Semi-supervised Hyperspectral Band Selection via Spectral-spatial Hypergraph Model

Xiao Bai, Zhouxiao Guo, Yanyang Wang, Zhihong Zhang, Jun Zhou

Abstract—Band selection is an essential step towards effective and efficient hyperspectral image classification. Traditional supervised band selection methods are often hindered by the problem of lacking enough training samples. To address this problem, we propose a semi-supervised band selection method that allows contribution from both labelled and unlabelled hyperspectral pixels. This method first builds a hypergraph model from all hyperspectral samples in order to measure the similarity among pixels. We show that hypergraph can captures relationship among pixels in both spectral and spatial domain. In the second step, a semi-supervised learning method is introduced to propagate class labels to unlabelled samples. Then a linear regression model with group sparsity constraint is used for band selection. Finally, hyperspectral pixels with selected bands are used to train a support vector machine classifier. The proposed method is tested on three benchmark datasets. Experimental results demonstrate its advantages over several other band selection methods.

Index Terms—Hyperspectral imaging, Band selection, Hypergraph, Image classification.

I. INTRODUCTION

Hyperspectral imagery has been used in a variety of remote sensing applications. It contains hundreds of spectral bands with nanoscale spectral resolution, which provides valuable information for target detection and classification. However, such abundant wavelength indexed bands also contain redundant or correlated information which influence the effectiveness and efficiency of data modelling [1]. An intuitive way of solving this problem is selecting a subset of bands from the original data [2].

Band selection methods can be categorized as supervised or unsupervised. Supervised learning methods use labelled data to determine the importance of bands. Class labels are often generated by manual annotation or ground survey. The learning outcome can be a searching strategy, or a class separability criterion. For instance, Yang et al. [3] proposed a sequential forward selection (SFS) searching

strategy for band selection. In [4], Rick Archibald et al introduced an embedded feature selection algorithm that is tailored to operate with support vector machines. In [5], bands were selected according to transformed divergence measure and the maximum likelihood accuracy on the test data.

Unsupervised methods, on the contrary, do not need to learn a predictive model from training data. Classic unsupervised methods, for example, principal component analysis [6] and affinity propagation [7], often use a priori knowledge of the scene to measure the statistical dependence between bands in order to evaluate the contribution of each band in the classification. Some methods, such as linear prediction [8] and geometrical feature selection [9], search for distinctive spectral signatures based on different considerations of band properties. Ghamary Asl et al. [10] proposed a method to select distinct and informative bands in a prototype space which is constructed by clustering of raw image data. Bands are selected based on either their orthogonal distance to the diagonal of the prototype space or their angle distance that is related to the correlation between neighboring bands. In unsupervised band selection, searching algorithms are required in most cases [11], which can be speeded up via parallel computing techniques [12]. The evaluation of the effectiveness of band selection approaches can be done at either the band level [13] or the whole data cube level [14].

Despite their advantages in implementation, unsupervised methods can not match the classification accuracy of the supervised methods. As a consequence, semi-supervised band selection methods, which combine the advantages of both supervised and unsupervised methods, have attracted the attention from the research community. Bai et al. [15] proposed a semi-supervised learning method that estimates the vectors of mean values and covariance matrices for each class under the assumption of a Gaussian mixture model. Chen et al combined Fisher's criteria and graph Laplacian to explore both labeled and unlabeled samples simultaneously [16]. Moreover, mutual information has also been used as a measurement criterion via graph regulation term [17]. These methods have shown their effectiveness when the amount of labelled data is insufficient.

Recently, hypergraph based methods have been proposed for pattern recognition [18] and image processing [19]. In [20] and [21], hypergraph construction methods have been introduced for hyperspectral band selection. These methods use spectral mutual information to build a hypergraph which is used for label propagation. However, both

This research is supported under China NSFC Projects (No.61105022) and Australian Research Council's DECRA Projects funding scheme (project ID DE120102948)

Xiao Bai and Zhouxiao Guo are with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China. (e-mail: baixiao.buaa@gmail.com.)

Yanyang Wang is with the School of Aeronautic Science and Engineering, Beihang University, Beijing 100191, China. (e-mail: wangyanyang@buaa.edu.cn.)

Zhihong Zhang is with the Software School of Xiamen University, Siming District, China

Jun Zhou is with the School of Information and Communication Technology, Griffith University, Nathan, Queensland 4111, Australia.

methods have not considered spatial information between pixels in the hyperspectral imagery. In this paper, we propose a semi-supervised band selection method based on hypergraph model which combines both spectral and spatial properties of data. Based on pairwise sample similarity generated from traditional undirected graph, this method first constructs a hypergraph model of all samples, both labeled and unlabeled, in a hyperspectral image. Each vertex of the hypergraph is a pixel, and edges between vertices show the similarity between pixels in both spectral feature space and spatial space. In this way, a matrix can be calculated on the hyperspectral image, which captures multiple relationships between data samples. In the second step, a projection matrix is learned in a semi-supervised manner, which propagates the class information from labeled data to unlabeled data. Band selection is implemented by solving a sparse group constrained linear regression problem on the whole hyperspectral image. This enables selection of not only individual band, but also groups of bands that are useful for image classification. This method is assessed via experiments on three hyperspectral datasets. The results show that joint spectral-spatial hypergraph construction in affinity matrix calculation and the sparse group constraints in band selection can effectively improve the classification performance.

