Hyperspectral Band Selection via Optimal Neighborhood Reconstruction

Qi Wang, *Senior Member, IEEE,* Fahong Zhang, and Xuelong Li*, *Fellow, IEEE*

***Abstract*—Band selection is one of the most important technique in the reduction of hyperspectral image (HSI). Different from traditional feature selection problem, an important characteristic of it is that there is usually strong correlation between neighboring bands, i.e., bands with close indexes. Aiming to fully exploit this prior information, a novel band selection method called optimal neighborhood reconstruction (ONR) is proposed. In ONR, band selection is considered as a combinatorial optimization problem. It evaluates a band combination by assessing its ability to reconstruct the original data, and applies a noise reducer to minimize the influence of noisy bands. Instead of using some approximate algorithms, ONR exploits a recurrence relation underlies the optimization target to obtain the optimal solution in an efficient way. Besides, we develop a parameter selection approach to automatically determine the parameter of ONR, ensuring it is adaptable to different data sets. In experiments, ONR is compared with some state-of-the-art methods on six HSI data sets. The results demonstrate that ONR is more effective and robust than the others in most of the cases.**

***Index Terms*—Hyperspectral band selection, least square, dictionary learning, sparse representation.**

I. Introduction

Recently, hyperspectral image (HSI) has attracted wide attention due to its ability of providing more abundant visual information, which is unreachable for traditional RGB image. A hyperspectral image is a 3-D data cube that consists of hundreds of spectral bands, and each of them records the reflectance of the scene to a specific electromagnetic wave. Though HSI processing techniques have been successfully applied in various of applications such as medical imaging processing [1], production quality inspection [2], and environmental monitoring [3], there are still troubles cannot be well resolved caused by the high dimensionality of HSI, mainly existing in efficient data storage, transmission, processing and so forth. As an effective approach to handle the high dimensionality, HSI reduction is considered as an important research field in HSI processing.

Similar to traditional reduction techniques, HSI reduction can be categorized into feature exaction [4, 5] and feature selection [6–8] (also known as band selection). The former combines some bands to generate new features, while the latter just drops some redundant bands. Though feature exaction can well preserve the discriminative information of the entire HSI, it will destroy the physical meaning of HSI data and make the extracted features hard to be interpreted. Due to this reason, band selection is more preferable in many cases. Based on whether the labeled samples are utilized, band selection can be further divided into supervised [9], semi-supervised [10], and unsupervised [11–13] methods. For supervised and semi-supervised methods, the label information can benefit the feature evaluation process [14], and hence helps to achieve better performance. Nevertheless, since HSI data are always hard to be labeled, these two methods are not very practical in real applications. As a result, the unsupervised methods, which are not restricted by labeled samples, have broader prospects for applications.

As a specialization of feature selection, band selection problem has some unique characteristics, including the following ones.

- Some features in HSI are contaminated by noises. Owing to atmospheric condition, sensor noise and other factors [15], noises are often unavoidable. And in some cases, the noises do not evenly distribute into all the bands but accumulate on a few of them, forming the so called noisy bands. So how to deal with this kind of noises and minimize the negative effect becomes an important issue for HSI band selection.
- There is a *correlated neighborhood property* (CNP) among the features, i.e., neighboring bands usually have stronger correlation. With the increase of band number, the signal curves of the ground objects are getting more and more smooth, which means samples drawn from the same ground object have similar characteristic on two neighboring bands, and hence will increase the similarity or correlation between them. Previous works that have utilized this property include [16], where a hierarchy of groups of adjacent bands are built, and [17], where a piece-wise constant function is optimized to separate the whole bands into multiple intervals.

In this paper, we take into consideration the above characteristics and propose an optimal neighborhood reconstruction (ONR) algorithm to tackle the band selection problem. The contributions of it are listed as follows.

1) We model the relationship between the selected bands and the entire data set from the perspective of linear reconstruction. By exploiting the correlated neighborhood property and introducing a noise reduction mechanism, a better characterization and interpretation of HSI data can be achieved with

*X. Li is the corresponding author. This work was supported by the National Key R&D Program of China under Grant 2018YFB1107403, National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, 61871470, and 61761130079.

Q. Wang, F. Zhang and X. Li are with the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@gmail.com, zfh@mail.nwpu.edu.cn and li@nwpu.edu.cn).

respect to their spectral correlation and noise distribution.

2) We propose an efficient optimization approach to search for the optimal band subset. Different from the previous optimization techniques that are based on approximate algorithms, the proposed approach can achieve an exact solution in a more efficient way.

3) We label and publish a new HSI data set named Leaves for testing[1]. Compared to the current public data sets, Leaves has larger spatial size and higher spectral dimensionality.

II. Related Work

Some representative unsupervised band selection methods will be discussed in this part. According to the employed selection strategy, unsupervised band selection can be divided into group-wise and point-wise methods.

A. Group-wise Method

Group-wise methods first separate the bands into groups, and select one band in each group to form the band subset. For this kind of method, two important issues must be resolved. One is how to separate the bands, and another is how to select the representative bands in each group. For the first issue, various kinds of clustering algorithms have been adopted to separate the bands, such as hierarchical clustering [18], spectral clustering [19, 20], low-rank representation [21–23] and affinity propagation [24]. As for the second issue, some of the clustering algorithms [24] can naturally determine the cluster centers and select them. For the others, a common approach is to select the bands that are closest to the centers of each group [21]. Apart from it, [18] proposes to select the bands that have the largest similarities with the others in the same group, and [25] selects the bands that have the largest information entropy.

Group-wise selection focuses on the reduction of the correlation among bands. Through the separation process, the selected bands are distributed to different groups and so can carry more distinctive information. Nevertheless, though many strategies have been proposed to select the representative bands in each group, the selection process is still individual, without considering the interaction among bands explicitly. Hence group-wise methods can usually provide a plausible and stable, but hardly outstanding result.

B. Point-wise Method

In point-wise methods, band selection task is considered as a combinatorial optimization problem, where the indexes of the desired bands are considered as a series of discrete optimization variable and some searching strategies are used to seek the optimal solution. According to the type of the objective function and the searching strategy, point-wise selection can be further divided into ranking-based, greedy-based and evolutionary-based methods.

Ranking-based methods first score the bands according to some ranking criterion such as band variance [26] and minimum constrained energy [27]. Since the solution is to simply select the bands with higher scores, ranking-based methods neglect the interaction among them and hence usually provide a band set with high dependency.

Greedy-based methods first design an objective function such like maximum ellipsoid volume [28] and minimum estimated abundance covariance [29], and search for the desired band subset via some greedy strategies, e.g., sequential forward selection (SFS) [29] and sequential backward selection (SBS) [28]. Greedy-based methods are rather efficient at the expense of insufficient search and calculation towards the optimum.

Evolutionary-based methods apply some evolutionary algorithms to search for the optimal solution. These algorithms include immune clone [30–32], particle swarm [29] and firefly algorithm [33], etc. Compared to greedy-based methods, these methods are more effective in searching the optimum but also take more computational cost and are hard to be tuned.

Compared to group-wise methods, the directive searching for the desired band subset encourages point-wise methods to capture the relationship among bands more exactly. However, without the preliminary grouping procedure, it is difficult to design an objective function which can make sure the selected band subset are always with low correlation.

III. Optimal Neighborhood Reconstruction

This section presents the motivation and some implementation details of optimal neighborhood reconstruction (ONR) method. First, the band selection problem is formulated and the objective function of ONR is proposed. Second, the algorithm to optimize the objective function is introduced. Third, a parameter selection strategy is given to adaptively set the parameter of ONR. Finally, the computational complexity of ONR is analysed.

A. Objective Function

We first define some notations that will be used throughout the paper. To distinguish different variables, we represent matrices, vectors and scalars by bold uppercase, bold lowercase and non-bold italic font (lowercase or uppercase) of characters respectively. A HSI can be defined as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_d] \in \mathbb{R}^{n\times d}$, where $\boldsymbol{x}_j$ is the jth band vector whose l_2 norm is scaled to 1, d is the number of bands and n is the number of pixels in each band. Suppose the finally selected bands are specified by $\boldsymbol{x}_{b_1}, \boldsymbol{x}_{b_2}, ..., \boldsymbol{x}_{b_m}$ where $\boldsymbol{b} = (b_1, b_2, ..., b_m)^T$ denotes the index vector. We assume there is $1 \le b_1 < b_2 < ... < b_m \le d$, in which m denotes the number of bands.

One rational criterion to examine the discrimination of a band subset is to assess its ability in reconstructing the whole bands. From this perspective, a prototype of our objective function is given as follows:

$$\min_{\boldsymbol{b},\boldsymbol{W}} \quad \mathcal{L}(\boldsymbol{E}) \quad s.t. \quad \boldsymbol{X} = [\boldsymbol{x}_{b_1}, \boldsymbol{x}_{b_2}, ...\boldsymbol{x}_{b_m}]\boldsymbol{W} + \boldsymbol{E}. \quad (1)$$

Here $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_d] \in \mathbb{R}^{m\times d}$ is the weight matrix, $\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_d] \in \mathbb{R}^{d\times n}$ is the error matrix and $\mathcal{L}$ is a function to evaluate the reconstruction error. According to Eq. (1), each column of $\boldsymbol{X}$ can be represented by:

$$\boldsymbol{x}_j = [\boldsymbol{x}_{b_1}, \boldsymbol{x}_{b_2}, ..., \boldsymbol{x}_{b_m}]\boldsymbol{w}_j + \boldsymbol{e}_j.$$

[1]The data set and code of this paper can be found on http://crabwq.github.io.

