection. We also present a new method to estimate the number of selected bands. Moreover, an efficient algorithm is developed to solve the resulting optimization problem.

A. Structural Subspace Clustering

We build our STSC method on the typical self-representation model which employs the input data \mathbf{X} as the dictionary, i.e., $\mathbf{X} \approx \mathbf{X}\mathbf{A}$. However, due to the effect of noise and within-cluster variability, the resulting coefficients matrix $\mathbf{A} \in \mathbb{R}^{B \times B}$ often deviates from the optimal solution [41, 42]. Here, we propose to decompose the coarse coefficients matrix \mathbf{A} as a combination of a desirable coefficients matrix $\mathbf{L} \in \mathbb{R}^{B \times B}$ and a sparse error matrix $\mathbf{S} \in \mathbb{R}^{B \times B}$, i.e., $\mathbf{A} = \mathbf{L} + \mathbf{S}$. The sparse matrix \mathbf{S} represents the error between \mathbf{A} and \mathbf{L} , and \mathbf{L} can be seen as a refined matrix of \mathbf{A} . Our STSC model aims to optimize the following objective function:

$$\begin{aligned} \arg \min_{\mathbf{A}, \mathbf{L}, \mathbf{S}} \Phi(\mathbf{X} - \mathbf{X}\mathbf{A}) + \lambda \Psi(\mathbf{L}) + \beta \|\mathbf{S}\|_1 \\ \text{s.t. } \mathbf{A} = \mathbf{L} + \mathbf{S}, \end{aligned} \quad (2)$$

where λ and β are two regularization parameters and $\Psi(\mathbf{L})$ is a structural regularization for \mathbf{L} which is detailed in the following subsection. We utilize the Frobenius norm $\|\cdot\|_F^2$ for the function $\Phi(\cdot)$. The regularization $\Psi(\mathbf{L})$ encodes the desirable local and global structure of data. Clearly, in (2) the data fidelity term with respect to \mathbf{A} and the regularization term with respect to \mathbf{L} are linked by the constraint $\mathbf{A} = \mathbf{L} + \mathbf{S}$. The coarse coefficients matrix \mathbf{A} may fit well in the data fidelity term but may not respect the desirable structure in the low-dimensional subspaces. This decomposition strategy allows an adaptive adjustment under the guidance of the structural regularization. Compared with the common formulation (1), our model is more general. In particular, when β is sufficiently large, the high penalty on the sparsity term $\|\mathbf{S}\|_1$ will promote \mathbf{S} to be zero, which leads to $\mathbf{A} = \mathbf{L}$ and our model reduces to (1).

After solving (2), we construct the similarity matrix as $\mathbf{W} = (|\mathbf{L}| + |\mathbf{L}|^T)/2$ and apply it in the spectral clustering [43] to obtain the clustering results of spectral bands. Finally, the bands that are closest to the centroids within each cluster are selected as the representatives.

B. Structural Regularizations

Neighbouring bands in a HSI tend to be highly similar. This is the inherent local property of hyperspectral bands. Some

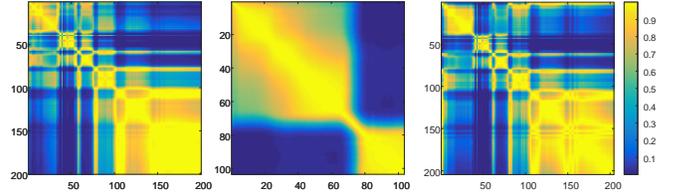


Fig. 1. Correlations of spectral bands in three typical HSIs: *Indian Pines* (left), *Pavia University* (middle) and *Salinas* (right).

bands that are not adjacent may also show high similarity due to the analogous spectral response of objects in the corresponding spectrum. This is a global property of hyperspectral bands. We take three typical HSIs as examples to show these properties: *Indian Pines*, *Pavia University* and *Salinas*. Let W_{ij} denote the similarity between i -th and j -th bands, which is calculated by $W_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / \sigma^2}$ where \mathbf{x}_i is the i -th spectral band. The similarity matrix shown in Fig. 1 reveals that high similarities of bands occur mostly in the diagonal regions which correspond to neighbouring bands in a narrow range of wavelength. Besides, high similarities also appear in the regions that are far way from the diagonal. This reveals the global property of bands where some non-adjacent bands have strong correlations.

Observing these properties, we propose a structural regularization $\Psi(\mathbf{L})$, which incorporates the local and global information simultaneously, to guide the learning of the coefficients matrix \mathbf{L} . Previous works [12, 29, 30, 44–49] demonstrate the effectiveness of using low-rank regularization in capturing the global property of data. Here, we adopt a nuclear norm based low-rank regularization $\|\mathbf{L}\|_*$ as one component of the structural regularization, with the goal of seeking a representation of the input data which is low rank. This promotes the participations of non-adjacent bands from the same subspace in the approximation of a query band [29], facilitating the preservation of global property of data.

Regarding the local correlations of bands, we take into account this prior information from two points of view. Firstly, since the neighbouring bands in the input matrix \mathbf{X} are often similar, their coefficients vectors in the subspace representation ideally should be close, by which the original geometric data structure can be well preserved in the low-dimensional subspaces without distortions. This results in the local smoothness of coefficients matrix in the horizontal direction. It should be noted that not all the neighbouring bands are similar as shown in Fig. 1. The coefficients vectors of these significantly different bands should be dissimilar, resulting in thereby huge differences (edges) in the ideal coefficients matrix along samples direction. As only a few adjacent bands pronounce such huge dissimilarities, the edges are sparse. To this end, we propose a $\ell_{2,1}$ norm based joint total variation to account for the prior information of local property in the input data:

$$\|\nabla_x \mathbf{L}\|_{2,1} = \sum_j \sqrt{\sum_i (\nabla_x L_{i,j})^2}, \quad (3)$$

where $\nabla_x \mathbf{L}$ is the first-order forward difference matrix of \mathbf{L}

in horizontal direction with each entry $\nabla_x L_{i,j} = L_{i,j+1} - L_{i,j}$. We take a periodical operation for the matrix boundaries. The joint total variation regularization in (3) promotes column sparsity of the difference matrix $\nabla_x \mathbf{L}$, which leads to only a few columns of $\nabla_x \mathbf{L}$ to be non-zero and most columns to be zero. This guarantees that most of the neighbouring columns in \mathbf{L} are close.

We observe that data properties are also reflected in the dictionary (which is equal to the input data). Hence, neighbouring atoms are highly correlated. We want to exploit this local structure in the dictionary to improve the clustering. Ideally, atoms in the dictionary shall act in a consistent way in the representation-related problems, which means that the similar atoms shall response similarly in the subspace representation while dissimilar atoms yield different responses. Coefficients values measure the responses of atoms in the subspace representation, indicating the contributions of atoms. For instance, let $L_{i,j}$ represent the desirable response of atom \mathbf{x}_i to input data \mathbf{x}_j . Assume the neighbouring atom \mathbf{x}_{i+1} is close to \mathbf{x}_i , we infer that its response to input data \mathbf{x}_j , i.e., $L_{i+1,j}$, is similar to $L_{i,j}$. This results in the smoothness of coefficients matrix in vertical direction. While a few neighbouring atoms are significantly different as discussed before, the associated coefficients, e.g., $L_{i,j}$ and $L_{i,j+1}$, are considerably dissimilar, resulting in large differences (edges) in the desirable coefficients matrix. In consideration of the local property of dictionary, we adopt a 1-D total variation as one component of the structural regularization:

$$\|\nabla_y \mathbf{L}\|_1 = \sum_j \sum_i |\nabla_y L_{i,j}|, \quad (4)$$

where $\nabla_y \mathbf{L}$ is the first-order forward difference matrix of \mathbf{L} in vertical direction with each entry $\nabla_y L_{i,j} = L_{i+1,j} - L_{i,j}$. The regularization (4) results in a sparse solution for each column of $\nabla_y \mathbf{L}$, where most of the elements are zero or near zero. As $\nabla_y \mathbf{L}$ is the first-order difference matrix, most of the neighbouring entries for each column of \mathbf{L} , i.e., $L_{i,j}$ and $L_{i+1,j}$, will be close.

Finally, we obtain the proposed structural regularization $\Psi(\mathbf{L})$ as follows:

$$\Psi(\mathbf{L}) = \lambda_1 \|\nabla_x \mathbf{L}\|_{2,1} + \lambda_2 \|\nabla_y \mathbf{L}\|_1 + \lambda_3 \|\mathbf{L}\|_*, \quad (5)$$

where λ_1 , λ_2 and λ_3 are three regularization parameters. Then, we formulate completely the proposed structural subspace clustering model as follows:

$$\arg \min_{\mathbf{A}, \mathbf{L}, \mathbf{S}} \frac{1}{2} \|\mathbf{X} - \mathbf{XA}\|_F^2 + \lambda_1 \|\nabla_x \mathbf{L}\|_{2,1} + \lambda_2 \|\nabla_y \mathbf{L}\|_1 + \lambda_3 \|\mathbf{L}\|_* + \beta \|\mathbf{S}\|_1, \quad s.t. \quad \mathbf{A} = \mathbf{L} + \mathbf{S}. \quad (6)$$

C. Estimation of the Number of Selected Bands

In this section, we propose a new method to estimate the number of selected bands by making use of eigenvalue gaps of graph Laplacian. The ideal similarity matrix has a block-diagonal structure because only the bands that are from the same cluster are connected. The number of connected components in the similarity matrix is associated with the

spectral property of the Laplacian matrix \mathbf{L}_W according to the following Theorem:

Theorem 1. [43] *Let \mathbf{W} be an undirected graph with non-negative weights. Then the multiplicity c of the eigenvalue 0 of \mathbf{L}_W equals the number of connected components (blocks) in the graph.*

Let $\sigma_i(\mathbf{L}_W)$ be the eigenvalues of \mathbf{L}_W in the decreasing order. We define the eigenvalue gaps d_i as the absolute difference of two adjacent values of $\sigma_i(\mathbf{L}_W)$, i.e., $d_i = |\sigma_{i+1}(\mathbf{L}_W) - \sigma_i(\mathbf{L}_W)|$. According to Theorem 1, $\sigma_i(\mathbf{L}_W) > 0$ for $i = 1, \dots, B - c$ and $\sigma_i(\mathbf{L}_W) = 0$ for $i = B - c + 1, \dots, B$ in the ideal case. Thus, there is a relatively big eigenvalue gap between $\sigma_c(\mathbf{L}_W)$ and $\sigma_{c+1}(\mathbf{L}_W)$, resulting in a larger value of d_c . To find the number c such that all $\sigma_1(\mathbf{L}_W), \dots, \sigma_{B-c}(\mathbf{L}_W)$ are large, but $\sigma_{B-c+1}(\mathbf{L}_W)$ is relatively small, we estimate c by making use of the first-order difference of eigenvalues, i.e., d_i , as follows:

$$c = B - k + 1 \quad (7)$$

where k is the smallest index which satisfies that $d_{n_k} = d_k / \sigma_k(\mathbf{L}_W)$ is no less than a threshold value δ . d_{n_i} is a normalized value of d_i in the range of $[0, 1]$ for the purpose of setting adaptively the number of bands with a proper δ regardless of the types of hyperspectral images. The benefit of normalization will be shown in the experiments. According to (7), we can see our method provides a computationally efficient solution for the estimation of the number of selected bands.