The contributions of this paper lies in four aspects. Firstly, we present a hypergraph model that generalizes the traditional graph model for band selection. We show that hypergraph is a better approach to characterize the similarity between hyperspectral pixels. This lead to more reliable label propagation in semi-supervised band selection. Secondly, we introduce spectral and spatial information into hyperedge calculation. Such joint spectral spatial analysis has demonstrated significant advantages over the traditional spectral only analysis. Thirdly, a sparse group constrained linear regression model is used for band selection. This model selects not only individual bands, but also a group of bands that are discriminative for image classification. Fourthly, we perform comprehensive experiments to compare the proposed method with several alternative methods in the literature. In-depth analysis on key steps is provided, which allows better understanding of the proposed method.

The rest of this paper is organized as follows. Section II introduces the hypergraph construction and sample similarity calculation method. Section III introduces the label propagation and sparse group constrained linear regression methods for band selection. Section IV describes the datasets used in the experiments and the image classification results.

II. HYPERGRAPH MODEL

In this section, we first introduce traditional un-directed graph representation. Then we propose a spectral-spatial hypergraph model to calculate the similarities among hyperspectral image pixels. Hypergraph is a special graph representation where each edge can connect more than two vertices. Compared with traditional un-directed graph,

hypergraph can better characterize the complex relationship inside the hyperspectral data.

A. Graph on Hyperspectral Samples

Given a hyperspectral image $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{N \times M}$ where $\mathbf{x}_i$ represents a vector of spectral responses at a pixel or sample location in a hyperspectral image. N is the number of pixels and M is the number of spectral bands. Let $G = (\mathbb{V}, \mathbb{E}, \mathbf{W})$ be a weighted un-directed graph defined on $\mathbf{X}$. $\mathbb{V}$ is a set of vertices consist of N samples in $\mathbf{X}$, and $\mathbb{E}$ are the edges connecting vertices. $\mathbf{W}$ is a matrix whose entries correspond to the weights of edges, each of which characterizing the similarity between two vertices connected by an edge.

When constructed from a hyperspectral image, the edge weight matrix $\mathbf{W}$ of the graph can be defined as follows:

$$\mathbf{W}_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / \sigma} \quad (1)$$

where σ is a scalar parameter, which is the variance of the average distance between samples. In this definition, $\mathbf{x}_i$ and $\mathbf{x}_j$ can be either labeled or unlabeled samples in $\mathbf{X}$. If two samples are close to each other, i.e., the Euclidean distance between two spectral vectors is small, a large edge weight is generated. Otherwise, a small edge weight is obtained.

B. Hypergraph Model

The weights of graph defined above only describe the relationship between two samples in a hyperspectral imagery. In many cases, the support on similarity computation comes from more than two samples. For example, each pixel in a hyperspectral image may be grouped with its neighboring pixels, or with several very similar pixels in the spectral feature space. Then the similarity calculation between arbitrary pixel pair can be converted to calculating the similarity between two groups of pixels.

To facilitate such group-wise similarity calculation, we introduce the concept of hypergraph. A hypergraph is a generalization of a graph in which an edge can connect more than two vertices. Therefore, it enables better description on the complex relationship between pixels [22].

Let a hypergraph $\tilde{G}$ be defined as $\tilde{G} = (\mathbb{V}, \tilde{\mathbb{E}}, \tilde{\mathbf{W}})$. $\mathbb{V} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ is a set of vertices with the same definition as in the traditional un-directed graph, $\tilde{\mathbb{E}}$ is a set of non-empty subsets of $\mathbb{V}$ called hyperedges. $\tilde{\mathbf{W}}$ is a set of weight functions which correspond to each hyperedge. A hyperedge can connect an arbitrary number of vertices. In this paper, we only consider K -uniform hypergraph in which every hyperedge connects the same number of K vertices, i.e., a hyperedge $\tilde{\mathbb{E}}_i = \{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}\}$. K -uniform is a widely used option of constructing a hypergraph. It is desirable that each hyperedge contains a set of vertices that have some common properties [23]. Given a vertex $\mathbf{x}_i \in \mathbf{X}$, the selection of K vertices to construct $\tilde{\mathbb{E}}_i$ is based on the edge weight matrix $\mathbf{W}$ defined in equation (1). Let $\mathbf{x}_{i_K} = \mathbf{x}_i$, only $K - 1$ vertices with top $\mathbf{W}_{ij}$ values, i.e., those pixels closest to $\mathbf{x}_i$ in the spectral feature space, are selected to form $\{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{K-1}}\}$. Therefore, the number

of hyperedges is the same as the number of vertices. There are two reasons to use top $K - 1$ vertices to construct the hyperedges of a hypergraph. First, top $K - 1$ similar vertices share the highest common spectral property, compared to the rest of vertices. Second, using top $K - 1$ vertices saves the computation time than using all vertices.

Compared with the tradition graph definition, the hyperedge in a hypergraph can connect more than two vertices. Therefore, the definition of the weight for hyperedge should consider the relationship among multiple vertices connected by the edge. In this paper, the weight $\tilde{\mathbf{W}}(i)$ of hyperedge $\tilde{\mathbb{E}}_i$ consists of two terms that incorporate both spatial and spectral information in the hypergraph construction. The first term is $\tilde{\mathbf{W}}_s$ which is calculated after considering the spatial relationship between pixels, and the second term is mutual information $\tilde{\mathbf{W}}_m$ in the spectral feature space.