That means each $\boldsymbol{x}_j$ is considered to be correlated with all the selected bands, where the weight vector $\boldsymbol{w}_j$ characterizes this kind of correlation. However, this modeling strategy is too complex since too much bands are involved to reconstruct a single band, neglecting the sparsity among data set [34, 35]. As stated in Section I, $\boldsymbol{x}_j$ should have higher probability to be correlated with its neighborhood, rather than those which are distant to it. Hence, there are only a few of the selected bands having correlation with $\boldsymbol{x}_j$, and mostly are its neighborhoods. As an extreme case, we assume $\boldsymbol{x}_j$ is only correlated with two of its nearest neighborhoods, one is on its left side, and another is on its right side. Formally, for all $1 \leq j \leq d$, there exists a k which satisfies $b_k \leq j < b_{k+1}$ so that $\boldsymbol{x}_j$ is correlated with $\boldsymbol{x}_{b_k}$ and $\boldsymbol{x}_{b_{k+1}}$. With the above consideration, Eq. (1) can be rewritten as:

$$\begin{aligned} \min_{\boldsymbol{b},\boldsymbol{Z}} \quad & \mathcal{L}(\boldsymbol{E}), \\ s.t. \quad & \boldsymbol{x}_j = [\boldsymbol{x}_{b_k}, \boldsymbol{x}_{b_{k+1}}]\boldsymbol{z}_j + \boldsymbol{e}_j, \\ & for \quad b_k \leq j < b_{k+1}, \end{aligned} \tag{2}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_d] \in \mathbb{R}^{2\times d}$ is a simplified weight matrix. To make Eq. (2) meaningful for the cases that $j < b_1$ and $j > b_m$, we stipulate $b_0 = 0$, $b_{m+1} = d + 1$ and $\boldsymbol{x}_0 = \boldsymbol{x}_{d+1} = \boldsymbol{0}$.

The above analysis exploits the CNP to characterize the relationship between the selected bands and the entire HSI. Now the remaining problem is how to quantify the reconstruction error, i.e., how to choose $\mathcal{L}$. Since the number of the selected bands is limited and sometimes much less than d, to reconstruct the whole data set precisely is almost impossible. For arbitrary band combinations, there will always be bands that cannot be well reconstructed. They may be contaminated by noises or just have low correlation with the selected bands. What we suppose to evaluate via $\mathcal{L}$, however, is how many bands can be well reconstructed and in which extent they are reconstructed. If a large loss $\boldsymbol{e}_j$ is found when reconstructing band $\boldsymbol{x}_j$, it means $\boldsymbol{x}_j$ is inexactly reconstructed, and we are not willing to see this kind of inexactness greatly influence the searching for the optimal band subset. Hence with $\boldsymbol{e}_j$ increasing, the value of $\mathcal{L}$ shouldn't be increased endlessly. Based on the above consideration, the definition of $\mathcal{L}$ is given as:

$$\mathcal{L}(\boldsymbol{E}) = \sum_{j=1}^{d} g_\tau(\|\boldsymbol{e}_j\|_2),$$

where $\|\cdot\|_2$ is the l_2 norm, $g_\tau : \mathbb{R} \to \mathbb{R}$ is a noise reducer to limit the growth of $\mathcal{L}$:

$$g_\tau(x) = \begin{cases} x, & x \leq \tau, \\ \tau, & x > \tau, \end{cases}$$

and τ is a threshold to distinguish "exact" reconstruction from "inexact" reconstruction.

In summary, the final objective function turns to be:

$$\begin{aligned} \min_{\boldsymbol{b},\boldsymbol{Z}} \quad & \sum_{j=1}^{d} g_\tau(\|\boldsymbol{e}_j\|_2), \\ s.t. \quad & \boldsymbol{x}_j = [\boldsymbol{x}_{b_k}, \boldsymbol{x}_{b_{k+1}}]\boldsymbol{z}_j + \boldsymbol{e}_j, \\ & for \quad b_k \leq j < b_{k+1}. \end{aligned} \tag{3}$$

B. Optimization

By assigning each $\boldsymbol{x}_j$ to its belonging band interval which is determined by $\boldsymbol{b}$, Eq. (3) is equivalent into:

$$\min_{\boldsymbol{b},\boldsymbol{Z}} \quad \sum_{k=0}^{m} \sum_{j=b_k+1}^{b_{k+1}-1} g_\tau(\|\boldsymbol{x}_j - [\boldsymbol{x}_{b_k}, \boldsymbol{x}_{b_{k+1}}]\boldsymbol{z}_j\|_2). \tag{4}$$

Noting that each column of $\boldsymbol{Z}$ is independently optimized, Eq. (4) can be further reformulated as:

$$\min_{\boldsymbol{b}} \quad \sum_{k=0}^{m} \sum_{j=b_k+1}^{b_{k+1}-1} \min_{\boldsymbol{z}_j} \; g_\tau(\|\boldsymbol{x}_j - [\boldsymbol{x}_{b_k}, \boldsymbol{x}_{b_{k+1}}]\boldsymbol{z}_j\|_2). \tag{5}$$

Since g_τ is monotonically non-decreasing, the solution of the following equation must be the solution of Eq. (5) (this can be easily demonstrated via reduction to absurdity).

$$\min_{\boldsymbol{b}} \quad \sum_{k=0}^{m} \sum_{j=b_k+1}^{b_{k+1}-1} g_\tau(\min_{\boldsymbol{z}_j} \|\boldsymbol{x}_j - [\boldsymbol{x}_{b_k}, \boldsymbol{x}_{b_{k+1}}]\boldsymbol{z}_j\|_2). \tag{6}$$

Here we define two auxiliary variables $\boldsymbol{L} \in \mathbb{R}^{d\times d\times d}$ and $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ as:

$$L_{l,r,j} = \min_{\boldsymbol{z}} \|\boldsymbol{x}_j - [\boldsymbol{x}_l, \boldsymbol{x}_r]\ \boldsymbol{z}\|_2, \tag{7}$$

$$S_{l,r} = \sum_{j=l+1}^{r-1} g_\tau(L_{l,r,j}), \tag{8}$$

where $L_{l,r,j}$ is an element in $\boldsymbol{L}$, evaluating the loss to reconstruct $\boldsymbol{x}_j$ with $\boldsymbol{x}_l$ and $\boldsymbol{x}_r$ as the bases. $S_{l,r}$ is the loss to reconstruct band interval $\{\boldsymbol{x}_{l+1}, ..., \boldsymbol{x}_{r-1}\}$. With the above definition, Eq. (6) can be rewritten as:

$$\min_{\boldsymbol{b}} \sum_{k=0}^{m} S_{b_k, b_{k+1}}. \tag{9}$$

Until now, the original optimization problem has been converted to two subproblems as described in Eq. (7) and (9) respectively.

1) Solution to Eq. (7): Eq. (7) is actually a least square problem, whose solution is given by:

$$\boldsymbol{z}^* = ([\boldsymbol{x}_l, \boldsymbol{x}_r]^T[\boldsymbol{x}_l, \boldsymbol{x}_r])^{-1}[\boldsymbol{x}_l, \boldsymbol{x}_r]^T\boldsymbol{x}_j. \tag{10}$$

2) Solution to Eq. (9): We first define two matrices $\boldsymbol{D} \in \mathbb{R}^{(d+1)\times(m+1)}$ and $\boldsymbol{Q} \in \mathbb{R}^{(d+1)\times(m+1)}$ as follows.

$$D_{i,j} = \min_{b_1<...<b_j} \sum_{k=0}^{j-1} S_{b_k, b_{k+1}}, \text{ s.t. } b_j = i, \tag{11}$$

$$Q_{i,j} = \arg\min_{j-1\leq k\leq i} \; D_{k,j-1} + S_{k,i}. \tag{12}$$

Note that there should be $j \leq i$ since $b_j = i$. To solve Eq. (9), we present two theorems as follows:

Theorem 1. *The following equation holds for* $1 \leq i \leq d + 1$ *and* $1 \leq j \leq i$*:*

$$D_{i,j} = \min_{j-1\leq k\leq i} \; D_{k,j-1} + S_{k,i}. \tag{13}$$

Theorem 2. *If we set* $\boldsymbol{b}^* = [b_1^*, b_2^*, ..., b_m^*]^T$ *according to Eq. (14),* $\boldsymbol{b}^*$ *should be one of the solution to Eq. (9).*

$$b_j^* = Q_{b_{j+1}^*, j+1}. \tag{14}$$

The proof of Theorem 1 and 2 is given in appendix.