D. Optimization

To solve the resulting optimization problem (6), we develop an efficient algorithm based on the alternating direction method of multipliers (ADMM) [50]. First, we introduce three auxiliary variables $\mathbf{V}_1 = \nabla_x \mathbf{L}$, $\mathbf{V}_2 = \nabla_y \mathbf{L}$ and $\mathbf{V}_3 = \mathbf{L}$ to split the difficult optimization problem (6) into several sub-problems which can be solved separately. The problem (6) is reformulated equivalently by:

$$\arg \min_{\mathbf{A}, \mathbf{L}, \mathbf{S}, \mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3} \frac{1}{2} \|\mathbf{X} - \mathbf{XA}\|_F^2 + \lambda_1 \|\mathbf{V}_1\|_{2,1} + \lambda_2 \|\mathbf{V}_2\|_1 + \lambda_3 \|\mathbf{V}_3\|_* + \beta \|\mathbf{S}\|_1, \quad s.t. \quad \mathbf{A} = \mathbf{L} + \mathbf{S}, \quad \nabla_x \mathbf{L} = \mathbf{V}_1, \quad \nabla_y \mathbf{L} = \mathbf{V}_2, \quad \mathbf{L} = \mathbf{V}_3. \quad (8)$$

We derive the augmented Lagrangian function of (8) as follows:

$$\begin{aligned} \mathcal{L} = & \frac{1}{2} \|\mathbf{X} - \mathbf{XA}\|_F^2 + \lambda_1 \|\mathbf{V}_1\|_{2,1} + \lambda_2 \|\mathbf{V}_2\|_1 \\ & + \lambda_3 \|\mathbf{V}_3\|_* + \beta \|\mathbf{S}\|_1 + \langle \mathbf{Y}_1, \mathbf{A} - \mathbf{L} - \mathbf{S} \rangle \\ & + \langle \mathbf{Y}_2, \nabla_x \mathbf{L} - \mathbf{V}_1 \rangle + \langle \mathbf{Y}_3, \nabla_y \mathbf{L} - \mathbf{V}_2 \rangle + \langle \mathbf{Y}_4, \mathbf{L} - \mathbf{V}_3 \rangle \\ & + \frac{\mu}{2} (\|\mathbf{A} - \mathbf{L} - \mathbf{S}\|_F^2 + \|\nabla_x \mathbf{L} - \mathbf{V}_1\|_F^2 \\ & + \|\nabla_y \mathbf{L} - \mathbf{V}_2\|_F^2 + \|\mathbf{L} - \mathbf{V}_3\|_F^2), \end{aligned} \quad (9)$$

where $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ and \mathbf{Y}_4 are Lagrange multipliers and μ is a penalty parameter.

We then solve all the unknown variables iteratively by updating one while fixing others. The sub-problem with respect to \mathbf{A} is formulated by:

$$\arg \min_{\mathbf{A}} \frac{1}{2} \|\mathbf{X} - \mathbf{X}\mathbf{A}\|_F^2 + \frac{\mu}{2} \|\mathbf{A} - \mathbf{L} - \mathbf{S} + \frac{\mathbf{Y}_1}{\mu}\|_F^2. \quad (10)$$

By setting the derivative of (10) with respect to \mathbf{A} to zero, we update \mathbf{A} as follows:

$$\mathbf{A} = (\mathbf{X}^T \mathbf{X} + \mu \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{X} + \mu (\mathbf{L} + \mathbf{S} - \frac{\mathbf{Y}_1}{\mu})). \quad (11)$$

The sub-problem with respect to \mathbf{L} is to minimize the following objective function:

$$\arg \min_{\mathbf{L}} \frac{1}{2} (\|\mathbf{A} - \mathbf{L} - \mathbf{S} + \frac{\mathbf{Y}_1}{\mu}\|_F^2 + \|\nabla_x \mathbf{L} - \mathbf{V}_1 + \frac{\mathbf{Y}_2}{\mu}\|_F^2 + \|\nabla_y \mathbf{L} - \mathbf{V}_2 + \frac{\mathbf{Y}_3}{\mu}\|_F^2 + \|\mathbf{L} - \mathbf{V}_3 + \frac{\mathbf{Y}_4}{\mu}\|_F^2). \quad (12)$$

Denote by \mathbf{H}_x and \mathbf{H}_y the difference operators in horizontal and vertical directions, we reformulate the difference matrices $\nabla_x \mathbf{L}$ and $\nabla_y \mathbf{L}$ as vector forms by $\mathbf{H}_x \mathbf{L}(\cdot)$ and $\mathbf{H}_y \mathbf{L}(\cdot)$, respectively, where $\mathbf{L}(\cdot)$ represents a reshaped vector from the matrix \mathbf{L} . We derive the equivalent problem of (12) as follows:

$$\begin{aligned} \arg \min_{\mathbf{L}(\cdot)} & \frac{1}{2} (\|\mathbf{A}(\cdot) - \mathbf{L}(\cdot) - \mathbf{S}(\cdot) + \frac{\mathbf{Y}_1(\cdot)}{\mu}\|_F^2 \\ & + \|\mathbf{H}_x \mathbf{L}(\cdot) - \mathbf{V}_1(\cdot) + \frac{\mathbf{Y}_2(\cdot)}{\mu}\|_F^2 \\ & + \|\mathbf{H}_y \mathbf{L}(\cdot) - \mathbf{V}_2(\cdot) + \frac{\mathbf{Y}_3(\cdot)}{\mu}\|_F^2 \\ & + \|\mathbf{L}(\cdot) - \mathbf{V}_3(\cdot) + \frac{\mathbf{Y}_4(\cdot)}{\mu}\|_F^2). \end{aligned} \quad (13)$$

We set the derivative of (13) to zero and obtain

$$\begin{aligned} & (\mathbf{H}_x^T \mathbf{H}_x + \mathbf{H}_y^T \mathbf{H}_y + 2\mathbf{I}) \mathbf{L}(\cdot) \\ & = \mathbf{A}(\cdot) - \mathbf{S}(\cdot) + \frac{\mathbf{Y}_1(\cdot)}{\mu} + \mathbf{H}_x^T (\mathbf{V}_1(\cdot) - \frac{\mathbf{Y}_2(\cdot)}{\mu}) \\ & + \mathbf{H}_y^T (\mathbf{V}_2(\cdot) - \frac{\mathbf{Y}_3(\cdot)}{\mu}) + \mathbf{V}_3(\cdot) - \frac{\mathbf{Y}_4(\cdot)}{\mu}. \end{aligned} \quad (14)$$

As \mathbf{H}_x and \mathbf{H}_y are two convolutions, the above problem can be solved by using the fast Fourier transform (FFT) method:

$$\mathbf{L}(\cdot) = \mathcal{F}^{-1} \left[\frac{\mathbf{P}}{2 + \mathcal{F}(\mathbf{H}_x)^2 + \mathcal{F}(\mathbf{H}_y)^2} \right], \quad (15)$$

where $\mathbf{P} = \mathcal{F}(\mathbf{A}(\cdot) - \mathbf{S}(\cdot) + \mathbf{Y}_1(\cdot)/\mu + \mathbf{H}_x^T (\mathbf{V}_1(\cdot) - \mathbf{Y}_2(\cdot)/\mu) + \mathbf{H}_y^T (\mathbf{V}_2(\cdot) - \mathbf{Y}_3(\cdot)/\mu) + \mathbf{V}_3(\cdot) - \mathbf{Y}_4(\cdot)/\mu)$, and $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ are the operators of the FFT and the inverse FFT, respectively.

The objective function with respect to \mathbf{S} is formulated by

$$\arg \min_{\mathbf{S}} \beta \|\mathbf{S}\|_1 + \frac{\mu}{2} \|\mathbf{A} - \mathbf{L} - \mathbf{S} + \frac{\mathbf{Y}_1}{\mu}\|_F^2 \quad (16)$$

We introduce the following soft-thresholding operator:

$$\mathcal{R}_{\Delta}(x) = \begin{cases} \text{sgn}(x)(|x| - \Delta) & |x| \geq \Delta \\ 0 & \text{otherwise,} \end{cases} \quad (17)$$

then we derive the solution of (16) as [51–54]

$$\mathbf{S} = \mathcal{R}_{\frac{\beta}{\mu}}(\mathbf{A} - \mathbf{L} + \frac{\mathbf{Y}_1}{\mu}). \quad (18)$$

The sub-problem with respect to \mathbf{V}_1 is obtained by the following formula:

$$\arg \min_{\mathbf{V}_1} \lambda_1 \|\mathbf{V}_1\|_{2,1} + \frac{\mu}{2} \|\nabla_x \mathbf{L} - \mathbf{V}_1 + \frac{\mathbf{Y}_2}{\mu}\|_F^2. \quad (19)$$

Denote by \mathbf{v}_1^i the i -th column of \mathbf{V}_1 and \mathbf{q}^i the i -th column of $\nabla_x \mathbf{L} + \mathbf{Y}_2/\mu$, we solve (19) in a column-wise manner as follows:

$$\arg \min_{\mathbf{v}_1^i} \lambda_1 \|\mathbf{v}_1^i\|_2 + \frac{\mu}{2} \|\mathbf{q}^i - \mathbf{v}_1^i\|_F^2. \quad (20)$$

The optimal solution to (20) is formulated by

$$\mathbf{v}_1^i = (1 - \lambda_1/\mu/\|\mathbf{q}^i\|_2)_+ \mathbf{q}^i, \quad (21)$$

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

We update the variable \mathbf{V}_2 by minimizing the following objective function:

$$\arg \min_{\mathbf{V}_2} \lambda_2 \|\mathbf{V}_2\|_1 + \frac{\mu}{2} \|\nabla_y \mathbf{L} - \mathbf{V}_2 + \frac{\mathbf{Y}_3}{\mu}\|_F^2. \quad (22)$$

The solution of (22) is derived as follows:

$$\mathbf{V}_2 = \mathcal{R}_{\frac{\lambda_2}{\mu}}(\nabla_y \mathbf{L} + \frac{\mathbf{Y}_3}{\mu}). \quad (23)$$

The sub-problem with respect to \mathbf{V}_3 is shown by

$$\arg \min_{\mathbf{V}_3} \lambda_3 \|\mathbf{V}_3\|_* + \frac{\mu}{2} \|\mathbf{L} - \mathbf{V}_3 + \frac{\mathbf{Y}_4}{\mu}\|_F^2. \quad (24)$$

To solve this problem, we first introduce a soft-thresholding operator $\mathcal{D}_{\delta}(\mathbf{X})$. Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the singular value decomposition (SVD) of a matrix \mathbf{X} , i.e., $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where \mathbf{U} and \mathbf{V} are the left and right singular vectors, respectively, and $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values σ_i . The operator $\mathcal{D}_{\delta}(\mathbf{X})$ is defined as follows:

$$\mathcal{D}_{\delta}(\mathbf{X}) := \mathbf{U} \mathcal{D}_{\delta}(\mathbf{\Sigma}) \mathbf{V}^T, \quad (25)$$

where $\mathcal{D}_{\delta}(\mathbf{\Sigma}) = \text{diag}(\max((\sigma_i - \delta), 0))$.