Formally, the weight of a hyperedge can be defined as

$$\tilde{\mathbf{W}}(i) = t\tilde{\mathbf{W}}_s(i) + (1 - t)\tilde{\mathbf{W}}_m(i) \quad (2)$$

where t is a parameter in $[0, 1]$, which determines the trade-off between spectral and spatial information of the pixels used to construct a hyperedge. Analysis on how t influences the band selection and image classification performance is given in Section IV-A. In practice, the optimal value of t can be determined by cross validation.

The spatial information $\tilde{\mathbf{W}}_s(i)$ of hyperedge $\tilde{\mathbb{E}}_i$ is computed as follows

$$\tilde{\mathbf{W}}_s(i) = \sum_{\{a,b\} \subset \{i_1, \dots, i_K\}} \frac{2w_{a,b}}{K * (K - 1)} \quad (3)$$

where

$$w_{a,b} = e^{-||c_a - c_b||^2 / \sigma} \quad (4)$$

In the above equation, c_a and c_b are the pixel coordinates of two samples indexed by a and b in the hyperedge, $||c_a - c_b||$ is the spatial distance between these two samples, and σ is a scale parameter controlling the decay rate of the distance. The definition of $\tilde{\mathbf{W}}_s(i)$ reflects the overall spatial distance among vertices in a hyperedge. In cases when two pixels have similar spectral information but are distant from each other, according to Equation (4), a small weight $w_{a,b}$ will be generated. Since the spatial information $\tilde{\mathbf{W}}_s(i)$ is summed over all weights as presented in Equation (3), this small weight does not generate significant impact to the final calculation.

The second component in Equation (2) is the mutual information $\tilde{\mathbf{W}}_m(i)$ of hyperspectral pixels, which is computed as follows

$$\tilde{\mathbf{W}}_m(i) = K \frac{I_{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}}}{H(\mathbf{x}_{i_1}) + \dots + H(\mathbf{x}_{i_K})} \quad (5)$$

where $I_{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}}$ is the multivariate mutual information among hyperspectral pixels indexed by $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}$. It can be calculated as follows [24]

$$I_{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}} = \sum_{k=1}^K (-1)^{k-1} \sum_{\substack{\mathbb{X}_s \subset \{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}\} \\ |\mathbb{X}_s| = k}} H(\mathbb{X}_s) \quad (6)$$

In Equation (6), $\mathbb{X}_s$ is a subset of $\{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}\}$, and $H(\mathbb{X}_s)$ represents the Shannon entropy of a discrete random variable with possible values and probability mass function. If $\mathbb{X}_s$ is in the form of single variable $\mathbf{x}_{i_k} = [\mathbf{x}_{i_{k1}}, \dots, \mathbf{x}_{i_{kM}}]^T$, where M is the total number of hyper-spectral bands, the entropy can be explicitly calculated as follows

$$\begin{aligned} H(\mathbf{x}_{i_k}) &= \sum_{m=1}^M P(\mathbf{x}_{i_{km}}) I(\mathbf{x}_{i_{km}}) \\ &= - \sum_{m=1}^M P(\mathbf{x}_{i_{km}}) \log_2(\mathbf{x}_{i_{km}}) \end{aligned} \quad (7)$$

When there are more than two variables, this expands to

$$\begin{aligned} H(\mathbb{X}_s = [\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_d}]) \\ = - \sum_{\mathbf{x}_{i_1}} \dots \sum_{\mathbf{x}_{i_d}} P(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_d}) \log_2[P(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_d})] \end{aligned} \quad (8)$$

where d is the number of vertices in random variable $\mathbb{X}_s$. $P(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_d})$ is the probability of $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_d}$ occurring together. $P(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_d}) \log_2[P(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_d})]$ is defined to be 0 if $P(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_d}) = 0$.

It is clear that the greater the value of $I_{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}}$ is, the more relevant the K vertices are.

Algorithm 1: Hypergraph construction algorithm

Data: Dataset $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{N \times M}$ with each row being a spectral feature vector in the hyperspectral imagery.

Compute similarity matrix: $\mathbf{W}_{ij} = e^{-||\mathbf{x}_i - \mathbf{x}_j||^2 / \sigma}$;

for $i = 1$ to N **do**

1. Arrange $\mathbf{W}_{ij}$ in descending order;
2. Select the first $K - 1$ vertices and $\mathbf{x}_i$ to construct a hyperedge $\tilde{\mathbb{E}}_i = \{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}\}$;
3. Calculate $\tilde{\mathbf{W}}_s(i)$ using Equation (3);
4. Calculate $\tilde{\mathbf{W}}_m(i)$ using Equation (5);
5. Calculate $\tilde{\mathbf{W}}(i) = t\tilde{\mathbf{W}}_s(i) + (1 - t)\tilde{\mathbf{W}}_m(i)$;

end

The hypergraph construction method is summarized in Fig. 1 and Algorithm 1. There is no widely accepted method to handle the operation of hypergraph weight matrix. In our method, we convert the hypergraph weight matrix into an affinity matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ in order to facilitate the computation in the later steps. Weights of all hyperedges containing $\mathbf{x}_a$ and $\mathbf{x}_b$ are added up to generate the weight of $\mathbf{M}_{a,b}$. Therefore, $\mathbf{M}_{a,b}$ is computed as

$$\begin{aligned} \mathbf{M}_{a,b} &= \sum_i \tilde{\mathbf{W}}(i) \\ \text{s.t. } \mathbf{x}_a, \mathbf{x}_b &\in \tilde{\mathbb{E}}_i \end{aligned} \quad (9)$$

