

Article

# Discriminant Analysis with Graph Learning for Hyperspectral Image Classification

Mulin Chen <sup>1</sup>, Qi Wang <sup>1,2,\*</sup> and Xuelong Li <sup>3,4</sup>

<sup>1</sup> School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, 710072, China

<sup>2</sup> Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an, 710072, China

<sup>3</sup> Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, 710119, China

<sup>4</sup> University of Chinese Academy of Sciences, Beijing, 100049, China

\* Correspondence: crabwq@gmail.com; Tel.: +86-029 8849-5715-8002

Academic Editor: name

Version April 19, 2019 submitted to MDPI

**Abstract:** Linear Discriminant Analysis (LDA) is a widely-used technique for dimensionality reduction, and have been applied in many practical applications, such as hyperspectral image classification. Traditional LDA assumes that the data obeys the Gaussian distribution. However, in real-world situations, the high-dimensional data may be with various kinds of distributions, which restricts the performance of LDA. To reduce this problem, we propose the *Discriminant Analysis with Graph Learning* (DAGL) method in this paper. Without any assumption on the data distribution, the proposed method learns the local data relationship adaptively during the optimization. The main contributions of this research are threefold: (1) the local data manifold is captured by learning the data graph adaptively in the subspace; (2) the spatial information within the hyperspectral image is utilized with a regularization term; (3) an efficient algorithm is designed to optimize the proposed problem with proved convergence. Experimental results on hyperspectral image datasets show that promising performance of the proposed method, and validates its superiority over the state-of-the-arts.

**Keywords:** Hyperspectral image classification; Linear Discriminant Analysis; Graph learning; Sparse learning

## 1. Introduction

Hyperspectral Image (HSI) provides hundreds of spectral bands for each pixel and conveys numerous surface information. Hyperspectral image classification aims to distinguish the land-cover types of each pixel, and the spectral bands are considered as features. However, the great number of bands significantly increases the computational complexity [1]. Moreover, some bands are highly correlated, leading to the feature redundancy problem. Consequently, it is critical to perform dimensionality reduction before classification. The goal of dimensionality reduction is to project the original data into a low-dimensional subspace while preserving the valuable information.

Dimensionality reduction techniques can be roughly classified into two categories: feature selection [2,3] and feature extraction [4–9]. Feature selection methods select the most relevant feature subset from the original feature space, while feature extraction methods exploit the low-dimensional subspace which contains valuable information. Compared to feature selection, feature extraction is able to create meaningful features through the transformation of the original ones. Consequently, plenty of techniques have been put forward on feature extraction [9–13]. Principal Component Analysis (PCA) [14] and Linear Discriminant Analysis (LDA) [15] are the most popular feature extraction

31 methods. PCA learns the feature subspace by maximizing the variance of the feature matrix. While  
32 LDA learns a linear transformation that minimizes the within-class distance and maximizes the  
33 between-class discrepancy. In this research, we mainly focus on LDA because it is able to use the prior  
34 knowledge and shows better performance in real-world applications [11].

35 Though achieving good performance in many tasks, LDA has four major drawbacks on processing  
36 HSI data. Firstly, LDA suffers from the *ill-posed* problem [12]. LDA needs to compute the inverse  
37 matrix of the within-class scatter  $S_w$ . When the data dimensionality exceeds the number of training  
38 samples,  $S_w$  is irreversible. So LDA cannot handle the HSI data with great number of spectral  
39 bands. Secondly, the feature dimensionality reduced by LDA is less than the class number, namely  
40 *over-reducing* problem [13]. Taking KSC dataset [16] for example, the class number is thirteen, and the  
41 rank of the between-class scatter  $S_b$  is at most twelve. So LDA could find at most twelve projection  
42 directions, which may be insufficient for retaining the useful information. Thirdly, LDA neglects  
43 the spatial smoothness. In HSIs, the pixels within a spatial neighborhood region usually belong to  
44 the same class. But LDA just focuses on the pixels' distances in the feature space, and ignores the  
45 spatial aspect. Fourthly, LDA assumes that the data samples are Gaussian-distributed, and share equal  
46 covariances in all the classes. However, HSI data seldom obeys the Gaussian distribution [20], and the  
47 local data structure may be inconsistent with the global structure. Therefore, LDA is unable to find the  
48 local classification boundary for the HSI data.

49 In the past decades, many variants of the original LDA are proposed, trying to enhance its  
50 performance from different views. Bandos et al. [18] proposed the Regularized LDA (RLDA), which  
51 employs a regularized within-class scatter to tackle the ill-posed problem. Kumar and Agrawal [19]  
52 presented the two-dimensional exponential discriminant analysis for data with small sample size. The  
53 Semi-supervised Discriminant Analysis (SDA) method [21] utilizes the unlabelled data to extend the  
54 training set. Wan et al. [13] and Nie et al. [10] developed the full rank between-class scatter matrix  
55 to mitigate the over-reducing problem. To enforce the spatial consistency, Yuan et al. [22] and Wang  
56 et al. [23] constructed a scatter matrix from a small neighborhood, and took it as a regularization  
57 term. With the above methods, the ill-posed and over-reducing problem are alleviated, and the spatial  
58 correlation between pixels can be preserved. However, the exploration of the local data structure  
59 remains to be an open issue. Some graph-based methods [10,11,24,29] defined the scatter matrices  
60 according to the predefined affinity graph, which may be seriously affected by the noise. Ly et al. [37]  
61 performed graph learning and discriminant analysis separately, so the data graph is also fixed during  
62 the optimization. Recently, Wang et al. [9] and Wu et al. [31] proposed to learn the data graph in the  
63 subspace. However, they neglect the similar samples from different classes, which largely determine  
64 the classification boundary.