Then we obtain the following solution of (24) by

$$\mathbf{V}_3 = \mathcal{D}_{\frac{\lambda_3}{\mu}}(\mathbf{L} + \frac{\mathbf{Y}_4}{\mu}). \quad (26)$$

We update other variables as follows, respectively:

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{Y}_1 + \mu(\mathbf{A} - \mathbf{L} - \mathbf{S}) \\ \mathbf{Y}_2 &= \mathbf{Y}_2 + \mu(\nabla_x \mathbf{L} - \mathbf{V}_1) \\ \mathbf{Y}_3 &= \mathbf{Y}_3 + \mu(\nabla_y \mathbf{L} - \mathbf{V}_2) \\ \mathbf{Y}_4 &= \mathbf{Y}_4 + \mu(\mathbf{L} - \mathbf{V}_3). \end{aligned} \quad (27)$$

Let $e = \max(e_1, e_2, e_3, e_4)$ with $e_1 = \|\mathbf{A} - \mathbf{L} - \mathbf{S}\|_{\infty}$, $e_2 = \|\nabla_x \mathbf{L} - \mathbf{V}_1\|_{\infty}$, $e_3 = \|\nabla_y \mathbf{L} - \mathbf{V}_2\|_{\infty}$ and $e_4 = \|\mathbf{L} - \mathbf{V}_3\|_{\infty}$. The above steps are iteratively updated until the stop criterion is satisfied, that is, $e < \varepsilon$ or the number of iterations exceeds a predefined value, i.e., the maximum number of iterations *MaxIter*. We set $\varepsilon = 1e-6$ and *MaxIter* = 100. Algorithm 1 summarizes the complete process to solve the optimization problem (6).

Algorithm 1 The solver to the proposed model in (6)

```

1: Input:  $\mathbf{X}$ ,  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$ ,  $\beta$ 
2: Initialize  $\mathbf{L}$ ,  $\mathbf{S}$ ,  $\{\mathbf{V}_i\}_{i=1}^3$ ,  $\{\mathbf{Y}_i\}_{i=1}^4$  as  $\mathbf{0}$  and  $\mu = 1$ 
3: while not converged do
4:   Update  $\mathbf{A}$  by (11)
5:   Update  $\mathbf{L}$  by (15)
6:   Update  $\mathbf{S}$  by (18)
7:   Update  $\mathbf{V}_1$  by (21)
8:   Update  $\mathbf{V}_2$  by (23)
9:   Update  $\mathbf{V}_3$  by (26)
10:  Update other variables by (27)
11: end while
12: Output: Matrix  $\mathbf{L}$ 

```

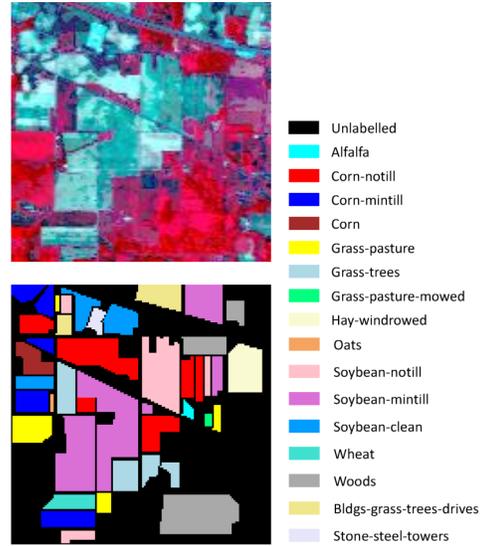
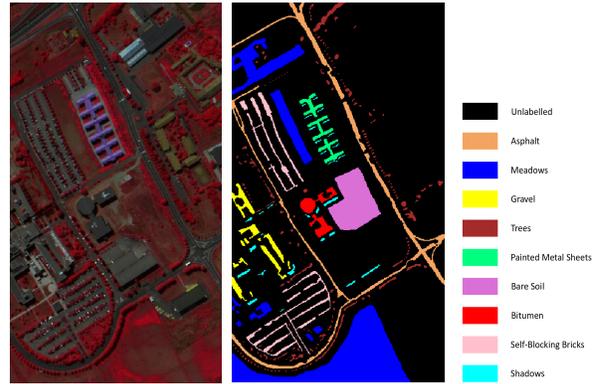
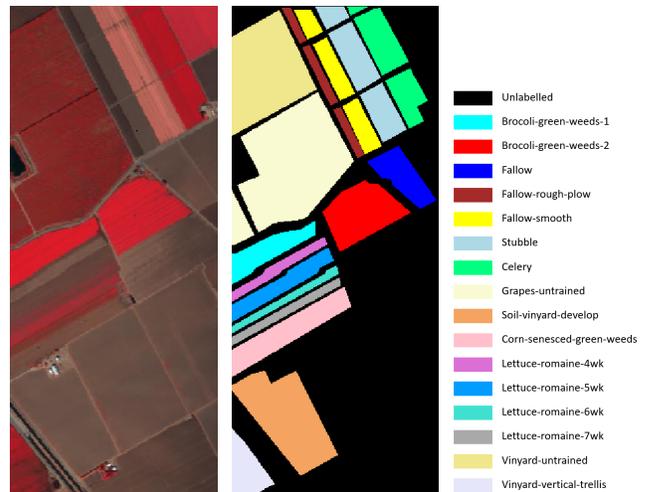
Next, we analyse the computational complexity of Algorithm 1. The updates of \mathbf{A} and \mathbf{L} have time complexity of $\mathcal{O}(B^3)$ and $\mathcal{O}(B^2 \log(B^2))$, respectively. The time complexity is $\mathcal{O}(B^2)$ for updating \mathbf{V}_1 and $\mathcal{O}(B^3)$ for updating \mathbf{V}_3 . The computational complexity for updating \mathbf{S} , \mathbf{V}_2 and $\{\mathbf{Y}_i\}_{i=1}^4$ is neglectable. Thus, the overall time complexity of Algorithm 1 is $\mathcal{O}(J(B^3 + B^2 \log(B^2)))$, where J is the number of iterations.

IV. EXPERIMENTS

The performance of band selection is often evaluated in classification task [23], where the selected bands are fed into classifiers. In our experiments, SVM [55] and K-nearest-neighbors (KNN) [56] are two employed classifiers. The benchmark methods for comparison consist of one ranking-based method maximum-variance principal component analysis (MVPCA) [24], one hybrid-based method FDPC [35], the saliency-based band selection method (SBSS) [57] and five clustering-based methods optimal clustering framework (OCF) [13], adaptive subspace partition strategy (ASPS) [58], ISSC [36], L2-LRSC [29] and FLLRSC [30]. Among the clustering-based methods, ISSC, L2-LRSC and FLLRSC select bands based on subspace clustering and have achieved the state-of-the-art performance. We utilize overall accuracy (OA), average accuracy (AA) and Kappa coefficient (κ) as quantitative assessments. Ten percent of labeled samples per class are randomly selected as training data and the rest are employed for testing. We repeat all experiments ten times and average results are reported. We implement the SVM with radial basis function kernel and the KNN by using the optimization toolbox in MATLAB 2016b. The number of neighbors of KNN is set to 3 as in [9, 58].

A. Datasets

1) *Indian Pines*: The *Indian Pines* (*IndianP*) was acquired by the Airborne/Visible Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines region in North-western Indiana on June 12, 1992, with 20-m spatial resolution per pixel. The image of size $145 \times 145 \times 220$ contains 16 classes and 10249 pixels are labeled as shown in Table I. In the experiment, we remove 20 spectral bands in 104-108, 150-163 and 200 due to water absorption. The false-color composite image and ground truth are shown in Fig. 2.

Fig. 2. False color image (top) and ground truth (bottom) of *Indian Pines*.Fig. 3. False color image (left) and ground truth (right) of *Pavia University*.Fig. 4. False color image (left) and ground truth (right) of *Salinas*.

2) *Pavia University*: The *Pavia University* (*PaviaU*) was captured 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

TABLE I

CLASS NAME AND THE AMOUNT OF LABELED SAMPLES IN *IndianP* (TEN PERCENT OF LABELLED SAMPLES PER CLASS ARE USED FOR TRAINING).

| No. | Class name | Train | Test | Total |
|-------|--------------------------|-------|------|-------|
| 1 | Alfalfa | 5 | 41 | 46 |
| 2 | Corn-notill | 143 | 1285 | 1428 |
| 3 | Corn-mintill | 83 | 747 | 830 |
| 4 | Corn | 24 | 213 | 237 |
| 5 | Grass-pasture | 49 | 434 | 483 |
| 6 | Grass-trees | 73 | 657 | 730 |
| 7 | Grass-pasture-mowed | 3 | 25 | 28 |
| 8 | Hay-windrowed | 48 | 430 | 478 |
| 9 | Oats | 2 | 18 | 20 |
| 10 | Soybean-notill | 98 | 874 | 972 |
| 11 | Soybean-mintill | 246 | 2209 | 2455 |
| 12 | Soybean-clean | 60 | 533 | 593 |
| 13 | Wheat | 21 | 184 | 205 |
| 14 | Woods | 127 | 1138 | 1265 |
| 15 | Bldgs-grass-trees-drives | 39 | 347 | 386 |
| 16 | Stone-steel-towers | 10 | 83 | 93 |
| Total | | 1031 | 9218 | 10249 |

TABLE II

CLASS NAME AND THE AMOUNT OF LABELED SAMPLES IN *PaviaU* (TEN PERCENT OF LABELLED SAMPLES PER CLASS ARE USED FOR TRAINING).

| No. | Class name | Train | Test | Total |
|-------|----------------------|-------|-------|-------|
| 1 | Asphalt | 664 | 5967 | 6631 |
| 2 | Meadows | 1865 | 16784 | 18649 |
| 3 | Gravel | 210 | 1889 | 2099 |
| 4 | Trees | 307 | 2757 | 3064 |
| 5 | Painted Metal Sheets | 135 | 1210 | 1345 |
| 6 | Bare Soil | 503 | 4526 | 5029 |
| 7 | Bitumen | 133 | 1197 | 1330 |
| 8 | Self-Blocking Bricks | 369 | 3313 | 3682 |
| 9 | Shadows | 95 | 852 | 947 |
| Total | | 4281 | 38495 | 42776 |

classes and 42776 labeled samples as shown in Table II. Fig. 3 shows the false color image and ground truth.