To show how the entries of the affinity matrix are calculated, we give the following example. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two pixels whose affinity is to be calculated. Assume $K = 4$, i.e., a hyperedge contains four vertices. If three hyperedges

Fig. 1. Construction of hypergraph.

contain both $\mathbf{x}_1$ and $\mathbf{x}_2$, i.e., $\tilde{\mathbb{E}}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$, $\tilde{\mathbb{E}}_2 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5\}$, $\tilde{\mathbb{E}}_3 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_5, \mathbf{x}_7\}$, $\tilde{\mathbb{E}}_4 = \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_7\}$, then $\mathbf{M}_{1,2} = \tilde{\mathbf{W}}_1 + \tilde{\mathbf{W}}_2 + \tilde{\mathbf{W}}_3$.

III. BAND SELECTION AND IMAGE CLASSIFICATION

In this section, we move on to the band selection method. After the hypergraph has been transformed into an affinity matrix, our goal is to convert the high dimensional data $\mathbf{X} \in \mathbb{R}^{N \times M}$ to low dimensional data $\mathbf{Y} \in \mathbb{R}^{N \times D}$ ($D \ll M$), where M and D are the number of bands before and after band selection, respectively. This dimensionality reduction problem can be modeled as a special subspace learning task, where the projection matrix is constrained to be a selection matrix $\mathbf{S}$, such that $\mathbf{Y} = \mathbf{XS}$. To utilize both labeled and unlabeled data, a semi-supervised learning method, which consists of label propagation and selection matrix learning, is introduced. Finally, all bands are sorted based on sparse matrix $\mathbf{S}$, and the top bands are selected as the final outcome. In summary, this stage has four steps, which include label propagation, sparse matrix $\mathbf{S}$ estimation, band selection, and image classification.

A. Label Propagation

Given a dataset of pixels $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_l, \dots, \mathbf{x}_N] \in \mathbb{R}^{N \times M}$ with l labeled samples and $u = N - l$ unlabeled samples, the goal of label propagation is to estimate which classes the unlabeled samples belong to.

We adopted the label propagation method by Nie et al. [25], which not only achieves the goal of semi-supervised learning, but also discovers the latent novel class in the data. Define an initial label matrix $\mathbf{L} \in \mathbb{R}^{N \times (C+1)}$, where C is the number of classes of the data. An additional class label $C+1$ is added to record the outliers. The initial value of label matrix is defined as follows:

$$\mathbf{L}_{i,j} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ is labeled as } l_i = j \text{ or } j = C+1 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Let $\mathbf{F} \in \mathbb{R}^{N \times (C+1)}$ denotes the set of classification function defined on $\mathbf{X}$. Since $\mathbf{F}$ works on both labelled and unlabelled data, it can be considered as a label propagation matrix, i.e., $\mathbf{F}_{ij}$ represents the probability that samples i belong to class j . In every propagation step, each data absorbs

a fraction of the label information from its neighborhood and also retains some label information of its initial state. To propagate labels, we define a matrix $\mathbf{P}$ as $\mathbf{P} = \mathbf{D}^{-1}\mathbf{M}$, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{M}_{i,j}$, and $\mathbf{M}$ is the affinity matrix calculated from hypergraph. This is equivalent to the normalization process for matrix $\mathbf{M}$. Then the label of $\mathbf{x}_i$ at time $t+1$ can be updated from its previous estimation $\mathbf{F}(t)$

$$\mathbf{F}(t+1) = \lambda \mathbf{P} \mathbf{F}(t) + (1-\lambda) \mathbf{L}. \quad (11)$$

where λ is a parameter in $[0, 1]$. The above steps iterate until convergence. Let $\mathbf{F}^*$ denotes the limit of the sequence $\mathbf{F}(t)$, the final estimation function $\mathbf{F}^*$ can be computed directly using an analytical solution without iterations

$$\mathbf{F}^* = \lim_{t \rightarrow \infty} \mathbf{F}(t) = (1-\lambda)(\mathbf{I} - \lambda \mathbf{P})^{-1} \mathbf{L} \quad (12)$$

where $\mathbf{I}$ is an identity matrix. Hence, we can use $\mathbf{F}^*$ as the classification function so that the labels of hyperspectral samples can be predicted in one step.

B. Sparse Matrix $\mathbf{S}$

After the label propagation matrix $\mathbf{F}^*$ is calculated, the next step is to learn the sparse selection matrix $\mathbf{S}$. Here we build a linear classifier $\mathbf{y} = \mathbf{S}^T \mathbf{x} + b$ where $\mathbf{x}$ is a sample in the hyperspectral imagery and b is a bias term. If $\mathbf{y}$ is close to $\mathbf{g}_j$ where $\mathbf{g}_j = [0, \dots, 0, 1, 0, \dots, 0]^T$ which means the j th element in $\mathbf{g}_j$ is 1, $\mathbf{x}$ will be classified into class j . Assuming that the result of classification from the linear classifier equals the above semi-supervised learning matrix $\mathbf{F}^*$, we can define a regression function as follows

$$\arg \min_{\mathbf{S}, b} \sum_{i=1}^N \sum_{j=1}^C \mathbf{F}_{i,j}^* \|\mathbf{S}^T \mathbf{x}_i + b - t_j\|^2 \quad (13)$$

In order to find an optimal subset of bands, we adopt the Lasso method [26] which is an L_1 -constrained fitting solution for statistics and data mining. It minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. This forces many entries of $\mathbf{S}$ to be zeros.