According to Theorem 1, the jth column of $\boldsymbol{D}$ only depends on the $(j-1)$th column of $\boldsymbol{D}$. In other words, once we have known the $(j-1)$th column of $\boldsymbol{D}$, we can achieve its jth column by enumerating all the possible values of k according to Eq. (13). Besides, since it is easy to know there is $D_{i,1} = S_{0,i}$ according to the definition of $\boldsymbol{D}$, the first column of $\boldsymbol{D}$ can be calculated in advance, which means the whole matrix $\boldsymbol{D}$ is solvable. Furthermore, since $\boldsymbol{D}$ is solvable, $\boldsymbol{Q}$ can be known as well, which indicates $\boldsymbol{b}^*$ can be obtained according to Eq. (14).

The pseudo code for the optimization of Eq. (9) is shown in Algorithm 1. Note that the algorithm we introduce here assumes $\boldsymbol{S}$ is already achieved. Since the calculation of $\boldsymbol{S}$ is correlated with the selection of τ, we will leave it to Section III-C.

Algorithm 1 Optimization of Eq. (9)

Input: All bands $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_d]$, number of bands m, auxiliary variable $\boldsymbol{S}$.

Set $D_{i,1} \leftarrow S_{0,i}$ for each $1 \leq i \leq m+1$.
Set $D_{i,j} \leftarrow \infty$ for each $1 \leq i \leq d+1$ and $1 < j \leq m+1$.
for $j \leftarrow 2$ **to** $m+1$ **do**
  for $i \leftarrow j$ **to** $d+1$ **do**
    for $k \leftarrow j-1$ **to** $i-1$ **do**
      if $D_{k,j-1} + S_{k,i} < D_{i,j}$ **then**
        $D_{i,j} \leftarrow D_{k,j-1} + S_{k,i}$.
        $Q_{i,j} \leftarrow k$.
      end if
    end for
  end for
end for
$b^*_{m+1} \leftarrow d+1$.
for $j \leftarrow m$ **to** 1 **do**
  $b^*_j \leftarrow Q_{b^*_{j+1}, j+1}$.
end for

Output: The indexes of m selected bands $b^*_1, b^*_2, ..., b^*_m$.

C. Parameter Selection

This subsection discusses about the selection of the parameter τ. First, the motivation to develop the parameter selection approach is given. Then the detailed procedure to set τ is provided. In the last, short analysis is made to help interpret the approach.

About parameter τ, it should be pointed out that it largely influences the performance of ONR. When τ is too small, only bands which have very low reconstruction errors will contribute to the objective function. So to deal with it, ONR will select a series of bands which are highly correlated to ensure some of the bands meet the required condition. When τ is too large, ONR tends to select the noisy bands, since once they are selected, their own reconstruction errors will be reduced to 0 directly, which leads to a large decrease to the objective function. However, both of the above two situations are not desired to be seen, so how to adaptively find a feasible τ is very important.

The proposed parameter selection approach is inspired by two concerns. The first one is that we should first distinguish noisy bands from clean bands. Since this task seems easier than to directly determine τ, and once accomplished, it helps to find the feasible τ, e.g., we can just simply set τ to the minimal possible reconstruction error of noisy bands to reduce the risk to select them. The second concern is that τ should be adjusted according to the band selection result. If τ is just determined based on some prior information but without feedback on how it influences the algorithm, there will be too much uncertainty in the approach. In consideration of these, two steps will be conducted to set τ.

1) Recognition of noisy bands: This step finds out the noisy bands according to the below procedure:

- Calculate the minimal possible reconstruction error (exclude self-reconstruction) $\boldsymbol{J} \in \mathbb{R}^d$ of each band:

$$J_i = \min_{j<i<k} L_{j,k,i}. \tag{15}$$

  Then sort $\boldsymbol{J}$ in ascending order to get $\tilde{\boldsymbol{J}}$.
- Calculate the histogram $\boldsymbol{H} \in \mathbb{R}^h$ of $\tilde{\boldsymbol{J}}$. Here h is the number of bins. Generally, clean bands have lower reconstruction errors, so they tend to accumulate on the first several bins of $\boldsymbol{H}$.
- Define a window size w, and find the minimal index i^* of $\boldsymbol{H}$ which satisfies:

$$\frac{\sum_{k=1}^{2w+1} H_k - \sum_{k=i^*-w}^{i^*+w} H_k}{\sum_{k=1}^{2w+1} H_k} > \epsilon. \tag{16}$$

  Here w defines a sliding window to smooth $\boldsymbol{H}$. ϵ identifies the threshold of the decreasing ratio of $\boldsymbol{H}$. When Eq. (16) is satisfied, the number of bands contained in the current bin (H_{i^*}) are considered much lesser than the first bin, which means these bands and bands in succeeding bins are probably noisy bands. So, the bands contained in the first i^*-1 bins of $\boldsymbol{H}$ are identified as clean bands, while the others are noisy bands. We name the indexes of clean bands as $\boldsymbol{\xi} = (\xi_1, \xi_2, ...)^T$.

2) Grid search: After the noisy bands are recognized, the parameter τ will be set according to the following procedures.

- Solve Eq. (6) with $\tau = \infty$, and set τ_{max} to the largest reconstruction error in this case:

$$\tau_{max} = \max_{1 \leq j \leq d} L_{b^\infty_k, b^\infty_{k+1}, j},\ s.t.\ b^\infty_k \leq j < b^\infty_{k+1}. \tag{17}$$

  Here b^∞ corresponds to the solution of Eq. (6) when $\tau = \infty$.
- Evenly split $(0, \tau_{max}]$ into t sub-intervals $(0, \tau_1]$, $(\tau_1, \tau_2]$, ...,$(\tau_{t-1}, \tau_t]$ ($\tau_{max} = \tau_t$).
- Solve Eq. (6) with $\tau = \tau_1, \tau_2, ..., \tau_t$ sequentially, and check the reconstruction error for each clean band ξ_i:

$$E_{\xi_i} = L_{b^\tau_k, b^\tau_{k+1}, \xi_i},\ s.t.\ b^\tau_k \leq \xi_i < b^\tau_{k+1}. \tag{18}$$

  Once more than a percentage γ of clean bands has reconstruction error lower than τ, stop searching. The current τ is just the desired one, and the current b^τ is the indexes of the final selected bands.

Intuitively, the first step aims to find an index of $\tilde{\boldsymbol{J}}$ that indicates the rapid growth of it (some examples are shown in Fig. 2). Since we assume that noisy bands have much larger reconstruction errors than clean bands, this phenomenon may indicate the occurrence of noisy bands. The second step aims to find the minimal value of τ such that most of the clean bands can be well reconstructed, so this τ is not too strict for the clean bands and also not too loose for the noisy bands. It is worth noting that when $\tau > \tau_{max}$, the solution of Eq. (6) will remain unchanged even if τ is changed.

The overall procedure of selecting τ involves 5 parameters: h, w, ϵ, t and γ as described above. After slightly parameter tuning, they are set as $h = 0.6d, w = 5, \epsilon = 0.6, t = 100$ and $\gamma = 95\%$. Specifically, h, w and ϵ are tuned according to the curve of $\tilde{\boldsymbol{J}}$ so that the results agree with our intuition (like what is shown in Fig. (2)). t and γ are tuned according to ONR's experimental performance. We want to highlight here is that these parameters will be fixed under different experimental settings so that the performance gain obtained by parameter tuning will be minimized.

The pseudo code of ONR with adaptive parameter selection is given in Algorithm 2.

Algorithm 2 Optimal Neighborhood Reconstruction

Input: All bands $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_d]$, number of bands m.

1: Set parameter $w = 5$, $h = 0.6 \cdot d$, $\epsilon = 0.6$, $t = 100$,
2: Get $\boldsymbol{L}$ according to Eq. (7) and Eq. (10).
3: Get $\boldsymbol{J}$ according to Eq. (15) and sort it to get $\tilde{\boldsymbol{J}}$.
4: Compute the histogram $\boldsymbol{H}$ of $\tilde{\boldsymbol{J}}$.
5: Find the index i according to Eq. (16) and achieve the indexes of clean bands $\boldsymbol{\xi}$.
6: Get $\boldsymbol{S}^{\infty}$ with $\tau = \infty$ according to Eq. (8).
7: Conduct Algorithm 1 with $\boldsymbol{S}^{\infty}$, and get τ_{max} according to Eq. (17).
8: **for** $\tau = \frac{\tau_{max}}{t}$, $\frac{2\tau_{max}}{t}$, ..., τ_{max} **do**
9: Get $\boldsymbol{S}^{\tau}$ using the current τ.
10: Conduct Algorithm 1 with $\boldsymbol{S}^{\tau}$, and get $\boldsymbol{b}^{\tau}$.
11: Calculate E_{ξ_i} for each ξ_i in $\boldsymbol{\xi}$ using Eq. (18).
12: **if** more than 95% of the elements of $E_{\xi_i} < \tau$ **then**
13: $\boldsymbol{b}^* = \boldsymbol{b}^{\tau}$.
14: Break.
15: **end if**
16: **end for**

Output: The indexes of m selected bands b^*.

D. Complexity of ONR

This section discuss about some issues concerning the time complexity of ONR. The first part presents a trick to accelerate ONR, and the second part gives detailed analyses toward the time complexity. Note that the acceleration trick only improves the efficiency, but does not affect the band selection result.