3) *Salinas*: The *Salinas* was collected by the AVIRIS sensor over the Salinas Valley, CA, USA. The geometric resolution is 3.7 m per pixel. The image of size $512 \times 217 \times 224$ contains 16 ground-truth classes with in total 54129 labeled pixels as shown in Table III. Twenty bands in 108-112, 154-167 and 224 are removed due to water absorption. We show the false color image and ground truth for the dataset *Salinas* in Fig. 4. All three data sets are accessible by http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.

B. Comparisons With the State-of-the-Art

We vary the number of selected bands c in the range of [5,50] with a step size of 5, and report the classification results of SVM and KNN in Figs. 5-7. We also report the classification results when using all the spectral bands for reference. Since SBSS selects automatically the spectral bands based on local extrema points of a Hessian matrix of HSI, the chosen bands are fixed. We do not include its results in Figs. 5-7 but in the following experiments. For STSC, We set $\lambda_1 = 0.1, \lambda_2 = 10^{-6}, \lambda_3 = 5 \times 10^{-4}, \beta = 10^{-6}$ for *IndianP*,

TABLE III

CLASS NAME AND THE AMOUNT OF LABELED SAMPLES IN *Salinas* (TEN PERCENT OF LABELLED SAMPLES PER CLASS ARE USED FOR TRAINING).

| No. | Class name | Train | Test | Total |
|-------|---------------------------|-------|-------|-------|
| 1 | Broccoli-green-weeds-1 | 201 | 1808 | 2009 |
| 2 | Broccoli-green-weeds-2 | 373 | 3353 | 3726 |
| 3 | Fallow | 198 | 1778 | 1976 |
| 4 | Fallow-rough-plow | 140 | 1254 | 1394 |
| 5 | Fallow-smooth | 268 | 2410 | 2678 |
| 6 | Stubble | 396 | 3563 | 3959 |
| 7 | Celery | 358 | 3221 | 3579 |
| 8 | Grapes-untrained | 1128 | 10143 | 11271 |
| 9 | Soil-vinyard-develop | 621 | 5582 | 6203 |
| 10 | Corn-senesced-green-weeds | 328 | 2950 | 3278 |
| 11 | Lettuce-romaine-4wk | 107 | 961 | 1068 |
| 12 | Lettuce-romaine-5wk | 193 | 1734 | 1927 |
| 13 | Lettuce-romaine-6wk | 92 | 824 | 916 |
| 14 | Lettuce-romaine-7wk | 107 | 963 | 1070 |
| 15 | Vinyard-untrained | 727 | 6541 | 7268 |
| 16 | Vinyard-vertical-trellis | 1181 | 1626 | 1807 |
| Total | | 5418 | 48711 | 54129 |

$\lambda_1 = 10^{-5}, \lambda_2 = 0.1, \lambda_3 = 10^{-4}, \beta = 10^{-6}$ for *PaviaU* and $\lambda_1 = 10^{-6}, \lambda_2 = 10^{-3}, \lambda_3 = 10^{-3}, \beta = 10^{-6}$ for *Salinas*. Figs. 5-7 reveal the following important observations.

1) Our method STSC outperforms other methods in most cases in terms of OA, AA and κ for both SVM and KNN classifiers. The diagrams show that the performance evolutions of SVM in terms of OA, AA and κ are in consistence with that of KNN for all the methods, but SVM always yields higher accuracies and κ compared to KNN. Generally, for all the methods the classification performances of the two classifiers increase when more selected band are utilized. This is intuitively reasonable as the spectral signatures of ground objects often become more distinguishable while more spectral bands are adopted. Rare cases are also observed such as OCF in the dataset *IndianP* (Fig. 5) where the OA decreases from 75.2% to 71.8% with the increasing amount of bands from $c = 30$ to $c = 35$. This can be attributed to the suboptimal bands clustering, resulting in the highly redundant selected bands. In Figs. 5-7, we can observe that for most of the methods the performance increases dramatically at the beginning, i.e., $c \leq 20$, and then saturates to a stable point when the number of bands is sufficiently large. Noticeably, the saturated points normally locate the areas where $c \ll B$ and often reach comparable or even better performances compared to that using all bands. Our method outperforms the ‘‘ALL bands’’ in terms of OA in SVM when $c \geq 25$ on the dataset *IndianP*, $c \geq 30$ on the dataset *PaviaU* and $c \geq 35$ on the dataset *Salinas*. This demonstrates the effectiveness of our band selection method.

2) The ranking-based method MVPCA in most cases perform worse than others, especially when only a few bands are selected such as $c < 25$. MVPCA selects representative bands based on the ranking score from high to low in the criterion of variance of bands, which results in the selection of bands with larger variance in particular. It is true that bands with large variance sometimes are more informative than those with smaller variance. However, noisy bands normally show

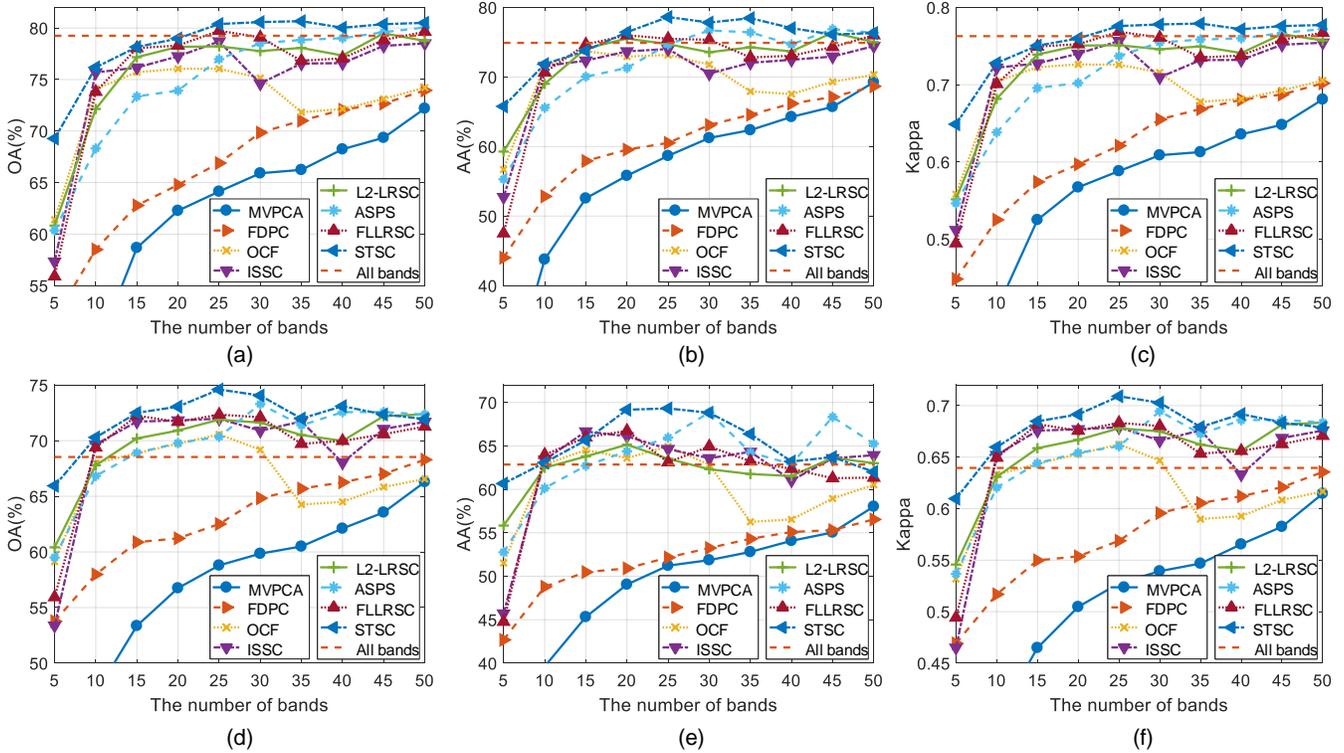


Fig. 5. The influence of band selection on the classification performance in classifiers SVM and KNN on the dataset *IndianP*. (a) OA in SVM, (b) AA in SVM, (c) κ in SVM, (d) OA in KNN, (e) AA in KNN and (f) κ in KNN.

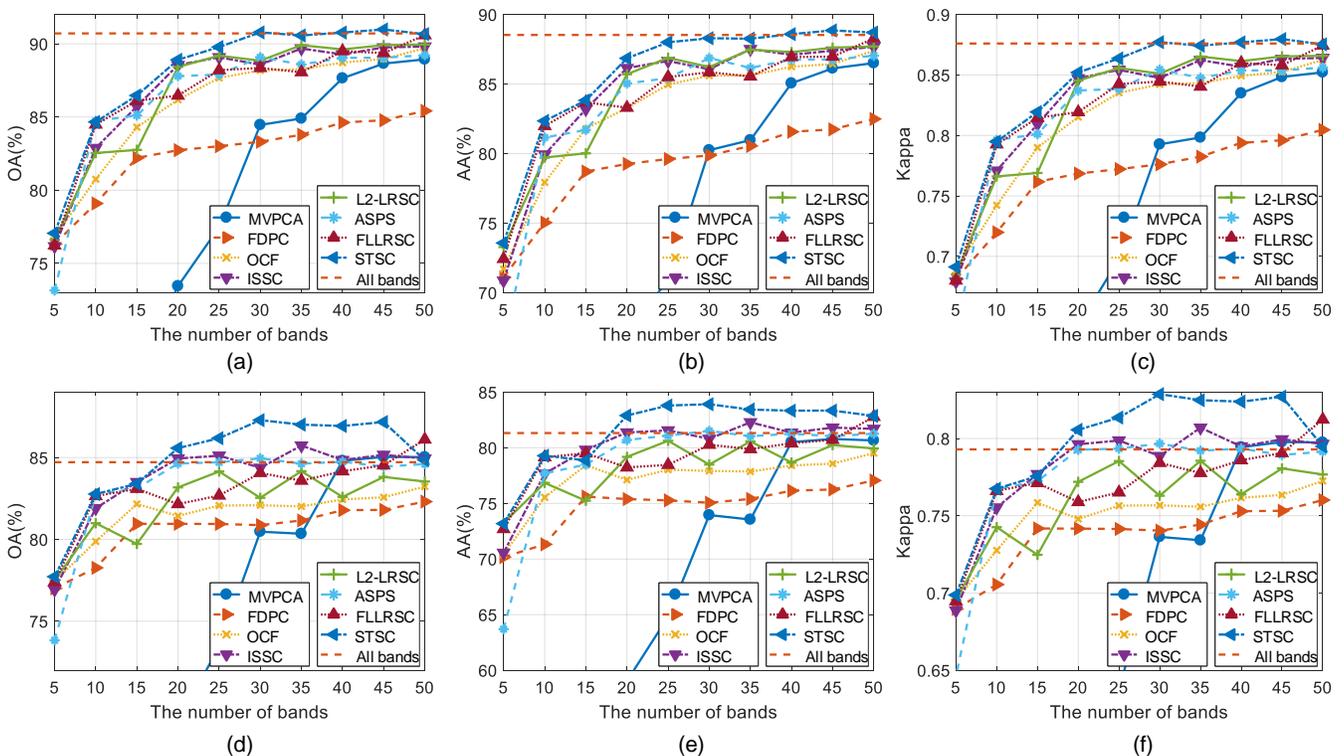


Fig. 6. The influence of band selection on the classification performance in classifiers SVM and KNN on the dataset *PaviaU*. (a) OA in SVM, (b) AA in SVM, (c) κ in SVM, (d) OA in KNN, (e) AA in KNN and (f) κ in KNN.