Considering that neighboring spectral bands may be highly correlated, we also group the neighboring bands and

solve the band selection problem using the group LASSO method [27]. To do so, we uniformly divide the spectral bands into R groups. Then the group Lasso imposes a combination of L_1 and L_2 penalties on the regression coefficients. Therefore, the loss function is designed to eliminate the impact of the outlier as well as to achieve the minimal reconstruction error, while encouraging only some bands in the useful band groups be selected.

The regression function can be written as

$$\arg \min_{\mathbf{S}, b} \sum_{i=1}^N \sum_{j=1}^C \mathbf{F}_{i,j}^* ||S^T \mathbf{x}_i + b - t_j||^2 + \theta \sum_{k=1}^R d_k |S_k|_2 \quad (14)$$

where $|S_k|_2$ is given below

$$|S_k|_2 = \left(\sum_{j \in k} S_j^2 \right)^{\frac{1}{2}} \quad (15)$$

where $\mathbf{F}_{i,j}^*$ is the element from the label propagation matrix in Equation (12) and θ is a parameter in $[0, 1]$. d_k is a set of strictly positive fixed weights. Therefore, the sum over R groups is essentially the L_1 penalty term of the group sparsity.

C. Band selection

Note the goal of band selection is to select D bands from a total of M band candidates. Here, we use band score to evaluate every band from the selection matrix $\mathbf{S}$, which is defined as

$$bandscore(j) = \max_k |S_{j,k}| \quad (16)$$

where j is the band index, $S_{j,k}$ is the j th row of $\mathbf{S}_k$. After getting scores of all the bands, we sort the bands in descending order and select the top D bands as the outcome.

D. Image Classification

Finally, we perform image classification based on the selected bands which are the same for all samples. A support vector machine (SVM) classifier is applied to the selected bands of each sample to be classified in the hyperspectral remote sensing images. The classification process is carried out for each sample individually. SVM is used here because of its good performance in the nonlinear separable problem and small training sample sets. In this step, the Gaussian radial basis function (RBF) [28] is chosen as the kernel function, which is defined as

$$K_{\mathbf{F}}(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}||) \quad (17)$$

where γ is a parameter inversely proportional to the width of the Gaussian kernel. $\mathbf{x}$ and $\mathbf{x}_i$ are samples and support vectors whose dimension is D .

IV. EXPERIMENTS

Having presented our method in the previous sections, we now turn our attention to demonstrating its utility for classification. Three real-world hyperspectral image datasets have been used for experiments. They are APhi dataset, Indian Pines dataset, and Washington DC Mall dataset. We compared the overall classification accuracy of our method with those from several alternative methods in the literature. These methods include band selection methods based on semi-supervised learning, unsupervised learning, and uniform sampling, as follows

- Uniform sampling (Uniform) which selects bands uniformly.
- Maximum-Variance PCA (MVPCA) [6] which ranks bands based on their importance and their correlation with other bands.
- Mutual Information-based Information Divergence (ID) [2] which selects bands based on mutual information.
- FCGL [29] which is a semi-supervised algorithm that combines Fisher's criteria and graph Laplacian to explore both labeled and unlabeled samples at the same time.

In the experiments, the values of some key parameters were determined by cross validation on the training data. In particular, $K = 9$ in Equation (9), $\lambda = 0.92$ in Equation (12), and $\theta = 0.25$ in Equation (14). The kernel parameter γ in Equation (17) was obtained by cross-validation on a subset of the training samples [30]. In the group LASSO solution for the regression model in Equation (13), each group consisted of 20 neighboring bands.

A. APhi Dataset

The APhi (Airborne Push hyperspectral Imager) data is $210 \times 150 \times 64$ in size and covers 455nm to 805nm in spectral range. It was acquired from an altitude of approximately 20km with 18m in spatial resolution. The number of bands in the APhi dataset is 64. Fig. 2 displays the 26th band of APhi imagery. The ground truth of APhi data is shown in Fig. 3. The land cover types of APhi dataset are paddy, bamboo, tea, pachyrhizus, caraway and water, respectively.

Fig. 2. Data cub of APhi dataset.

Fig. 3. Ground truth of APhi dataset.

The performance of band selection is evaluated by the classification accuracy. To this end, both labeled and unlabeled samples were randomly selected to train a semi-supervised classifier. Labels of samples were firstly propagated to unlabeled samples, which were then used to train an SVM classifier. In the APhi dataset, we randomly selected 340 labeled samples and 2410 unlabeled samples for label propagation and classifier learning and testing. Table I displays the detailed number of samples for each class during the label propagation and classification step.

TABLE I
NUMBER OF LABELED AND UNLABELED SAMPLES FROM APhi DATASET USED IN THE EXPERIMENTS.

Class	Labeled samples	Unlabeled samples
Paddy	100	800
Bamboo	50	280
Tea	50	280
Pachyrhizus	30	250
Caraway	50	200
Water	60	800
Total	340	2410

In the label propagation step, for all 6 land-cover classes, 12% samples of each class were randomly selected as the labeled samples and the remaining ones were used as the unlabeled samples. Both labeled and unlabeled samples were randomly chosen from the dataset for five times. The same training and testing samples were also used in the classification step.