1) Acceleration Trick: This trick is to accelerate the calculation of each element in $\boldsymbol{L}$. We first define a covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} \in \mathbb{R}^{d\times d}$, and a notation $\boldsymbol{\Delta}^{(l,r)} = ([\boldsymbol{x}_l\ \boldsymbol{x}_r]^T [\boldsymbol{x}_l\ \boldsymbol{x}_r])^{-1} \in \mathbb{R}^{2\times 2}$. Then Eq. (10) can be rewritten as:

$$\begin{aligned} \boldsymbol{z}^* &= \begin{bmatrix} \Sigma_{l,l} & \Sigma_{l,r} \\ \Sigma_{r,l} & \Sigma_{r,r} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{l,j} \\ \Sigma_{r,j} \end{bmatrix} \\ &= \boldsymbol{\Delta}^{(l,r)} \begin{bmatrix} \Sigma_{l,j} \\ \Sigma_{r,j} \end{bmatrix}. \end{aligned} \tag{19}$$

Here $\Sigma_{i,j}$ is the i-row j-column element of $\boldsymbol{\Sigma}$. By substituting Eq. (19) into Eq. (7), we get:

$$\begin{aligned} L^2_{l,r,j} &= \|\boldsymbol{x}_j - [\boldsymbol{x}_l\ \boldsymbol{x}_r]\ \boldsymbol{z}^*\|_2^2 \\ &= \Sigma_{j,j} - \begin{bmatrix} \Sigma_{l,r}\ \Sigma_{r,j} \end{bmatrix}\ \boldsymbol{\Delta}^{(l,r)} \begin{bmatrix} \Sigma_{l,j} \\ \Sigma_{r,j} \end{bmatrix}. \end{aligned} \tag{20}$$

Eq. (20) implies that once $\boldsymbol{\Sigma}$ and $\boldsymbol{\Delta}^{(l,r)}$ are obtained, each element of $\boldsymbol{L}$ can be achieved with only a fixed number of scalar computations.

2) Complexity Analysis: As can be seen from Algorithm 2, the most time consuming part of ONR is the calculation of $\boldsymbol{L}$ in line 2 and the main iteration from line 8 to 16. The other procedures like the computation of $\tilde{\boldsymbol{J}}$ or $\boldsymbol{H}$ are instead with little costs, so they are not considered here.

a) The complexity to calculate $\boldsymbol{L}$. Originally, we need to calculate each element of $\boldsymbol{L}$ according to Eq. (10) and substitute the result to Eq. (7) in order to achieve $\boldsymbol{L}$. Eq. (10) actually involves dot product of vectors for four times, scalar product of vectors for four times and matrix inversion (of a 2×2 matrix) for one time. The corresponding complexity is $O(8n + 2^3) = O(n)$. Similarly, substituting the result into Eq. (7) also costs $O(n)$. So computing $\boldsymbol{L}$ directly costs $O(d^3 n)$. If the above trick is used, $\boldsymbol{\Sigma}$ should be achieved at first, which costs $O(d^2 n)$. Once $\boldsymbol{\Sigma}$ is attained, $\boldsymbol{\Delta}^{(l,r)}$ for $1 \le l < r \le d$ can be calculated within $O(d^2)$ ($\boldsymbol{\Delta}^{(l,r)}$ is only a 2×2 matrix). Then since the second line of Eq. (20) only involves a fixed number of scalar computations, each element of $\boldsymbol{L}$ can be computed within $O(1)$. Consequently, the complexity to attain $\boldsymbol{L}$ can be reduced from $O(d^3 n)$ to $O(d^2 n + d^3)$ owing to the acceleration trick.

b) The complexity of main iteration. The most time-consuming part in main iteration is the calculation of $\boldsymbol{S}$ in line 9 and the conduction of Algorithm 1 in line 10. It is easy to see the computational cost of Algorithm 1 is $O(d^2 m)$, so iterating over line 9 for t times costs $O(td^2 m)$ in total. When referring to $\boldsymbol{S}$, $O(td^3)$ is needed if we directly calculate it according to Eq. (8). However, the complexity can be reduced to $O(td^2 m + d^3 log(d))$ because the value of τ increases progressively. The key idea to achieve this reduction is to decide how the $\boldsymbol{S}$ obtained in the last iteration should be updated in the current iteration. To accomplish this, we need to find the elements in $\boldsymbol{L}$ whose values are between the last τ and the current τ and adjust $\boldsymbol{S}$ according to each of them. We will not list all the details here because they are too tedious and somehow trifling. One can refer to our code to find more information[2].

Summarizing the above two procedures, the total complexity of ONR is $O(d^2 n + td^2 m + d^3 log(d))$. Generally there are $n \gg tm$ and $n \gg dlog(d)$, so the complexity is approximate

[2]If this paper is finally accepted, we will publish our code.

to $O(d^2n)$. Moreover, the $O(d^2n)$ here refers to the naive algorithm to compute matrix product when we calculate $\mathbf{\Sigma}$. So if faster algorithms like [36] can be used, the efficiency of ONR can be further improved.

IV. Experiment

To examine the effectiveness of ONR, experiments are conducted on several real-world HSI data sets. First, the experimental settings are introduced. Then the results of classification experiments are shown to see whether ONR is superior to the other state-of-the-art methods. Finally the computational efficiency of different methods is compared and analysed.

A. Experimental Setup

The subsection includes description of data sets, introduction of comparison methods and parameter settings.

1) Data Set: Six real-world HSI data sets are used in the experiments, namely Indian Pines, Pavia University, Salinas, Kennedy Space Center (KSC), Botswana and Leaves.

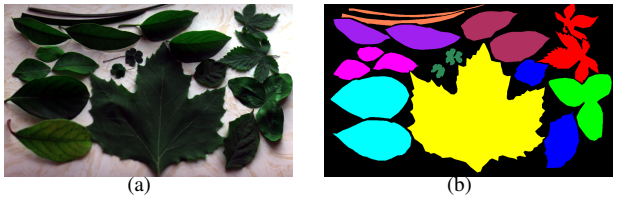
- **Indian Pines** is captured by AVIRIS sensor in Northwestern Indiana in 1992. It contains 145×145 pixels, 224 bands and 16 classes of interest. The wavelengths of bands range from 0.4 μm to 2.5 μm. 24 bands are removed due to water absorption, and 200 bands are used in the experiments.
- **Pavia University** is acquired by ROSIS sensor in Pavia, northern Italy, in 2002. It contains 610×340 pixels, 9 classes of interest and 103 bands, with wavelengths range from 0.43 μm to 0.86 μm.
- **Salinas** is collected by AVIRIS over Salinas Valley, California in 1998. It has 512×217 pixels, 224 bands and 16 classes of interest. The wavelengths of bands range from 0.4 μm to 2.5 μm. 20 water absorption bands are discarded and 204 bands are used in the experiments.
- **KSC** is acquired by AVIRIS over the Kennedy Space Center, Florida in 1996. It has 512×614 pixels, 224 bands and 13 classes of interest. The wavelengths of bands range from 0.4 μm to 2.5 μm. 48 absorption bands are removed and 176 bands are used in the experiments.
- **Botswana** is captured by NASA EO-1 satellite over Okavango Delta, Botswana in 2001. It consists of 1476×256 pixels, 242 bands and 14 classes of interest. The wavelengths of bands range from 0.4 μm to 2.5 μm. 97 water absorption bands are removed and 145 bands are used in the experiments.
- **Leaves**, as shown in Fig. 1, was a test image captured by GaiaField - an portable hyperspectral imager on May 30, 2018. It is a close shot of 10 different classes of leaves (see details in Table I). It has 1168×696 pixels and 520 bands, with wavelength range from 0.4μm to 1.0μm.

2) Comparison Method: There are a total of 7 band selection methods included as competitors. They are:

- **Uniform band selection (UBS)** [26] just simply selects the bands uniformly.
- **Ward's Linkage strategy Using Mutual Information (WaLuDi)** [18] first separates the bands into clusters via hierarchical clustering, where Kullback-Leibler divergence is adopted to capture the feature-level similarity. Then in each cluster, the band that has the highest similarity to the others is selected.
- **Enhanced fast density-peak-based clustering (E-FDPC)** [37] aims to find the potential clustering centers, which usually have two properties: large local density and large inter-cluster distance. The former one means there should be lots of data points surrounding the clustering centers, while the later one means the two clustering centers should be far from each other. E-FDPC then weights these two factors to rank the bands.
- **Normalized Cut based Optimal Clustering (NC-OC)** [38] is a group-wise selection method. It proposes a dynamic programming based optimization method to search for the optimal clustering result, and a rank on clusters strategy to select the representative bands in each cluster. Here we adopt the normalized cut (NC) as the clustering criterion and MVPCA as the ranking criterion.
- **Rank minimization band selection (RMBS)** [21] searches for the low-rank structure among bands via low-rank representation. Then it clusters the bands into groups and select one in each group to constitute the band subset.
- **Orthogonal projection-based band selection (OPBS)** [39] begins with a one-band subset which contains the band with maximum variance. Then it select the band which can maximize the orthogonal projection to the subspace spanned by the subset, and add it to the subset.

3) Parameter Setting: For the parameter settings about the comparative methods, most of them are free of parameter. For RMBS, as an exception, its parameter is tuned via a grid search in $\{10^{-4}, 10^{-3}, ..., 10^{3}, 10^{4}\}$ on Indian Pines, and fixed for the other data sets.