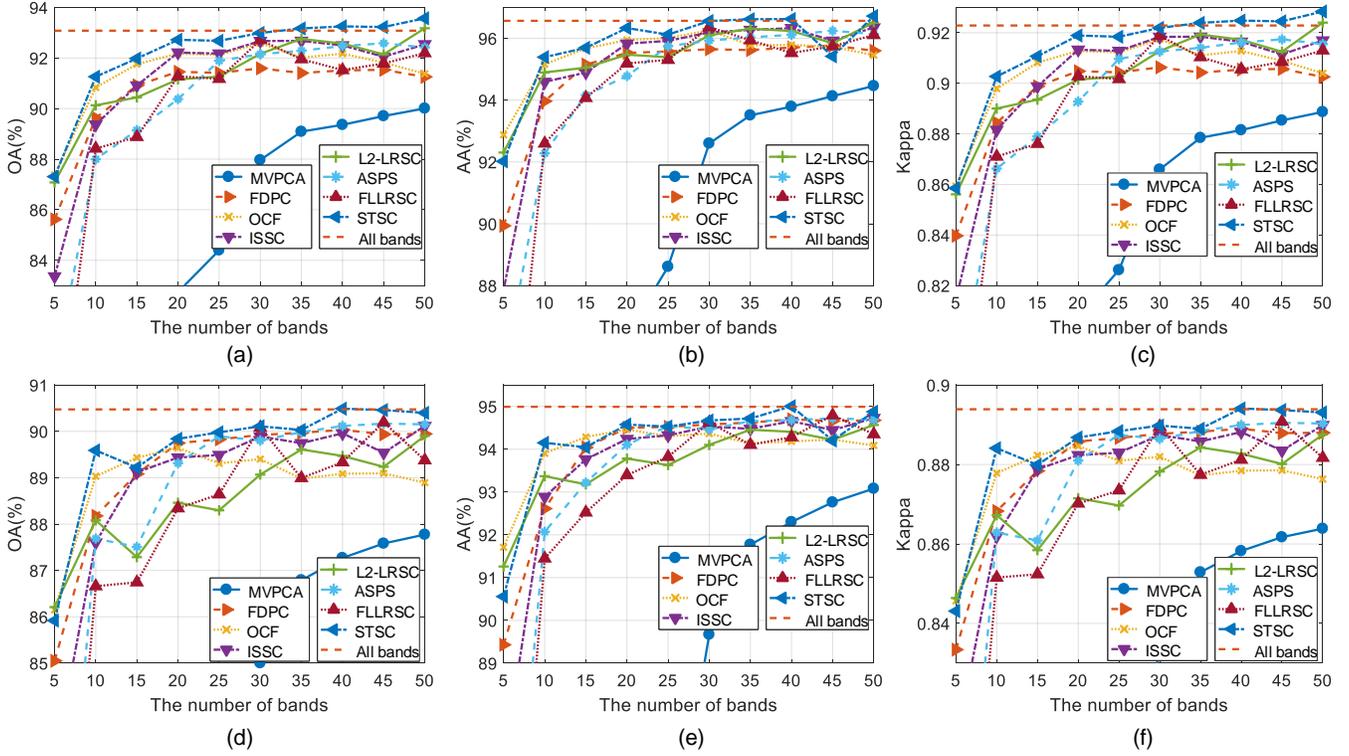


Fig. 7. The influence of band selection on the classification performance in classifiers SVM and KNN on the dataset *Salinas*. (a) OA in SVM, (b) AA in SVM, (c) κ in SVM, (d) OA in KNN, (e) AA in KNN and (f) κ in KNN.

big variance and might be viewed as representative bands based on this ranking criterion, which results in the sensitive performance of MVPCA to noise. Moreover, this method neglects the correlations of bands and thus leads to the selected bands to be highly redundant. This deteriorates thereby the classification performance of classifiers significantly especially when c is very small. For instance, when $c = 5$, the OAs of SVM are 31.7%, 37.8% and 52.2% on the datasets *IndianP*, *PaviaU* and *Salinas*, respectively, which are very poor compared to other methods. FDPC is a hybrid-based band selection method which integrates a density-based clustering algorithm with ranking approach. Results show that FDPC outperforms MVPCA in most cases of c in the three datasets, which benefits from the decorrelations of bands by measuring the local density and intracluster distance of bands. Compared with MVPCA, it is observed that clustering-based methods including OCF, ASPS, ISSC, L2-LRSC, FLLRSC and STSC perform better mostly in terms of OA, AA and κ in the three datasets.

3) Subspace clustering based methods, i.e., ISSC, L2-LRSC, FLLRSC and STSC, outperform another clustering based method OCF most of the time. OCF learns the cluster structure of bands by optimizing an objective function under the assumption that bands in the same cluster have contiguous wavelengths. This results in a contiguous bands division, which is conflict with the fact that non-adjacent bands can be in the same cluster. Consequently, the bands from different clusters might be of high similarity, leading to the high redundancy of the selected bands. This deteriorates the

classification performance of classifiers as shown in Fig. 5 where the performance decreases in terms of OA and κ when $c \geq 35$. Subspace clustering based methods recovery the cluster structure by optimizing representation-based problems with the core idea that each band can be represented as a linear combination of others in the same subspace to the query band. By incorporating prior information of the data with effective constraints, subspace clustering based methods offer a better capability to uncover the cluster structure of data in a more efficient and flexible way than OCF. ASPS segments the spectral bands into several continuous groups with a coarse-to-fine strategy and selects high-quality bands containing minimum noise in each refined group as representatives. Generally, it performs comparable to the subspace clustering based methods when a large number of bands are selected (for example $c = 50$). However, its performance is poor when c is small as shown in Figs. 6 and 7.

4) Among the subspace clustering based methods, STSC achieves the best results in terms of OA almost for each c in the SVM and KNN classifiers on the three datasets. The improvements can be mainly attributed to the incorporation of the local and global information in the data, which provides a comprehensive analysis of spectral bands and enables our model to learn better the cluster structure. Other methods such as ISSC, L2-LRSC and FLLRSC make use of the local and global prior information partially and thus yield sub-optimal band selection solutions, resulting thereby in worse classification performances in the classifiers. The results in Fig. 5 show that the performance improvement of STSC

TABLE IV
QUANTITATIVE EVALUATION MEASURED BY CLASSIFICATION FOR DIFFERENT BAND SELECTION METHODS (κ IN %)

| Datasets | Evaluations | MVPCA | FDPC | SBSS | OCF | ASPS | ISSC | L2-LRSC | FLLRSC | STSC | ALL |
|----------------|------------------|------------|------------|-------------------|------------|-------------------|------------|------------|-------------------|-------------------|------------|
| <i>IndianP</i> | SVM (OA) | 65.90±0.79 | 69.85±0.50 | 77.16±0.68 | 75.15±0.50 | 78.53±0.38 | 74.62±0.56 | 77.76±0.52 | 79.09±0.48 | 80.58±0.51 | 79.25±0.34 |
| | SVM (AA) | 61.23±1.61 | 63.09±1.54 | 73.16±1.50 | 71.78±1.46 | <u>76.69±1.23</u> | 70.48±1.50 | 73.56±1.13 | 75.34±2.09 | 77.82±0.89 | 74.91±1.65 |
| | SVM (κ) | 60.87±0.86 | 65.51±0.59 | 73.90±0.77 | 71.58±0.59 | 75.47±0.43 | 70.95±0.64 | 74.59±0.57 | 76.11±0.54 | 77.80±0.55 | 76.28±0.40 |
| | KNN (OA) | 59.85±0.81 | 64.83±0.69 | 70.15±0.74 | 69.19±0.50 | <u>73.29±0.56</u> | 70.89±0.62 | 71.62±0.58 | 72.11±0.44 | 74.06±0.65 | 68.53±0.63 |
| | KNN (AA) | 51.86±1.17 | 53.25±1.01 | 61.65±1.66 | 62.83±1.80 | 68.84±2.41 | 63.57±2.15 | 62.31±2.49 | 64.93±2.39 | <u>68.81±2.16</u> | 62.85±1.81 |
| | KNN (κ) | 53.93±0.90 | 59.56±0.81 | 65.79±0.84 | 64.67±0.58 | <u>69.46±0.63</u> | 66.61±0.67 | 67.50±0.65 | 68.02±0.48 | 70.28±0.74 | 63.95±0.72 |
| <i>PaviaU</i> | SVM (OA) | 84.46±0.19 | 83.31±0.21 | <u>89.71±0.16</u> | 88.18±0.35 | 89.09±0.25 | 88.58±0.19 | 88.85±0.24 | 88.35±0.17 | 90.78±0.20 | 90.71±0.15 |
| | SVM (AA) | 80.25±0.24 | 79.87±0.44 | <u>87.55±0.29</u> | 85.61±0.38 | 86.87±0.36 | 86.06±0.29 | 86.23±0.33 | 85.85±0.40 | 88.29±0.24 | 88.53±0.33 |
| | SVM (κ) | 79.27±0.25 | 77.62±0.29 | <u>86.28±0.21</u> | 84.22±0.47 | 85.44±0.34 | 84.75±0.25 | 85.11±0.32 | 84.46±0.23 | 87.71±0.27 | 87.60±0.20 |
| | KNN (OA) | 84.46±0.26 | 80.87±0.26 | <u>85.14±0.27</u> | 82.08±0.24 | 84.96±0.27 | 84.38±0.21 | 82.54±0.34 | 84.06±0.29 | 87.29±0.16 | 84.72±0.23 |
| | KNN (AA) | 73.95±0.52 | 75.06±0.48 | <u>81.83±0.31</u> | 77.92±0.43 | 81.52±0.39 | 80.81±0.33 | 78.49±0.52 | 80.27±0.38 | 83.91±0.34 | 81.32±0.27 |
| | KNN (κ) | 73.63±0.36 | 74.04±0.35 | <u>79.89±0.37</u> | 75.67±0.33 | 79.67±0.37 | 78.86±0.30 | 76.30±0.46 | 78.42±0.39 | 82.87±0.22 | 79.30±0.31 |
| <i>Salinas</i> | SVM (OA) | 87.96±0.26 | 91.59±0.19 | 92.11±0.13 | 92.59±0.17 | 92.15±0.19 | 92.67±0.10 | 92.17±0.20 | 92.71±0.14 | 92.99±0.17 | 93.08±0.20 |
| | SVM (AA) | 92.60±0.27 | 95.65±0.10 | 95.91±0.10 | 96.26±0.15 | 95.93±0.16 | 96.19±0.08 | 96.10±0.12 | <u>96.36±0.10</u> | 96.56±0.13 | 96.57±0.15 |
| | SVM (κ) | 86.59±0.29 | 90.63±0.21 | 91.21±0.15 | 91.74±0.19 | 91.26±0.21 | 91.84±0.11 | 91.27±0.22 | <u>91.88±0.16</u> | 92.19±0.19 | 92.29±0.22 |
| | KNN (OA) | 85.00±0.20 | 89.92±0.15 | 89.97±0.18 | 89.40±0.13 | 89.80±0.22 | 89.89±0.15 | 89.07±0.15 | <u>90.08±0.13</u> | 90.11±0.13 | 90.47±0.15 |
| | KNN (AA) | 89.67±0.22 | 94.58±0.10 | <u>94.63±0.14</u> | 94.36±0.11 | 94.46±0.18 | 94.53±0.11 | 94.10±0.16 | <u>94.62±0.11</u> | 94.67±0.10 | 94.99±0.12 |
| | KNN (κ) | 83.29±0.22 | 88.77±0.17 | 88.83±0.20 | 88.20±0.14 | 88.64±0.24 | 88.74±0.17 | 87.82±0.17 | 88.96±0.14 | 88.98±0.14 | 89.39±0.17 |

is more significant when c is smaller. For instance, when $c = 50$, the OA improvements of STSC in SVM are below 2% compared to ISSC, L2-LRSC and FLLRSC. While for the case of $c = 5$, the OA improvements are 12% for ISSC, 8% for L2-LRSC and 13% for FLLRSC. We can observe the similar results for the KNN classifier.