Fig. 4 presents the classification results compared with other band selection methods. To investigate the impact of different number of selected bands on the classification accuracy, we varied this number from 5 to 25 and compared the accuracy of different methods. It can be seen that our method has outperformed all alternative methods, followed by FCGL and MVPCA. Uniform sampling method has delivered the lowest classification accuracy. This shows the effectiveness of band selection from our method. When the number of bands increases, the classification accuracy also increases for each method. The proposed hypergraph based method is close to convergence when the band number

reaches 16. Table II lists the selected bands by four methods when the number of bands is set to 10.

TABLE II
SELECTED BANDS FROM APhi DATASET BY DIFFERENT METHODS.

MVPCA	8,15,20,21,26 37,54,55,59,60
Uniform	1,8,15,22,29 36,43,50,57,64
FCGL	1,10,20,21,26 37,39,52,54,60
Our method	1,3,10,20,26 28,32,39,52,59

Fig. 4. Classification accuracies on the APhi dataset by different methods.

Parameter t in Equation (2) controls the contribution from the spatial and spectral information in the hyper-edge weight calculation. To evaluate such contribution, we calculated the classification accuracies on the APhi data with changing t values, as shown in Fig. 5. This figure shows that the best classification performance is reached when $t = 0.425$. It suggests that both spatial and spectral information are important in the hypergraph construction step. This is reasonable because spatial and spectral relationships are two key relationships between hyperspectral pixels.

B. Indian Pines AVIRIS Dataset

Indian Pines AVIRIS dataset was collected by Airborne Visible/Infrared Imaging Spectrometer. We considered 9 land-cover classes which are corn-notill, corn-min, grass/pasture, grass/tree, hay-windrowed, soybeans-min, soybeans-notill, soybeans-clean, and woods. A pseudo-color image constructed by bands 26, 100, and 47 is shown in Fig. 6 whose ground truth is displayed in Fig. 7. After removing water absorption and noisy bands, 200 bands were used for the experiments.

On the Indian Pines dataset, we randomly selected 1081 samples as the training data and 5407 samples as the testing data. This process repeated 10 times to generate ten different random splits of datasets. The number of training and test samples in each class are shown in Table III.

Fig. 5. Classification accuracy on the APhi dataset with different contribution from spectral and spatial information in hypergraph construction.

Fig. 6. Pseudo-color image constructed from bands 26, 100 and 47 in the Indian Pines dataset.

Fig. 7. Ground truth of the Indian pines dataset.

TABLE III
NUMBER OF TRAINING AND TESTING SAMPLES FOR EACH CLASS IN THE INDIAN PINES DATASET.

Class	training samples	testing samples
Corn-notill	100	900
Corn-min	111	556
Grass/Pasture	140	280
Grass/Tree	100	250
Hay-windrowed	80	280
Soybeans-notill	100	800
Soybeans-min	300	1500
Soybeans-clean	50	410
Woods	100	800
Total	1081	5407

Table IV lists the indices of selected bands when the number of bands is set to 10. The table shows that some bands are unanimously selected by different band selection

methods, for example, band 11 in the Indian Pines data, which suggests that these bands are the most informative bands for hyperspectral classification.

TABLE IV
SELECTED BANDS FROM THE INDIAN PINES DATASET BY DIFFERENT METHODS.

MVPCA	1,11,26,32,35 97,37,184,185,186
Uniform	1,21,41,61,81 101,121,141,161,181
FCGL	4,11,15,21,29 79,82,94,128,186
Our method	2,7,11,25,32 61,76,90,128,134

Fig. 8 shows the change of accuracies when different number of bands are selected. Similarly, the best result is produced when the number of selected band exceeds 20. It can be seen that the proposed method has consistently outperformed MVPCA, Uniform, FCGL and ID methods. This figure also shows that when selecting more than 20 bands, the gain on the accuracy is not as significant as the gain on smaller number of bands. This suggests that 20 bands can be considered as the preferred number of bands to be selected.

Fig. 8. Classification results on the Indian Pines dataset when the number of selected bands varies.

C. Washington DC Mall dataset

We also performed experiments on the airborne hyperspectral HYDICE data captured at the Washington DC mall. It has 210 spectral bands from 400nm to 2500nm, and 1208 scan lines with 307 pixels in each scan line. The totally available classes are nine. Fig. 9 displays a pseudo-color image of the dataset and Fig. 10 shows the region we selected for experiments and the ground truth labels. In the experiments, we used 6 classes which are road, grass, shadow, trail, tree, and roof, respectively. The numbers of labeled samples and unlabeled samples in each class are shown in Table V. These were used to train the band selection methods and the SVM classifiers.

Fig. 11 shows the classification accuracy of each band selection method. Obviously, compared with other band

Fig. 9. Pseudo-color image of the HYDICE data on Washington DC Mall.

Fig. 10. Selected region and ground truth of the Washington DC Mall dataset.

TABLE V
NUMBER OF LABELED AND UNLABELED SAMPLES IN THE
WASHINGTON DC MALL DATASET.

Class	Labeled samples	Unlabeled samples
Road	50	200
Grass	100	900
Shadow	20	100
Trail	100	800
Tree	80	360
Roof	100	400
Total	450	2760

TABLE VI
SELECTED BANDS FROM THE WASHINGTON DC MALL DATASET BY
DIFFERENT METHODS.