For ONR, we have discussed about the selection of τ in



(a) (b)

Fig. 1. The true color image of Leaves data set and the ground truth. (a) True color image. (b) Ground truth.

TABLE I
Class name and sample number for each class of Leaves

Index	Class Name	# of Samples
1	Platanus acerifolia	165973
2	Duchesnea indica	29915
3	Shamrock	34746
4	Rugosa rose	31782
5	Creeping oxalis	4793
6	Ligustrum lucidum	73658
7	Ligustrum quihoui	21600
8	Sweet-scented osmanthus	32868
9	Ophiopogon japonicus	11843
10	Paederia foetida	38017

Section III-C. Here Fig. 2 and Fig. 3 illustrate the procedure to find τ more intuitively. In Fig. 2, it seems there is always a state that $\tilde{\boldsymbol{J}}$ starts to increase rapidly for all the data sets. This is agree with the assumption that the noisy bands are more unlikely to be well reconstructed and will have much larger reconstruction errors. In Fig. 3, it can be observed that when τ is too small (as the case in (a)), the distribution of the selected bands are over-concentrated. Most of the bands still have large reconstruction errors. When τ is too large (as the case in (b)), many bands with indexes at around 130 are selected. However, these bands contribute little to the overall reconstruction results since two contiguous bands of them only help one band between them to have error lower than τ. Besides, spectral curves of different classes are hard to be distinguished in this interval. It seems that (c) gives a proper estimation of τ, because most of the bands have errors lower than τ, and contribute a lot to the overall reconstruction.

B. Classification Experiment

To examine the effectiveness of ONR, classification experiments are conducted on the above mentioned 6 data sets. Support vector machine (SVM) [40] and k-nearest neighborhood (KNN) [41] are adopted as classifiers to examine the classification performance. For each data set, 10% of the samples are randomly selected to train the classifier, while the remaining 90% are used in testing. We run all the experiments 10 times individually to reduce the randomness. Fig. 4 and 5 plot the overall accuracy (OA) curves produced by SVM and KNN for all the data sets, where m varies from 3 to 30 each 3 interval. Table II and III list the OA values averaged over the cases when $m = 3, 6, ..., 30$ on different data sets, where the best results are in bold and the second best are underlined. Table IV lists the indexes of 15 bands selected by ONR for each data set.

As is shown in Fig. 4 and 5, in most of the time, the performance of ONR is superior to the others. Although some competitors achieve comparative results with ONR on some data sets, they are not very robust in all the cases. For example, OPBS achieves a satisfactory performance on Pavia University and Salinas data sets when SVM is employed. But it fails on KSC data set, worse than the lower bound of the figures. E-FDPC performs well on Indian Pines and Salinas data sets, but it still inferior to most of the competitors when referring to KSC or Botswana. As for ONR, it attains a much more robust performance, and even dominates all the other methods on some data sets such as Botswana and Leaves.

C. Statistical Test

In order to find whether the proposed ONR has significant differences with the comparative methods among various conditions, we conduct Wilcoxon signed-rank test [42] between ONR and the other methods. The data we consider here is averaged overall accuracies on different data sets using different classifiers, i.e., each method is represented by 12 data items collected from Table II and III. The null hypothesis is that ONR has no significant differences with the other methods. The level of significance considered here is $\alpha = 0.05$. Table V shows the p-values when comparing ONR to the other 5 methods. Since all the p-values are less than α, we can conclude that ONR has significant differences with all the other methods.

D. Efficiency Test

To verify the efficiency of ONR, we record the time costs for different methods to select 30 bands on all the data sets. The times reported here include all the necessary steps like pre-processing, e.g., for ONR, its time cost includes all the steps listed in Algorithm 2. We also report the running times to perform SVM classification using 30 bands and all bands as references.

All the methods are implemented by MATLAB R2016a, and are conducted using an Intel Core i7-6800k 3.40-GHZ CPU with 64-GB RAM,

According to the results shown in Table VI, though ONR is not as efficient as E-FDPC, it is faster than the other competitors in most of the times. Moreover, since ONR is still efficient on data sets with large n such as Pavia University, KSC and Botswana, it seems that it is less sensitive to the growth of the spatial dimension, and so can be applied to applications that involve images with large spatial sizes.

E. Discussion

According to the above experimental results, we would like to discuss about some interesting phenomena, and try to give some constructive suggestions to the design of band selection algorithm.

1) The influence of the uniformity of the selected bands. In the previous studies, a good band subset is supposed to has some degrees of uniformity, i.e., the indexes of the selected bands should have a uniform distribution to some extent. This is true to some extent, since we can see the performance of UBS is pretty stable. However, the performance of ONR leads to some different conclusions. As is illustrated in Table IV, the bands selected by ONR are not very uniform. For Leaves, the selected bands only cover about $\frac{3}{5}$ of the spectrum, (the indexes of selected bands range from 26 to 324 as shown in Table IV, and there are 520 bands in total), while for KSC, this ratio is reduced to $\frac{2}{5}$. Nevertheless, with such an uneven distribution of band indexes, ONR still achieves a good performance on those two data sets when 15 bands are selected. So we can learn that the optimal band subset may not be distributed very uniformly, since the ground objects may be undistinguishable with respect to some specific electromagnetic waves.

2) The reason why ONR works. There are two main reasons why ONR can achieve a promising performance. First, ONR assembles band decorrelation and noise reduction into a unified optimization procedure. Through optimizing a reconstruction-based criterion, it offers strong resistance when two highly correlated bands are selected simultaneously. By introducing a noise reducer g_τ, the influence of noisy bands is minimized, which makes ONR have robust performance on seriously polluted data sets. Second, the optimization method utilized in ONR can ensure the global optimality of the

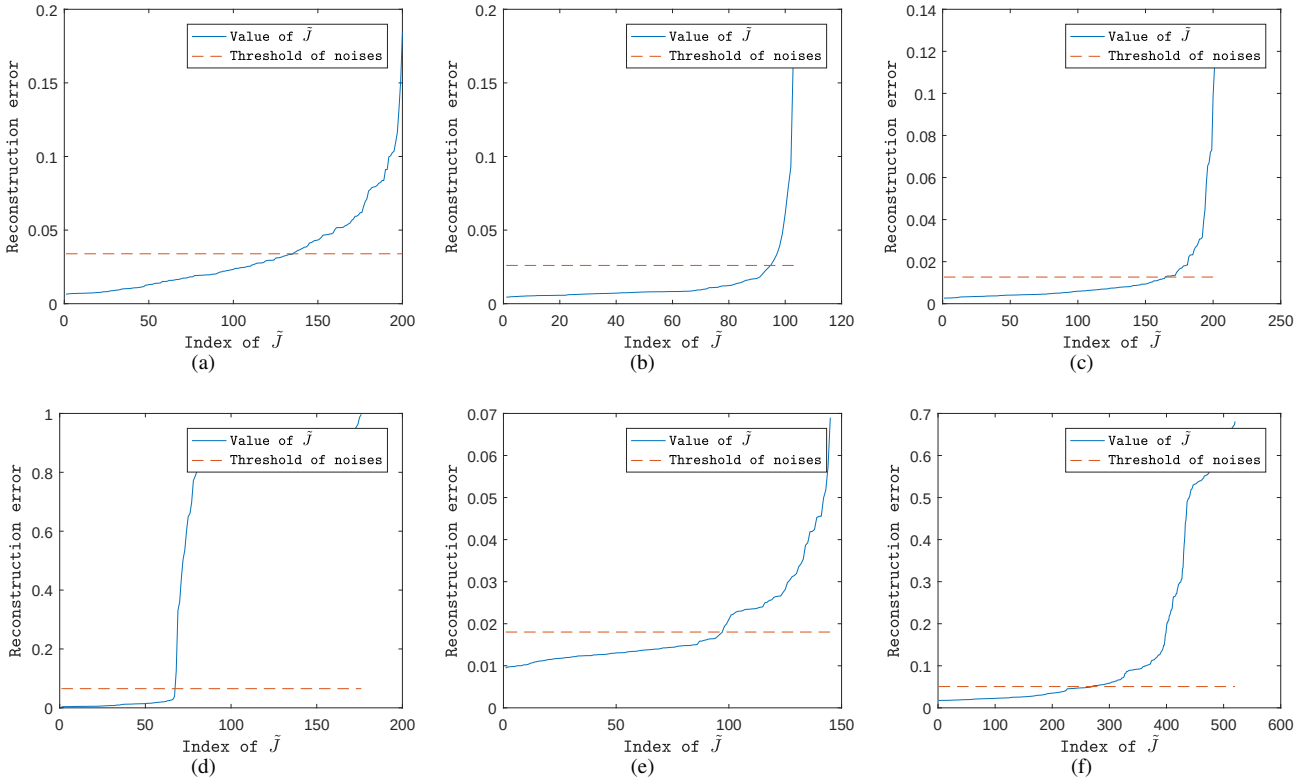


Fig. 2. The illustration to find the noisy bands for different data sets. (a)-(f) are the results for Indian Pines, Pavia University, Salinas, KSC, Botswana and Leaves data sets, respectively. The blue curve is the sorted minimal possible reconstruction error $\tilde{J}$, and the red line indicates the threshold to separate the noisy and clean bands in accordance with the approach in Section III-C. When J_i is larger than the threshold, X_i will be considered as a noisy band.