We report the results from Figs. 5-7 in the case of $c = 30$ in Table IV for a better presentation and discussion. The selected bands are shown in Tables V-VII. The results of SBSS are included where the number of selected bands is 60 on the dataset *IndianP*, 34 on the dataset *PaviaU* and 62 on the dataset *Salinas*. We also show the results of “ALL” that employs all the spectral bands for reference in Table IV. The best results (“ALL” is excluded) are marked in bold and suboptimal results are underlined. The results in Table IV show that STSC performs the best in most cases in terms of OA, AA and κ in the SVM and KNN classifiers. Compared with the results using all the bands, our method yields higher accuracies on the datasets *IndianP* and *Salinas* and comparable results on the dataset *Salinas* with 30 bands. This reveals the highly redundant information in HSIs and also demonstrates the advantages of band selection. Ranking-based method MVPCA produces the lowest accuracy and κ in *IndianP* and *Salinas*, which can be attributed to the high redundancy of selected bands as evidenced by the selected adjacent bands in Tables V and VII. FDPC also selects a number of adjacent bands, resulting in high redundancy of data and thereby a low classification accuracy in Table IV. OCF and ASPs perform analogously on the three datasets and fall behind mostly the subspace clustering based methods. Compared with ISSC, L2-LRSC and FLLRSC take into account respectively the local correlations of bands and hidden features in HSI, which results in often improved performances than ISSC as shown in Table IV. Compared with ISSC and L2-LRSC, our model is formulated in a more general model. The proposed structural regularization incorporates comprehensively the local and global property of data in the low-dimensional subspaces, which enables a better uncovering of the cluster structure. The results in Tables V-VII show that compared with other methods, the selected bands of STSC contain less neighbouring bands which are often

redundant, resulting in a better classification performance. SBSS is a parameter-free band selection method and views saliency bands as representative bands based on local extrema points of the Hessian matrix of HSI. SBSS selects 60, 34 and 62 bands as representative bands in *IndianP*, *PaviaU* and *Salinas*, respectively. In Table IV, we can see that our method even in the case of $c = 30$ still outperforms SBSS in terms of OA, AA and κ in the classifiers SVM and KNN.

C. The Effect of Different Training Data Sizes

We vary the amount of training data per class, p , in SVM and KNN within the set {1%, 5%, 10%, 20%, 30%, 40%, 50%} and report the results in Fig. 8 in the case of $c = 30$. The results show that our method STSC consistently outperforms all the reference methods on the three datasets regardless of the value of p . Most of the OA curves in each dataset evolve nearly in parallel, and we can conclude the same remarks present in Section IV-B. Unsurprisingly, all the classification performances in Fig. 8 get enhanced when using more training samples. Specifically, the OAs of all the methods increase rapidly at the beginning when $1\% \leq p \leq 10\%$ and then climb at a lower speed when $p > 10\%$. This experiment validates that the training data size in classifiers has little influence on the comparisons of band selection methods.

D. Parameter Study

To analyse the effect of the parameters in our model on the classification performance, we conduct experiments on the three datasets with SVM classifier by varying the values of $\lambda_1, \lambda_2, \lambda_3$ and β . In the experiments, we analyse each parameter by fixing others according to the parameter settings in Section IV-B. The amount of training data per class is set to 10% and $c = 30$. The results in Fig. 9 show that the performance of our method is quite stable in a certain range with respect to each parameter. Most of the results achieve comparable performance to the case using all the bands on the three datasets where the accuracies are 79.25, 90.71 and 93.08 in *IndianP*, *PaviaU* and *Salinas*, respectively.

TABLE V
SELECTED BANDS OF DIFFERENT METHODS ($c = 30$) IN *IndianP*

| Methods | Selected bands |
|---------|--|
| MVPCA | 13 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 39 41 42 43 44 45 48 49 50 51 52 |
| FDPC | 105 106 107 108 118 124 125 128 130 137 144 145 146 150 153 155 157 159 163 173 175 177 178 181 183 187 189 192 195 196 |
| SBSS | 2 7 9 11 14 16 19 22 25 32 35 39 42 44 48 50 52 54 56 62 66 68 72 74 79 83 85 90 94 96 105 109 115 118 122 125 129 131 133 135 |
| OCF | 138 142 146 149 154 158 161 164 167 169 173 175 177 179 181 183 188 192 194 198 |
| ASPS | 3 8 9 16 28 29 43 46 49 50 57 58 61 67 72 77 107 115 117 118 128 157 159 163 173 181 187 189 192 200 |
| ISSC | 8 10 11 15 27 36 47 49 58 70 72 78 81 94 102 105 116 119 128 132 143 146 151 159 165 175 184 190 195 199 |
| L2-LRSC | 16 21 33 39 52 57 64 70 74 78 80 84 86 89 90 100 104 111 116 121 127 135 141 147 154 163 170 180 186 193 |
| FLLRSC | 3 8 14 23 28 30 31 34 35 36 39 42 45 51 58 63 65 68 83 87 90 94 99 107 108 120 129 137 143 174 |
| STSC | 9 12 18 20 23 25 28 35 41 47 71 74 82 86 90 97 98 99 107 112 123 131 133 137 152 159 160 163 184 185 |
| STSC | 5 8 23 36 39 44 51 64 70 73 76 82 84 95 101 106 111 117 122 127 130 137 144 150 157 164 174 182 188 194 |

TABLE VI
SELECTED BANDS OF DIFFERENT METHODS ($c = 30$) IN *PaviaU*

| Methods | Selected bands |
|---------|---|
| MVPCA | 61 62 63 64 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 |
| FDPC | 15 19 29 30 31 33 34 35 36 37 38 39 40 41 42 43 44 48 52 53 55 56 58 61 88 90 92 93 94 99 |
| SBSS | 4 6 8 12 18 21 23 25 27 31 34 37 39 41 43 45 49 52 57 59 65 68 70 72 75 77 82 84 88 91 93 97 99 101 |
| OCF | 1 2 3 6 10 15 19 22 27 29 33 36 41 44 48 52 53 61 65 66 68 70 73 77 79 82 88 92 99 101 |
| ASPS | 5 9 11 15 17 20 25 27 30 34 37 39 43 50 51 53 58 59 60 69 72 74 80 84 86 90 92 95 98 101 |
| ISSC | 3 8 14 19 22 23 27 30 31 33 34 35 40 43 45 50 53 54 58 63 69 74 77 79 81 84 86 91 94 100 |
| L2-LRSC | 1 2 3 4 5 8 10 12 14 17 21 25 30 34 39 42 48 55 61 65 69 73 76 79 83 85 88 94 99 102 |
| FLLRSC | 3 4 6 7 8 9 10 11 12 13 14 15 16 19 23 29 33 39 49 54 59 63 68 76 83 85 86 90 100 103 |
| STSC | 7 12 15 19 22 25 29 31 34 36 38 41 45 49 51 56 59 63 67 69 73 76 79 83 84 85 86 89 92 95 |

TABLE VII
SELECTED BANDS OF DIFFERENT METHODS ($c = 30$) IN *Salinas*

| Methods | Selected bands |
|---------|---|
| MVPCA | 28 29 30 31 32 33 34 35 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 72 |
| FDPC | 11 20 24 32 55 68 75 82 88 89 93 94 95 96 113 121 126 128 131 141 148 161 166 177 181 185 190 192 196 199 |
| SBSS | 2 6 10 12 15 19 22 25 28 31 33 37 42 45 47 52 55 58 60 62 67 69 71 73 76 82 86 88 93 97 100 107 110 112 114 121 124 126 128 133 |
| OCF | 136 138 143 146 150 154 158 160 165 167 170 174 177 181 185 187 190 192 194 197 200 203 |
| ASPS | 5 7 11 14 20 24 32 34 37 38 39 40 42 46 50 55 57 62 64 68 75 79 88 113 121 128 131 135 137 166 |
| ISSC | 10 13 18 26 29 41 47 51 53 71 73 81 85 94 106 110 113 116 129 139 146 148 155 161 170 177 182 184 194 203 |
| L2-LRSC | 5 15 19 25 30 42 48 54 58 68 73 80 83 89 94 100 108 114 120 125 133 143 154 159 165 171 177 186 194 201 |
| FLLRSC | 3 4 7 9 13 16 20 24 31 33 40 44 48 53 58 65 70 77 80 86 99 104 111 124 128 139 151 162 181 195 |
| STSC | 6 26 40 41 53 57 68 76 81 84 86 90 91 95 99 114 116 124 125 126 135 141 142 156 158 165 183 195 198 201 |
| STSC | 6 12 21 27 31 37 40 43 47 51 58 63 68 76 81 88 97 111 122 134 144 154 159 164 170 177 183 191 197 201 |

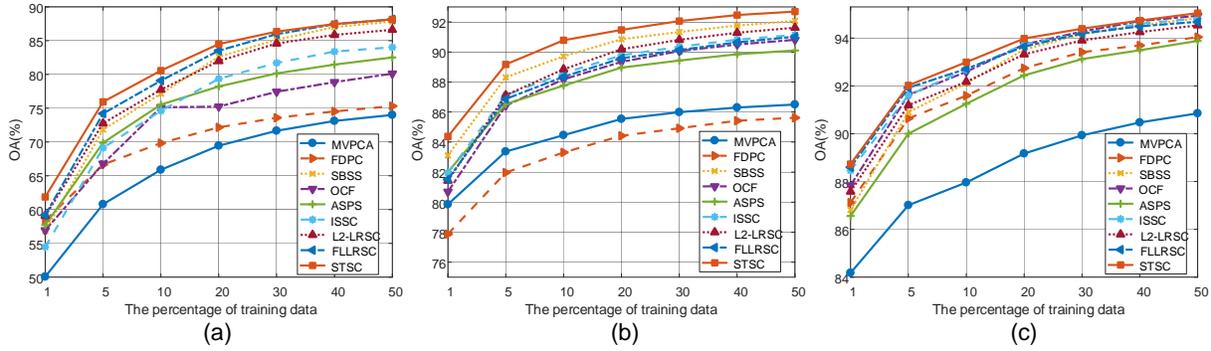


Fig. 8. Influence of training data size on the overall accuracy of SVM on the datasets: (a) *Indian Pines*, (b) *Pavia University* and (c) *Salinas*.