MVPCA	3,9,13,46,64 89,150,152,182,184
Uniform	21,41,61,81,101 121,141,161,181,201
FCGL	8,9,36,88,89 151,152,180,182,184
Our method	3,9,39,64,88 89,152,180,182,184

selection methods, which include both semi-supervised and unsupervised band selection options, our method can better predict the class labels of samples in the hyperspectral imagery. We can also observe that, when the number of selected band is the same, our method yields the best results in most of the cases. Note that the accuracy increases with more bands being selected. Our method converges when 30 bands are selected. The top 10 selected bands on this dataset are shown in Table VI.

Fig. 11. Classification accuracy on the Washington DC Mall dataset.

D. Efficiency comparison

Now we illustrate the computational costs of our method and other band selection methods. The codes for all methods were run on a laptop with Intel(R) Core(TM) i5 CPU. Table VII reports the time cost of each method on three benchmark datasets. From the table, we can see that ID is the fastest method and all other methods are much more time-consuming. This is because the ID method is an unsupervised method without spending time on label propagation. Our method is superior than the semi-supervised band selection method FCGL and unsupervised method MVPCA on both Indian Pines and APMI datasets. However, it is slower than others on the Washington DC Mall dataset. This is mainly because the construction of

Fig. 12. Correlation measures between the selected bands produced by ID, MVPCA, FCGL, and our method on three datasets. From left to right are the results on the APHI, the Indian Pines, and the Washington DC Mall datasets.

hypergraph requires larger number of testing and training samplings operations on the Washington DC Mall dataset than on the other two datasets.

TABLE VII
COMPUTATIONAL COST OF DIFFERENT BAND SELECTION METHODS

Time cost	ID	MVPCA	FCGL	Our method
Indian Pines	10s	3m10s	2m4s	2m2s
APHI	7s	2m6s	2m	1m25s
Washington DC Mall	11s	3m4s	4m	5m25s

E. Group Sparsity

We conducted an experiment to compare the basic sparsity and group sparsity methods on classification accuracy. For the group sparsity option, we uniformly divided the bands into groups of 20 neighboring bands and calculate the group regression coefficients. Then the top 10 bands were selected for the classification.

TABLE VIII
GROUP SPARSITY AND SPARSITY CLASSIFICATION RESULTS WITH 10 BANDS

Classification Accuracy	Sparsity	Group Sparsity
APHI	86.1%	86.4%
Indian Pines	66.4%	67.3%
Washington DC Mall	66.8%	67.8%

From Table VIII, we can see that the group sparsity method is slightly better than the basic sparsity method. This is because the group sparsity method not only determines which groups of bands shall be selected, but also further identified the most important bands from each selected group.

F. Graph vs Hypergraph

To compare the performance of using normal un-directed graph or hypergraph for classification, we run experiments on both Indian Pines and APHI datasets with these two options. The difference relies on Equation (9) where for normal un-directed graph we used $\mathbf{W}$ instead of $\bar{\mathbf{W}}$ to calculate matrix $\mathbf{M}$. Then both graph and hypergraph based similarity measures were combined with label propagation methods based on class extension as introduced in

TABLE IX
COMPARISON OF CLASSIFICATION RESULTS WHEN USING GRAPH AND HYPERGRAPH WITH 10 BANDS.

Method	Graph-Semi	Hypergraph-Semi
Indian Pines dataset	68.9%	75.5%
APHI dataset	70.5%	78.3%

Section III. 2000 samples were randomly selected from 5 classes for training and testing on each dataset. We only conducted experiments on the training datasets which contain labelled samples. 15% of samples in the training datasets were used as labeled samples, and the rest 85% were used as unlabeled samples in the process of label propagation. The accuracies of different method-dataset combinations are shown in Table IX. Hypergraph method is significantly superior than the traditional graph method in classification. This is due to the better measurement of similarities between hyperspectral pixels via capturing of both spectral and spatial relationship between multiple samples.

G. Correlation between bands

Finally, we evaluated the correlations of bands selected by different methods. Specifically, we used average correlation coefficient to assess the discriminant validity. Let $\mathbf{X}_a$ and $\mathbf{X}_b$ be two selected bands, their average correlation coefficient is defined as

$$r_{\mathbf{X}_a, \mathbf{X}_b} = \frac{\sum_{i=1}^N (\mathbf{X}_a(i) - \bar{\mathbf{X}}_a)(\mathbf{X}_b(i) - \bar{\mathbf{X}}_b)}{\sqrt{\sum_{i=1}^N (\mathbf{X}_a(i) - \bar{\mathbf{X}}_a)^2 \sum_{i=1}^N (\mathbf{X}_b(i) - \bar{\mathbf{X}}_b)^2}} \quad (18)$$

where $\bar{\mathbf{X}}_a$ and $\bar{\mathbf{X}}_b$ are the sample means of $\mathbf{X}_a$ and $\mathbf{X}_b$, respectively.

Fig. 12 shows the average correlation coefficients of ID, MVPCA, FCGL, and our method on the APHI, the Indian Pines, and the Washington DC Mall datasets, when the number of selected bands changes. From this figure we can see that the correlations of selected bands by our method are lower than all other methods in most conditions. This shows the effectiveness of our method which allows more independent bands be selected so that better classification performance can be achieved.