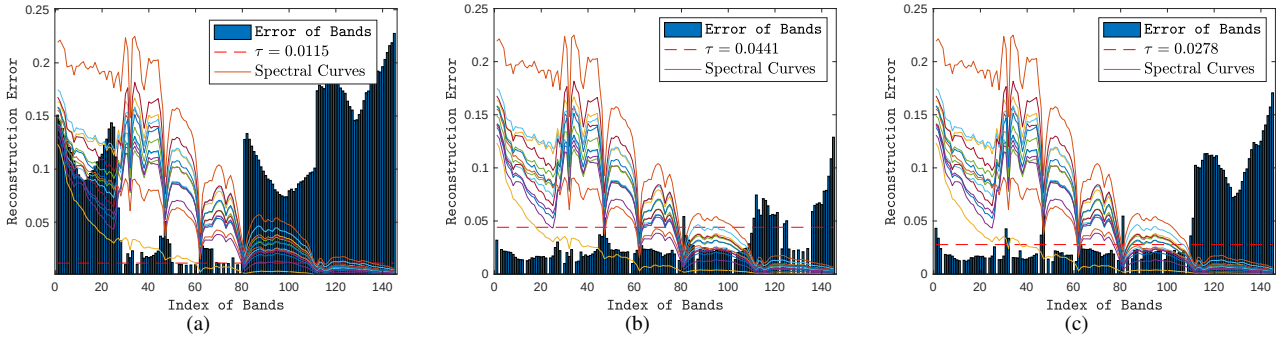


Fig. 3. The relationship between τ and the band selection results for Botswana data set. Spectral curves of 14 different classes are also plotted for comparison. (a)-(c) are the reconstruction errors of all the bands when $\tau = 0.0115, 0.0441$ and 0.0278 respectively. Each band is linearly reconstructed by two of the selected bands on its left and right. (c) is just the result produced by Algorithm 2. Once the reconstruction error of a band is 0, it means that band is one of the selected bands. The SVM classification accuracies concerning (a), (b) and (c) are 0.8573, 0.8840 and 0.9085 respectively.

TABLE II
OVERALL ACCURACY BY SVM WITH STANDARD VARIANCE AVERAGED OVER DIFFERENT m.

	Indian Pines	Pavia University	Salinas	KSC	Botswana	Leaves
UBS	$70.33 \pm 0.40\%$	$89.29 \pm 0.12\%$	$88.34 \pm 0.14\%$	$75.11 \pm 0.64\%$	$85.04 \pm 0.47\%$	$78.93 \pm 0.01\%$
E-FDPC	$73.55 \pm 0.40\%$	$88.54 \pm 0.13\%$	$90.81 \pm 0.11\%$	$72.62 \pm 0.61\%$	$86.43 \pm 0.48\%$	$82.30 \pm 0.02\%$
WaLuDi	$70.83 \pm 0.40\%$	$89.21 \pm 0.10\%$	$89.93 \pm 0.10\%$	$74.96 \pm 0.59\%$	$85.58 \pm 0.44\%$	$79.44 \pm 0.04\%$
NC-OC	$\underline{74.73 \pm 0.36\%}$	$\underline{89.38 \pm 0.12\%}$	$\mathbf{91.29 \pm 0.15\%}$	$\underline{75.25 \pm 0.65\%}$	$86.80 \pm 0.65\%$	$82.52 \pm 0.01\%$
RMBS	$72.86 \pm 0.56\%$	$88.61 \pm 0.08\%$	$90.89 \pm 0.16\%$	$73.35 \pm 0.57\%$	$85.59 \pm 0.77\%$	$\underline{84.41 \pm 0.01\%}$
OPBS	$67.48 \pm 0.67\%$	$88.99 \pm 0.12\%$	$91.08 \pm 0.13\%$	$51.12 \pm 0.35\%$	$\mathbf{87.51 \pm 0.51\%}$	$82.28 \pm 0.03\%$
ONR	$\mathbf{74.74 \pm 0.40\%}$	$\mathbf{89.51 \pm 0.13\%}$	$\underline{91.13 \pm 0.16\%}$	$\mathbf{75.34 \pm 0.61\%}$	$\underline{87.17 \pm 0.61\%}$	$\mathbf{86.74 \pm 0.01\%}$

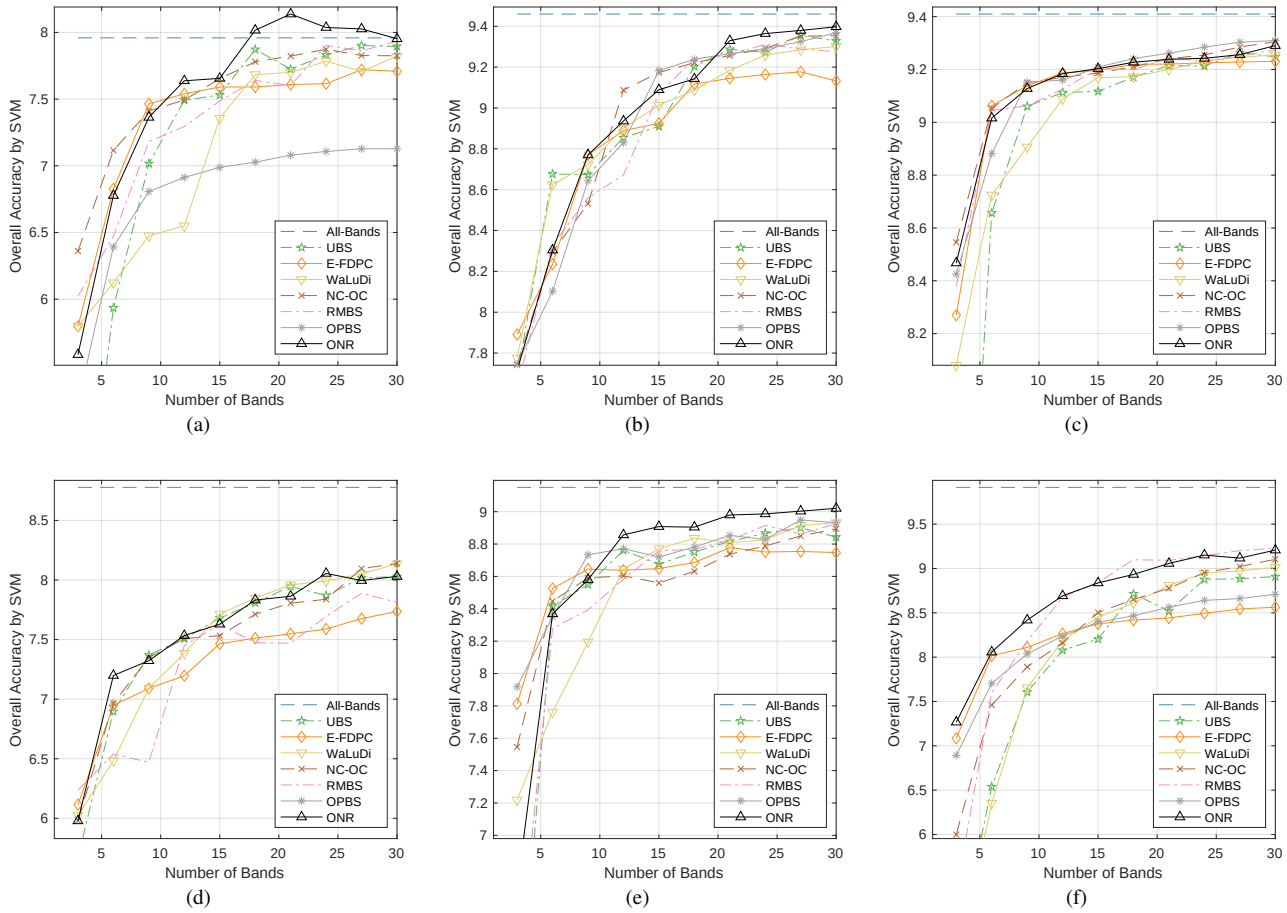


Fig. 4. Overall accuracy curves produced by SVM classifier. (a)-(f) OA curves for Indian Pines, Pavia University, Salinas, KSC, Botswana and Leaves data sets, respectively.