λ_1 and λ_2 control the local regularizations in our model. The results in *IndianP* in Fig. 9 (a) show that a larger λ_1 yields a higher OA. For parameter λ_2 , we also observe the similar results in *PaviaU* in Fig. 9 (b). This validates the importance of the incorporation of local regularizations in our model. We shall observe that the OA decreases slightly with larger λ_1 and λ_2 on the dataset *PaviaU* in Fig. 9 (a) and on the dataset *IndianP* in Fig. 9 (b). This is mainly caused by the too strong constraints on the local structure, which results in an

unbalanced learning result with respect to the local and global property of data. Fig. 9 (c) indicates that the OA is very stable when $\lambda_3 \leq 10^{-2}$ but drops significantly when $\lambda_3 > 10^{-2}$, showing the significance of the global regularization in our model. The parameter β controls the error between coefficients matrices \mathbf{A} and \mathbf{L} , and often yields a better performance with a relatively smaller value. When β is sufficiently large such as $\beta \geq 0.1$, the error matrix \mathbf{S} becomes a zeros matrix on the three datasets. In this case, our model reduces to a special

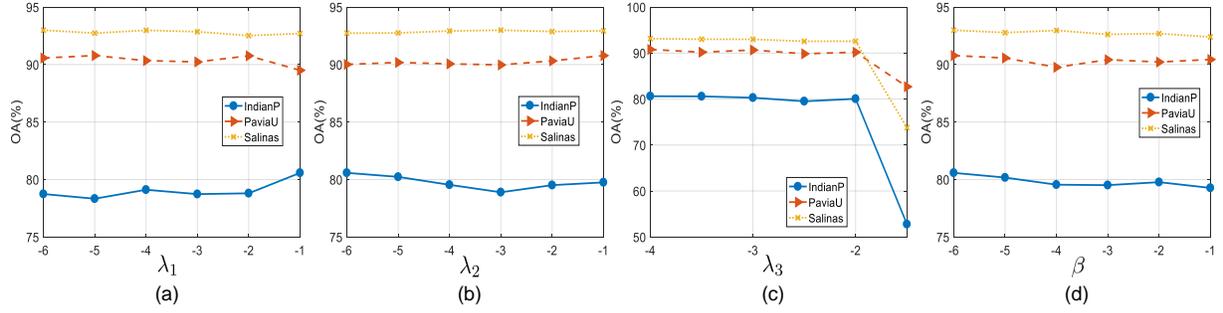


Fig. 9. Performance analysis in terms of OA by SVM with respect to parameters $\lambda_1, \lambda_2, \lambda_3$ and β on the three datasets (x-axis is in \log_{10}).

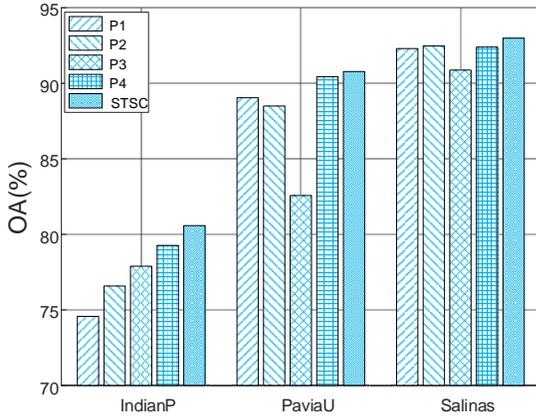


Fig. 10. Ablation study: OA in SVM for different variants of STSC.

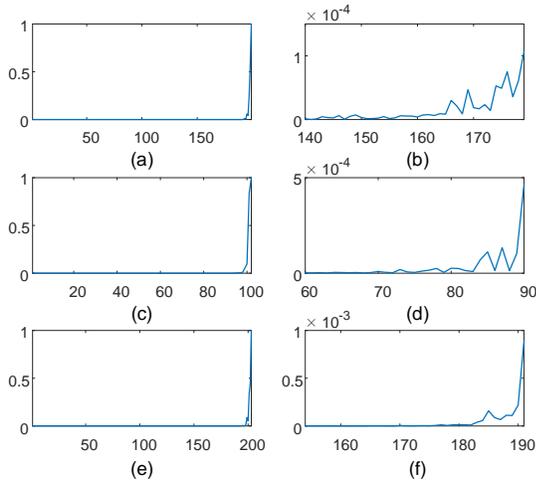


Fig. 11. The normalized eigenvalue gaps d_{n_i} on the datasets *IndianP* (top), *Pavia* (middle) and *Salinas* (bottom). The figures in the right are the enlarged details corresponding to the results in the left.

case where the regularization $\|\mathbf{S}\|_1$ in (6) is eliminated and the constraint $\mathbf{A} = \mathbf{L} + \mathbf{S}$ reduces to $\mathbf{A} = \mathbf{L}$, which allows the structural regularization $\Psi(\mathbf{L})$ functions equally to \mathbf{A} . We observe that this yields inferior performance in terms of OA compared to our original model, which demonstrates the efficacy of our decomposition strategy.

E. Ablation Study

We conduct ablation studies by removing one of the regularizations in (6), and denote by $P1, P2, P3$ and $P4$ the variants of the proposed model when $\|\nabla_x \mathbf{L}\|_{2,1}$, $\|\nabla_y \mathbf{L}\|_1$, $\|\mathbf{L}\|_*$ and $\|\mathbf{S}\|_1$ are removed, respectively. Note that we have $\mathbf{A} = \mathbf{L}$ when $\|\mathbf{S}\|_1$ is removed. We set $c = 30$ in the experiments. The results shown in Fig. 10 indicate that our STSC achieves consistently the best OA in SVM, demonstrating the effectiveness of each regularization. Compared with STSC, the variants, which produce the most severe OA drop, are $P1$ in *IndianP*, $P3$ in *PaviaU* and $P3$ in *Salinas*. This verifies the importance of incorporating local and global regularizations. It is also observed that when the term $\|\mathbf{S}\|_1$ is removed, the resulting performance of $P4$ is worse than the proposed STSC model, which is consistent with the previous conclusion in Section IV-D.

F. Determination of the Number of Selected Bands

We plot the normalized eigenvalue gaps d_{n_i} in Fig. 11 for the three datasets. The complete curves of d_{n_i} are shown in the left and the right are the corresponding details. We can see from Figs. 11 (a), (c) and (e) that all the curves originate at a wide flattened region, where d_{n_i} are close to zero, and then increase dramatically to 1. Despite some small fluctuations, all the curves grow gradually overall. According to (7), the number of representative bands is associated with the smallest index k that satisfies $d_{n_k} \geq \delta$. All the d_{n_i} are in the range of $[0,1]$, which allows us to determine the number of bands adaptively regardless of the type of HSIs. We set δ to 5×10^{-5} for our method and obtain $n = 27$ for *IndianP*, $n = 20$ for *PaviaU* and $n = 21$ for *Salinas*, by which our approach yields excellent classification performance that is comparable to the case using all the bands as shown in Figs. 5-7. Moreover, in Figs. 5-7 we can see that these estimated numbers are very close to the saturated points that achieve stable classification performances, which validates the effectiveness of our approach.

G. The Influence of Noise

We study the influence of noise on the performance of different band selection methods in *IndianP*. The number of bands is set to 30 and the SVM classifier is used. We consider three types of noise: Gaussian noise, impulse noise

TABLE VIII
THE OAS IN SVM FOR BAND SELECTION METHODS IN *Indian Pines* WITH DIFFERENT NOISE

| Data | MVPCA | FDPC | OCF | ASPS | ISSC | L2-LRSC | FLLRSC | STSC | ALL |
|----------------|------------|------------|------------|-------------------|------------|------------|------------|-------------------|------------|
| Original data | 65.90±0.79 | 69.85±0.50 | 75.15±0.50 | 78.53±0.38 | 74.62±0.56 | 77.76±0.52 | 79.09±0.48 | 80.58±0.51 | 79.25±0.34 |
| Gaussian noise | 48.23±0.49 | 48.09±0.66 | 62.39±0.66 | 62.34±0.64 | 65.28±0.54 | 65.24±0.54 | 63.30±0.40 | 67.36±0.54 | 72.85±0.47 |
| Impulse noise | 41.02±0.88 | 70.37±0.43 | 43.49±0.65 | 76.24±0.39 | 66.19±0.64 | 72.25±0.52 | 69.99±0.53 | 77.49±0.46 | 69.09±0.57 |
| Mixed noise | 40.38±0.44 | 50.80±0.60 | 35.56±0.64 | 64.55±0.74 | 59.06±0.55 | 58.92±0.89 | 58.00±0.57 | 64.04±0.60 | 66.36±0.73 |

TABLE IX
RUNNING TIME (IN SECONDS) FOR DIFFERENT BAND SELECTION METHODS

| Date sets | MVPCA | FDPC | SBSS | OCF | ASPS | ISSC | L2-LRSC | FLLRSC | STSC |
|----------------|-------------|-------------|------|------|------|-------------|---------|--------|------|
| <i>IndianP</i> | 0.06 | 0.07 | 1.64 | 1.03 | 1.05 | 0.06 | 0.87 | 20.51 | 1.06 |
| <i>PaviaU</i> | 0.57 | 0.20 | 3.77 | 0.76 | 0.46 | 0.08 | 0.63 | 18.86 | 0.53 |
| <i>Salinas</i> | 0.58 | 0.24 | 4.25 | 1.40 | 0.94 | 0.10 | 1.15 | 20.27 | 1.13 |

and a mixture of the former two. We add Gaussian noise such that signal-to-noise ratio (SNR) varies between 20 and 30 dB in each band, and we introduce impulse noise with 30% of corrupted pixels in bands 30-40 and 150-160. The results in Table VIII show that the accuracies of all methods decrease after adding the noise and the mixed noise tends to deteriorate the performance most severely except for the methods FDPC and ASPS. Our STSC method yields the highest OA in the cases after adding Gaussian noise and impulse noise. For the case of mixed noise, the performance of STSC is comparable to the best result of ASPS. We also show the classification results of “ALL” in Table VIII, which uses all the bands. We observe that for the method “ALL”, impulse noise results in a more significant OA drop than Gaussian noise compared with the results in the original data, which demonstrates that impulse noise has a more severe impact on the classification performance than Gaussian noise. It is noted that in the case of impulse noise our method outperforms significantly “ALL” by 8.4%, demonstrating the robustness of STSC to impulse noise. Other methods FDPC, ASPS, L2-LRSC and FLLRSC yield also higher accuracies than “ALL” in the case of impulse noise. A common idea among STSC, ASPS and L2-LRSC is that they consider the correlation of neighbouring bands in the band selection model, which facilitates a robust performance to impulse noise. ISSC does not consider such local correlation of bands, which in fact treats each band independently in the optimization process, yielding thereby a worse accuracy than STSC, ASPS and L2-LRSC in the case of impulse noise.