V. DISCUSSION AND CONCLUSION

In this paper, we have introduced a semi-supervised learning method as a useful means for band selection in hyperspectral imagery. With the introduction of hypergraph for hyperspectral pixel similarity calculation, better relationship between multiple samples can be captured than using the traditional graph method. We also show that distance in both spectral and spatial spaces can contribute to the construction of hypergraph. Finally, a label propagation method has been introduced in this paper for band selection with group sparsity constraint. The proposed method has been compared with several other methods in the literature. The experimental results have shown the advantages of the proposed method in terms of classification accuracy. In the future, we plan to further explore the structure of hypergraph so as to reduce the computation cost. Meanwhile, feature fusion will also be incorporated into the classifiers.

REFERENCES

- [1] P. Bajcsy and P. Groves, "Methodology for hyperspectral band selection," *Photogrammetric Engineering and Remote Sensing*, vol. 70, pp. 793–802, 2004.
- [2] C. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 6, pp. 1575–1585, 2006.
- [3] H. Yang, Q. Du, H. Su, and Y. Sheng, "An efficient method for supervised hyperspectral band selection," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 138–142, 2011.
- [4] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 4, pp. 674–677, 2007.
- [5] M. Riedmann and E.J. Milton, "Supervised band selection for optimal use of data from airborne hyperspectral sensors," in *IEEE International Geoscience and Remote Sensing Symposium*, 2003, pp. 1770–1772.
- [6] C. Chang, Q. Du, T. Sun, and M. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 6, pp. 2631–2641, 1999.
- [7] Y. Qian, F. Yao, and S. Jia, "Band selection for hyperspectral imagery using affinity propagation," *IET Computer Vision*, vol. 3, no. 4, pp. 213–222, 2009.
- [8] Q. Du and H. Yang, "A similarity-based unsupervised based band selection for hyperspectral image analysis," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 4, pp. 564–568, Oct. 2008.
- [9] X. Jia L. Wang and Y. Zhang, "A novel geometry-based feature-selection technique for hyperspectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 1, pp. 171–175, Jan. 2007.
- [10] M. Reza M. Ghamary and B. Mojaradi, "Unsupervised feature selection using geometrical measures in prototype space for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 7, pp. 3774–3787, July 2014.
- [11] P. Mitra, C. Murthy, and S. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301–312, 2002.
- [12] H. Yang, Q. Du, and G. Chen, "Unsupervised hyperspectral band selection using graphics processing units," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 3, pp. 660–668, 2011.
- [13] Y. Yuan, G. Zhu, and Q. Wang, "Hyperspectral band selection by multitask sparsity pursuit," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 2, pp. 631–644, Feb 2015.
- [14] K. Sun, X. Geng, L. Ji, and Y. Lu, "A new band selection method for hyperspectral image based on data quality," *IEEE Journal Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2697–2703, June 2014.
- [15] J. Bai, S. Xiang, and C. Pan, "Classification oriented semi-supervised band selection for hyperspectral images," in *Proceedings of the International Conference on Pattern Recognition*, 2012, pp. 1888–1891.
- [16] L. Chen, R. Huang, and W. Huang, "Graph-based semi-supervised weighted band selection for classification of hyperspectral data," in *2010 International Conference on Audio Language and Image Processing*, IEEE, 2010, pp. 1123–1126.
- [17] J. Feng, L. Jiao, X. Zhang, and T. Sun, "Hyperspectral band selection based on trivariate mutual information and clonal selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 7, pp. 4092–4105, July 2014.
- [18] R. C. Wilson P. Ren, T. Aleksic and E. R. Hancock, "A polynomial characterization of hypergraphs using the ihara zeta function," *Pattern Recognition*, vol. 44, no. 9, pp. 1941–1957, 2011.
- [19] R. C. Wilson X. Bai and E. R. Hancock, "Graph characteristics from the heat kernel trace," *Pattern Recognition*, vol. 42, pp. 2589–2606, 2009.
- [20] Z. Guo, X. Bai, Z. Zhang, and J. Zhou, "A hypergraph based semi-supervised band selection method for hyperspectral image classification," in *IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 3137–3141.
- [21] Z. Guo, H. Yang, X. Bai, Z. Zhang, and J. Zhou, "Semi-supervised hyperspectral band selection via sparse linear regression and hypergraph models," in *Geoscience and Remote Sensing Symposium (IGARSS)*, 2013, pp. 1474–1477.
- [22] S. Bulò and M. Pelillo, "A game-theoretic approach to hypergraph clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1312–1327, 2013.
- [23] X. Li, W. Hu, C. Shen, A. Dick, and Z. Zhang, "Context-aware hypergraph construction for robust spectral clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2588–2597, 2014.
- [24] T. Han, "Multiple mutual informations and multiple interactions in frequency data," *Information and Control*, vol. 46, no. 1, pp. 26–45, 1980.
- [25] F. Nie, S. Xiang, Y. Liu, and C. Zhang, "A general graph-based semi-supervised learning with novel class discovery," *Neural Computing and Applications*, vol. 19, no. 4, pp. 549–555, 2010.
- [26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B*, pp. 267–288, 1996.
- [27] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society B*, vol. 68, no. 1, pp. 49–67, 2006.
- [28] M. Pal and G. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2297–2307, 2010.
- [29] R. Huang L. Chen and W. Huang, "Graph-based semi-supervised weighted band selection for classification of hyperspectral data," *IEEE international conference on Audio Language and Image Processing (ICALIP)*, , no. 1123–1126, 2010.
- [30] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.