TABLE III
OVERALL ACCURACY BY KNN WITH STANDARD VARIANCE AVERAGED OVER DIFFERENT m

	Indian Pines	Pavia University	Salinas	KSC	Botswana	Leaves
UBS	$61.17 \pm 0.53\%$	$83.81 \pm 0.11\%$	$85.65 \pm 0.17\%$	$80.63 \pm 0.68\%$	$81.34 \pm 0.62\%$	$74.37 \pm 0.01\%$
E-FDPC	$\mathbf{67.46 \pm 0.37\%}$	$85.62 \pm 0.13\%$	$88.28 \pm 0.15\%$	$81.29 \pm 0.87\%$	$83.56 \pm 0.66\%$	$79.40 \pm 0.07\%$
WaLuDi	$64.67 \pm 0.37\%$	$83.93 \pm 0.08\%$	$87.52 \pm 0.14\%$	$81.02 \pm 0.70\%$	$81.79 \pm 0.62\%$	$75.47 \pm 0.01\%$
NC-OC	$66.03 \pm 0.35\%$	$85.44 \pm 0.15\%$	$88.18 \pm 0.15\%$	$81.21 \pm 0.76\%$	$83.09 \pm 0.73\%$	$79.40 \pm 0.06\%$
RMBS	$66.48 \pm 0.35\%$	$84.31 \pm 0.11\%$	$87.93 \pm 0.11\%$	$76.74 \pm 0.61\%$	$81.43 \pm 0.77\%$	$80.83 \pm 0.06\%$
OPBS	$62.52 \pm 0.55\%$	$83.37 \pm 0.12\%$	$87.19 \pm 0.16\%$	$54.90 \pm 0.45\%$	$82.99 \pm 0.57\%$	$79.41 \pm 0.02\%$
ONR	$65.69 \pm 0.27\%$	$\mathbf{86.76 \pm 0.14\%}$	$\mathbf{88.29 \pm 0.13\%}$	$\mathbf{81.99 \pm 0.77\%}$	$\mathbf{83.90 \pm 0.69\%}$	$\mathbf{84.30 \pm 0.03\%}$

TABLE IV
INDEXES OF 15 BANDS SELECTED BY ONR FOR ALL THE DATA SETS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Indian Pines	3	10	16	44	67	72	77	99	114	131	149	162	172	182	188
Pavia University	15	18	22	28	33	40	47	54	62	68	72	76	81	90	100
Salinas	6	14	19	30	44	56	70	103	113	118	133	162	173	185	194
KSC	6	10	15	18	24	28	30	32	34	37	42	52	63	72	77
Botswana	4	12	23	27	33	45	48	50	60	78	82	84	88	98	108
Leaves	26	40	61	98	115	146	172	219	238	251	261	269	281	300	324

TABLE V
THE P-VALUES WHEN COMPARING ONR TO THE OTHER METHODS.

	ONR→UBS	ONR→E-FDPC	ONR→WaLuDi	ONR→NC-OC	ONR→OPBS	ONR→RMBS
p-values	0.0005	0.0122	0.0005	0.0093	0.0049	0.0005

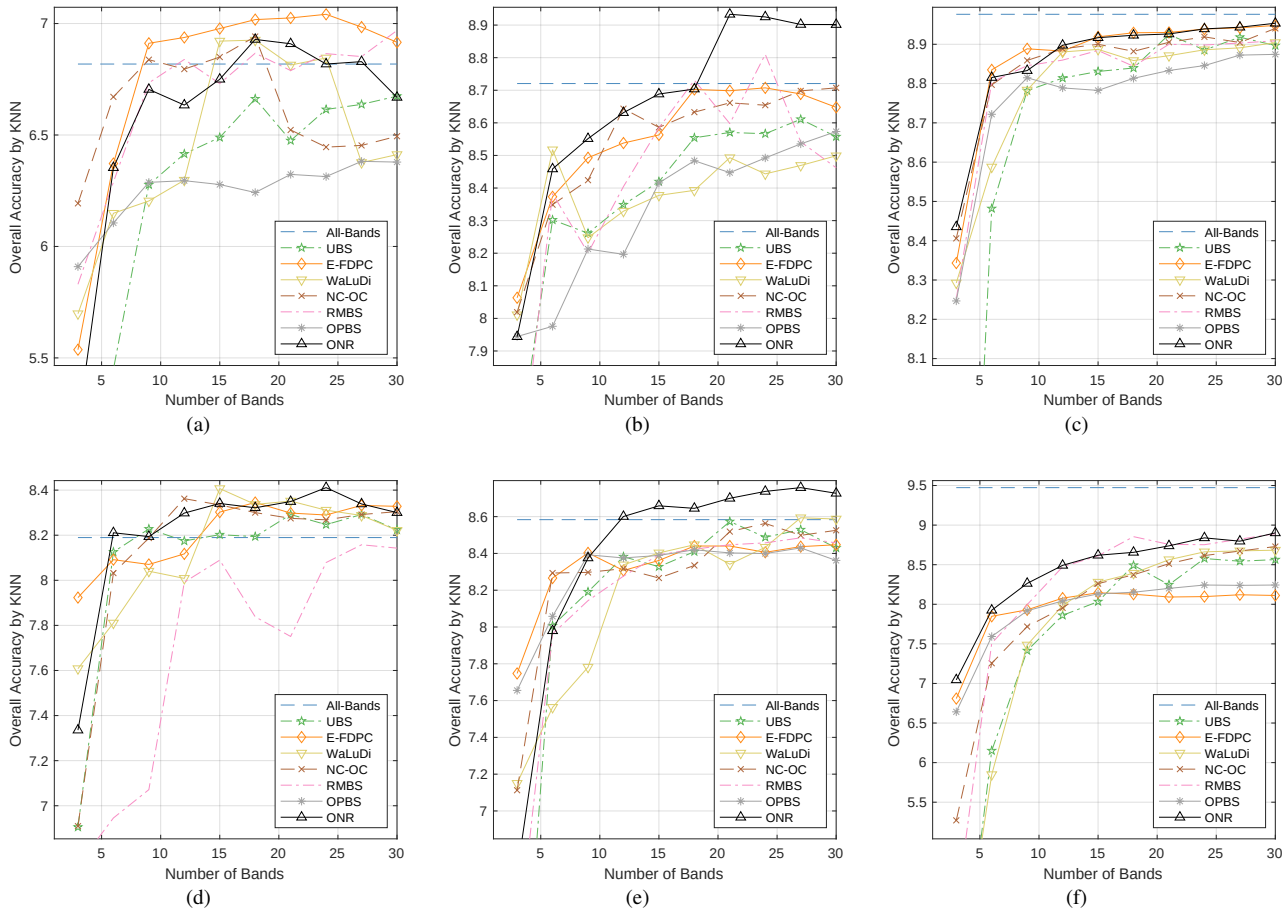


Fig. 5. Overall accuracy curves produced by KNN classifier. (a)-(f) OA curves for Indian Pines, Pavia University, Salinas, KSC, Botswana and Leaves data sets, respectively.

TABLE VI
RUNNING TIME (S) OF DIFFERENT BAND SELECTION METHODS WHEN 30 BANDS ARE SELECTED.
THE LAST TWO COLUMNS ARE THE RUNNING TIMES TO PERFORM SVM CLASSIFICATION USING 30 BANDS AND ALL BANDS.

	E-FDPC	WaLuDi	NC-OC	RMBS	OPBS	ONR	30 Bands SVM	All Bands SVM
Indian Pines	**0.06**	0.76	0.52	34.14	0.69	0.77	2.12	4.84
Pavia University	**0.21**	0.60	0.53	104.52	1.62	0.32	34.10	35.29
Salinas	**0.24**	1.15	0.92	118.96	2.19	0.85	23.10	49.53
KSC	**0.53**	1.68	1.58	324.50	5.49	0.63	10.61	17.18
Botswana	**0.56**	1.58	1.46	217.86	6.39	0.62	5.65	11.73
Leaves	**5.02**	14.64	15.92	4811.20	67.34	8.70	2980.21	9381.37

selection results, which may also improve the robustness of ONR.

3) Drawback of ONR. Although ONR performs well in the experiments, its main drawback is that it can not give a precise measurement towards the information implied in noisy bands. Noisy bands are not always useless, they may also be beneficial in some cases. But since they usually have little linear relation with the clean bands, ONR will refuse to select them in most of the time. With the growth of m, when we are allowed to select some of the noisy bands to improve the performance, ONR will lose its advantages since it can not distinguish which noisy band is more discriminative. Nevertheless, since the main purpose of band selection is to reduce the size of HSI data as much as possible, ONR is still very useful in related research topics.

V. CONCLUSION

This paper analyses the specificity of band selection compared to traditional feature selection problem, and presents a novel band selection method named ONR. ONR exploits the CNP of HSI data and applies an effective optimization method to achieve the exact solution of a reconstruction-based objective function.

ONR involves an noise identification mechanism to avoid the selection of noisy bands so it can achieve a more stable performance. However, since it cannot measure the information of noisy bands accurately, it will ignore some useful information hidden in them. The experiments demonstrate the

proposed algorithm has superior performance on various data sets compared to the state-of-the-art methods.

In the future, we will focus on how to model the non-linear relationship between noisy and clean bands, and better exploit the useful information.

APPENDIX

Proof to Theroem 1.

$$\begin{aligned} D_{i,j} &= \min_{b_1<\ldots<b_j} \sum_{k=0}^{j-1} S_{b_k,b_{k+1}}, \text{ s.t. } b_j = i \\ &= \min_{b_1<\ldots<b_{j-1}<i} \sum_{k=0}^{j-2} S_{b_k,b_{k+1}} + S_{b_{j-1},i} \\ &= \min_{j-1\le b_{j-1}<i} \min_{b_1<\ldots<b_{j-1}} \sum_{k=0}^{j-2} S_{b_k,b_{k+1}} + S_{b_{j-1},i} \\ &= \min_{j-1\le b_{j-1}<i} D_{b_{j-1},j-1} + S_{b_{j-1},i}. \\ &= \min_{j-1\le k<i} D_{k,j-1} + S_{k,i}. \end{aligned} \tag{21}$$

□

Intuitively, $D_{i,j}$ can be interpreted as the minimal loss to reconstruct the first i bands using j selected bands, satisfying that the ith band is selected. Theorem 1 reveals a recurrence relation that by enumerating all the possible values of b_{j-1}, $D_{i,j}$ can be converted to a series of simpler subproblems, which are to optimally reconstruct the first k bands using $j-1$ selected bands. The basic idea here is dynamic programming [17, 38].