H. Running Time

We show in Table IX the running time of different band selection methods on the three data sets. The number of selected bands is set to 30. All the methods were implemented in MATLAB R2016b on a laptop with an Intel[®] core-i7 6700HQ CPU with 16GB of RAM. The results show that ISSC is the fastest method, which can be mainly attributed to the closed-form solution. MVPCA, FDPC and ASPS are often more efficient than the representation-based methods L2-LRSC, FLLRSC and STSC. Due to the expensive computation of Hessian matrix, SBSS uses more running time in large data sets *PaviaU* and *Salinas* than *IndianP*. FLLRSC takes the longest running time because of the high computational cost

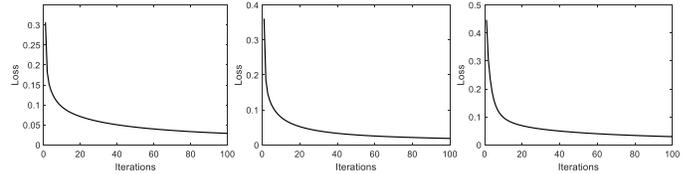


Fig. 12. The convergence curves of STSC in *IndianP* (left), *PaviaU* (middle) and *Salinas* (right).

from multiple SVDs. Compared with the benchmark methods, the running speed of the proposed method STSC is reasonable.

I. Convergence Study

We show in Fig. 12 the convergence curves of Algorithm 1 in three data sets. The loss of objective function is calculated by $L = \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{A}\|_F^2 + \lambda_1\|\nabla_x\mathbf{V}_1\|_{2,1} + \lambda_2\|\nabla_y\mathbf{V}_2\|_1 + \lambda_3\|\mathbf{V}_3\|_* + \beta\|\mathbf{S}\|_1$. The results reveal that the loss of our objective function monotonically decreases to a stable level. The loss drops sharply at the first several iterations and then tends to saturate at following iterations in all three data sets, demonstrating a stable convergence property of our algorithm.

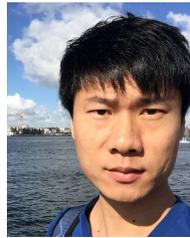
V. CONCLUSION

In this paper, we present a new unsupervised band selection method based on structural subspace clustering. Compared with the traditional subspace clustering model, we formulate a more general model where the coarse coefficients matrix, derived from the self-representation, is decomposed as a summation of a desirable coefficients matrix and a sparse matrix. To guide the learning of the desirable coefficients matrix, we take into account the prior information of bands in terms of local and global property, and propose a novel structural regularization. This allows a comprehensive analysis of the data in the lower-dimensional subspaces, facilitating thereby a better solution in subspace clustering model to uncover the intrinsic cluster structure of hyperspectral bands than the existing methods. Moreover, an adaptive approach with eigenvalue gap analysis on the Laplacian matrix is proposed to estimate the number of selected bands. We also develop an efficient algorithm based on the ADMM to solve the resulting optimization problem. Experiments on three benchmark datasets demonstrate that our method outperforms the state-of-the-art band selection methods.

REFERENCES

- [1] X. Bai, F. Xu, L. Zhou, Y. Xing, L. Bai, and J. Zhou, "Nonlocal similarity based nonnegative Tucker decomposition for hyperspectral image denoising," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 701–712, 2018.
- [2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, 1968.
- [3] M. D. Farrell and R. M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 192–195, 2005.
- [4] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Kernel principal component analysis for the classification of hyperspectral remote sensing data over urban areas," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, pp. 1–14, 2009.
- [5] Q. Du, "Modified Fisher's linear discriminant analysis for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 503–507, 2007.
- [6] Q. Wang, Z. Meng, and X. Li, "Locality adaptive discriminant analysis for spectral-spatial classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2077–2081, 2017.
- [7] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1185–1198, 2011.
- [8] Y.-J. Deng, H.-C. Li, L. Pan, L.-Y. Shao, Q. Du, and W. J. Emery, "Modified tensor locality preserving projection for dimensionality reduction of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 277–281, 2018.
- [9] A. Martínez-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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] L. Jiao, J. Feng, F. Liu, T. Sun, and X. Zhang, "Semisupervised affinity propagation based on normalized trivariable mutual information for hyperspectral band selection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2760–2773, 2014.
- [16] 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.
- [17] H. Su, B. Yong, and Q. Du, "Hyperspectral band selection using improved firefly algorithm," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 68–72, 2015.
- [18] X. Cao, C. Wei, Y. Ge, J. Feng, J. Zhao, and L. Jiao, "Semisupervised hyperspectral band selection based on dynamic classifier selection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1289–1298, 2019.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] S. Feng, Y. Itoh, M. Parente, and M. F. Duarte, "Hyperspectral band selection from statistical wavelet models," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2111–2123, 2017.
- [23] W. Sun and Q. Du, "Hyperspectral band selection: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 118–139, 2019.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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. 3906–3915, 2020.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Hyperspectral im-

- age clustering: Current achievements and future lines,” *IEEE Geosci. Remote Sens. Mag.*, 2020.
- [39] E. Elhamifar, G. Sapiro, and R. Vidal, “See all by looking at a few: Sparse modeling for finding representative objects,” in *Proc. IEEE CVPR*, 2012, pp. 1600–1607.
- [40] W. Sun, L. Zhang, L. Zhang, and Y. M. Lai, “A dissimilarity-weighted sparse self-representation method for band selection in hyperspectral imagery classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4374–4388, 2016.
- [41] 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.
- [42] 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. Remote Sens.*, vol. 12, no. 3, pp. 989–999, 2019.
- [43] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.
- [44] L. Zhou, X. Zhang, J. Wang, X. Bai, L. Tong, L. Zhang, J. Zhou, and E. Hancock, “Subspace structure regularized nonnegative matrix factorization for hyperspectral unmixing,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4257–4270, 2020.
- [45] 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.
- [46] L. Zhou, X. Bai, X. Liu, J. Zhou, and E. R. Hancock, “Learning binary code for fast nearest subspace search,” *Pattern Recognit.*, vol. 98, 2020.
- [47] H. Zhang, J. Cai, W. He, H. Shen, and L. Zhang, “Double low-rank matrix decomposition for hyperspectral image denoising and destriping,” *IEEE Trans. Geosci. Remote Sens.*, 2021.
- [48] H. Zhang, W. He, L. Zhang, H. Shen, and Q. Yuan, “Hyperspectral image restoration using low-rank matrix recovery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4729–4743, 2014.
- [49] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, “A convex formulation for hyperspectral image superresolution via subspace-based regularization,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, 2015.
- [50] 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.
- [51] M. A. Figueiredo and R. D. Nowak, “An EM algorithm for wavelet-based image restoration,” *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, 2003.
- [52] I. Daubechies, M. DeFrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, 2004.
- [53] P. L. Combettes and V. R. Wajs, “Signal recovery by proximal forward-backward splitting,” *Multiscale Model. Sim.*, vol. 4, no. 4, pp. 1168–1200, 2005.
- [54] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
- [55] N. Cristianini, J. Shawe-Taylor *et al.*, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.
- [56] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [57] P. Su, D. Liu, X. Li, and Z. Liu, “A saliency-based band selection approach for hyperspectral imagery inspired by scale selection,” *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 572–576, 2018.
- [58] 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.



Shaoguang Huang (Member, IEEE) received the M.S. degree in telecommunication and information system from Shandong University, Jinan, China, in 2015 and the Ph.D. degree in computer science engineering from Ghent University, Belgium, in 2019.

He is currently a Post-Doctoral Research Fellow with the Group for Artificial Intelligence and Sparse Modelling (GAIM), Ghent University, Belgium. His area of interests includes image processing, sparse representation, clustering, hyperspectral image analysis and machine learning.

Dr. Huang also serves as a Guest Editor for *Remote Sensing*. He is also a Reviewer of the international journals, including *IEEE Transactions on Geoscience and Remote Sensing*, *IEEE Transactions on Circuits and Systems for Video Technology*, *IEEE Transactions on Biomedical Circuits and Systems*, *Pattern Recognition*, *Remote Sensing*, *Computers & Geosciences*.



Hongyan Zhang (Senior Member, IEEE) received the B.S. degree in geographic information system and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, China, in 2005 and 2010, respectively.

He has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, since 2016. He is a Young Chang-Jiang Scholar appointed by the Ministry of Education of China.

He has authored/co-authored more than 100 research articles and eight patents. His research interests include image reconstruction for quality improvement, hyperspectral information processing and agricultural remote sensing.

Dr. Zhang scored first in the IEEE GRSS Data Fusion Contest in 2019 and 2021 organized by the IEEE Image Analysis and Data Fusion Technical Committee. He has served as the Session Chair of the 2016 IEEE IGARSS Conference and the 2015 IEEE WHISPERS Conference. He also serves as an Associate Editor of *Photogrammetric Engineering & Remote Sensing* and *Computers & Geosciences*. He is also a Reviewer of more than 30 international academic journals, including *IEEE Transactions on Geoscience and Remote Sensing*, *IEEE Transactions on Image Processing*, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *IEEE Geoscience and Remote Sensing Letters*.



Aleksandra Pižurica (Senior Member, IEEE) received the Diploma in electrical engineering from the University of Novi Sad, Serbia, in 1994, the Master of Science degree in telecommunications from the University of Belgrade, Serbia, in 1997, and the Ph.D. degree in engineering from Ghent University, Belgium, in 2002.

She is a Professor in statistical image modeling with Ghent University. Her research interests include the area of signal and image processing and machine learning, including multiresolution statistical image models, Markov Random Field models, sparse coding, representation learning, and image and video reconstruction, restoration, and analysis.

Prof. Pižurica received the scientific prize “de Boelpaep” for 2013–2014 awarded by the Royal Academy of Science, Letters and Fine Arts of Belgium for her contributions to statistical image modeling and applications to digital painting analysis. The work of her team has been awarded twice the Best Paper Award of the IEEE Geoscience and Remote Sensing Society Data Fusion contest, in 2013 and 2014. She has served as an Associate Editor for the *IEEE TRANSACTIONS ON IMAGE PROCESSING* (2012 – 2016), Senior Area Editor for the *IEEE TRANSACTIONS ON IMAGE PROCESSING* (2016 – 2019) and an Associate Editor for the *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY* (2016 – 2020). She was a Lead Guest Editor for the *EURASIP Journal on Advances in Signal Processing* Special Issue “Advanced Statistical Tools for Enhanced Quality Digital Imaging with Realistic Capture Models” in 2013.