Proof to Theroem 2. To prove Theorem 2, we will show that the following equation holds for $0 \le j \le m$:

$$D_{b^*_{j+1},j+1} = \sum_{k=0}^{j} S_{b^*_k,b^*_{k+1}}. \tag{22}$$

If this is true, we can verify that $\boldsymbol{b}^*$ is the solution to Eq. (9) by simply substituting $j = m$ to Eq. (22) (note that $b_{m+1} = d+1$ and $D_{d+1,m+1}$ is equivalent to Eq. (9)).

The proof is based on mathematical induction. When $j = 0$, Eq. (22) holds according to the definition of $\boldsymbol{D}$. When we assume Eq. (22) holds for $j = i-1$, for its successor $j = i$, we have Eq. (23) according to Theorem 1:

$$D_{b^*_{i+1},i+1} = \min_{i\le k<b^*_{i+1}} D_{k,i} + S_{k,b^*_{i+1}}. \tag{23}$$

Considering the definition of $\boldsymbol{Q}$ and Eq. (14), b^*_i should be one of the optimal arguments to Eq. (23), which means:

$$D_{b^*_{i+1},i+1} = D_{b^*_i,i} + S_{b^*_i,b^*_{i+1}}. \tag{24}$$

According to the induction assumption, there is:

$$D_{b^*_i,i} = \sum_{k=0}^{i-1} S_{b^*_k,b^*_{k+1}}. \tag{25}$$

Substituting Eq. (25) to Eq. (24), we get Eq. (22). □

REFERENCES

[1] H. Akbari, Y. Kosugi, K. Kojima, and N. Tanaka, "Detection and analysis of the intestinal ischemia using visible and invisible hyperspectral imaging," *IEEE Transaction on Biomed. Engineering*, vol. 57, no. 8, pp. 2011–2017, 2010.

[2] A. H. Saputro and W. Handayani, "Wavelength selection in hyperspectral imaging for prediction banana fruit quality," in *2017 International Conference on Electrical Engineering and Informatics (ICELTICs)*, 2017, pp. 226–230.

[3] F. Santini, L. Alberotanza, R. M. Cavalli, and S. Pignatti, "A two-step optimization procedure for assessing water constituent concentrations by hyperspectral remote sensing techniques: An application to the highly turbid venice lagoon waters," *Remote Sensing of Environment*, vol. 114, no. 4, pp. 887–898, 2010.

[4] B. Rasti, M. O. Ulfarsson, and J. R. Sveinsson, "Hyperspectral feature extraction using total variation component analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 6976–6985, 2016.

[5] L. Fang, N. He, S. Li, A. J. Plaza, and J. Plaza, "A new spatial-spectral feature extraction method for hyperspectral images using local covariance matrix representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3534–3546, 2018.

[6] A. Sellami, M. Farah, I. R. Farah, and B. Solaiman, "Hyperspectral imagery semantic interpretation based on adaptive constrained band selection and knowledge extraction techniques," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 4, pp. 1337–1347, 2018.

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

[8] 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 in Appl. Earth Observ. and Remote Sensing*, vol. 9, no. 9, pp. 4374–4388, 2016.

[9] G. Taşkın, H. Kaya, and L. Bruzzone, "Feature selection based on high dimensional model representation for hyperspectral images," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2918–2928, 2017.

[10] X. Cao, C. Wei, Y. Ge, J. Feng, J. Zhao, and L. Jiao, "Semi-supervised hyperspectral band selection based on dynamic classifier selection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 4, pp. 1289–1298, 2019.

[11] Y. Yuan, X. Zheng, and X. Lu, "Discovering diverse subset for unsupervised hyperspectral band selection," *IEEE Transaction on Image Processing*, vol. 26, no. 1, pp. 51–64, 2017.

[12] Q. Wang, Q. Li, and X. Li, "Hyperspectral band selection via adaptive subspace partition strategy," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 12, pp. 4940–4950, 2019.

[13] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Transaction on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1279–1289, 2016.

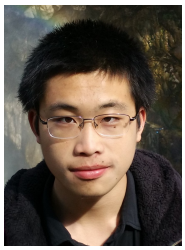
[14] M. Luo, F. Nie, X. Chang, Y. Yi, A. G. Hauptmann, and Q. Zheng, "Adaptive unsupervised feature selection with structure regularization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–13, 2017.

[15] S. Jia, Z. Ji, Y. Qian, and L. Shen, "Unsupervised band selection for hyperspectral imagery classification without manual band removal," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 531–543, 2012.

[16] A. Le Bris, N. Chehata, X. Briottet, and N. Paparoditis, "Extraction of optimal spectral bands using hierarchical band merging out of hyperspectral data," *International Archives of*

the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL-3/W3, pp. 459–465, 08 2015.
[17] A. C. Jensen and A. H. S. Solberg, "Fast hyperspectral feature reduction using piecewise constant function approximations," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, pp. 547–551, 2007.
[18] A. M. Usó, F. Pla, J. M. Sotoca, and P. García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geoscience and Remote Sensing*, vol. 45, no. 12-2, pp. 4158–4171, 2007.
[19] V. Kumar, J. Hahn, and A. M. Zoubir, "Band selection for hyperspectral images based on self-tuning spectral clustering," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 2013, pp. 1–5.
[20] S. X. Yu and J. Shi, "Multiclass spectral clustering," in *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, pp. 313–319 vol.1.
[21] G. Zhu, Y. Huang, S. Li, J. Tang, and D. Liang, "Hyperspectral band selection via rank minimization," *IEEE Geoscience and Remote Sensing Letters*, vol. PP, no. 99, pp. 1–5, 2017.
[22] W. He, H. Zhang, H. Shen, and L. Zhang, "Hyperspectral image denoising using local low-rank matrix recovery and global spatial-spectral total variation," *IEEE J Sel. Topics in Appl. Earth Observ. and Remote Sensing*, vol. 11, no. 3, pp. 713–729, 2018.
[23] W. He, H. Zhang, L. Zhang, and H. Shen, "Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration," *IEEE Trans. Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 178–188, 2016.
[24] Y. Qian, F. Yao, and S. Jia, "Band selection for hyperspectral imagery using affinity propagation," *Computer Vision Iet*, vol. 3, no. 4, pp. 213–222, 2009.
[25] M. Zhang, J. Ma, and M. Gong, "Unsupervised hyperspectral band selection by fuzzy clustering with particle swarm optimization," *IEEE Geoscience and Remote Sensing Letters*, vol. PP, no. 99, pp. 1–5, 2017.
[26] C. Chang, Q. Du, T. Sun, and M. L. G. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 37, no. 6, pp. 2631–2641, 1999.
[27] C. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geoscience and Remote Sensing*, vol. 44, no. 6, pp. 1575–1585, 2006.
[28] X. Geng, K. Sun, L. Ji, and Y. Zhao, "A fast volume-gradient-based band selection method for hyperspectral image," *IEEE Trans. Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 7111–7119, 2014.
[29] H. Su, Q. Du, G. Chen, and P. Du, "Optimized hyperspectral band selection using particle swarm optimization," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2659–2670, 2014.
[30] 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 Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2956–2969, 2015.
[31] 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 Recognition*, vol. 51, pp. 295–309, 2016.
[32] Y. Liu, Y. Chen, K. Tan, H. Xie, L. Wang, X. Yan, W. Xie, and Z. Xu, "Maximum relevance, minimum redundancy band selection based on neighborhood rough set for hyperspectral data classification," *Measurement Science and Technology*, vol. 27, no. 12, p. 125501, nov 2016.
[33] H. Su, B. Yong, and Q. Du, "Hyperspectral band selection using improved firefly algorithm," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 68–72, 2016.
[34] Z. Khan, F. Shafait, and A. Mian, "Joint group sparse pca for compressed hyperspectral imaging," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4934–4942, 2015.
[35] W. Sun and Q. Du, "Graph-regularized fast and robust principal component analysis for hyperspectral band selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–11, 2018.
[36] R. Raz, "On the complexity of matrix product," *Siam Journal on Computing*, vol. 32, no. 5, pp. 1356–1369, 2002.
[37] S. Jia, G. Tang, J. Zhu, and Q. Li, "A novel ranking-based clustering approach for hyperspectral band selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 88–102, 2016.
[38] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Trans. Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5910–5922, 2018.
[39] W. Zhang, X. Li, Y. Dou, and L. Zhao, "A geometry-based band selection approach for hyperspectral image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 1, no. 99, pp. 1–16, 2018.
[40] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transaction on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.
[41] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transaction on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[42] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.



Fahong Zhang received the B.E. degree in software engineering from the Northwestern Polytechnical University, Xi'an, China, in 2017. He is currently working toward the M.S. degree in computer science in the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include hyperspectral image processing and computer vision.

Xuelong Li (M'02-SM'07-F'12) is a Full Professor with the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China.