65 In this work, we propose a new method for supervised dimensionality reduction, termed as  
66 *Discriminant Analysis with Graph Learning* (DAGL). In order to exploit the data structure, the proposed  
67 method learns the data graph adaptively during learning of transformation matrix. And to guarantee  
68 the spatial smoothness, the samples within a small region are encouraged to share the same class  
69 label. With the proposed objective function, the proposed method does not have the ill-posed and  
70 over-reducing problem. The contributions made in paper are summarized as follows:

71 (1) The affinity graph is built according to the samples' distances in the subspace, so the local data  
72 structure is captured adaptively.

73 (2) The proposed formulation perceives the spatial correlation within HSI data, and avoids the  
74 ill-posed and over-reducing problem naturally.

75 (3) An alternative optimization algorithm is developed to solve the proposed problem, and its  
76 convergence is proved experimentally.

## 77 2. Linear Discriminant Analysis Revisited

In this section, the Linear Discriminant Analysis is briefly reviewed as the preliminary. Given an input data matrix  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$  ( $d$  is the data dimensionality and  $n$  is the number of samples), LDA defines the between-class scatter  $\mathbf{S}_b$  and within-class scatter  $\mathbf{S}_w$  as

$$\begin{aligned} \mathbf{S}_b &= \sum_{k=1}^c n_k (\mu_k - \mu)(\mu_k - \mu)^T, \\ \mathbf{S}_w &= \sum_{k=1}^c \sum_{\mathbf{x}_i \in k} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T, \end{aligned} \quad (1)$$

where  $n_k$  is the sample number in class  $k$ ,  $c$  is the class number,  $\mu_k$  is the mean of samples in class  $k$  and  $\mu$  is the mean of all the samples. With the above definitions, LDA aims to learn a linear transformation  $\mathbf{W} \in \mathbb{R}^{d \times m}$  ( $m \ll d$ ), which maximizes the between-class difference while minimizing the within-class separation:

$$\max_{\mathbf{W}} \text{Tr} \left( \frac{\mathbf{W}^T \mathbf{S}_b \mathbf{W}}{\mathbf{W}^T \mathbf{S}_w \mathbf{W}} \right), \quad (2)$$

78 where  $\text{Tr}()$  indicates the trace operator. With the optimal transformation  $\mathbf{W}^*$ , data sample  $\mathbf{x}_i$  can be  
79 projected to a  $m$ -dimensional feature vector  $\mathbf{W}^{*T} \mathbf{x}_i$ .

80 As shown in Eq. (1), LDA assumes that the data distribution is Gaussian and the between-class  
81 divergence can be reflected by the subtraction of the mean. This assumption is unsuitable for HSI data,  
82 and makes LDA insensitive to the local manifold.

## 83 3. Discriminant Analysis with Graph Learning

84 In this section, the Discriminant Analysis with Graph Learning (DAGL) method is introduced,  
85 and an optimization method is proposed to get the optimal solution.

### 86 3.1. Graph learning

87 In real-world tasks, such as HSI classification, the local manifold may be inconsistent with the  
88 global structure. So it is necessary to take the local data relationship into consideration.

In the past decades, numerous algorithms are proposed to explore the data structure. Some of them [25–28] first construct an affinity graph with various kernels (Gaussian kernel, linear kernel, 0-1 weighting), and then perform clustering or classification according to the spectral of the predefined graph. However, the choice of kernel scales and categories is still an open issue. Therefore, the graph learning methods [32–37] are developed to learn the data graph automatically. One of the most popular graph learning techniques is Sparse Representation [32,33], which aims to learn a sparse graph from the original data. Sparse Representation assumes that a data sample can be roughly represented by the linear combination of the others. Defining a coefficient matrix  $\mathbf{S} \in \mathbb{R}^{n \times n}$ , the optimal  $\mathbf{S}$  should minimize the reconstruction error as follows

$$\begin{aligned} \min_{\mathbf{S}} \sum_{j=1}^n \left\| \mathbf{x}_j - \sum_{i \neq j} \mathbf{x}_i \mathbf{S}_{ij} \right\|_2^2 \\ = \min_{\mathbf{S}} \|\mathbf{X} - \mathbf{X}\mathbf{S}\|_F^2. \\ \text{s.t. } \sum_{i=1}^n \mathbf{S}_{ij} = 1, \mathbf{S} \geq 0, \end{aligned} \quad (3)$$

89 If  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are similar,  $\mathbf{S}_{ij}$  will be large. So  $\mathbf{S}$  can be considered as the affinity graph.

### 3.2. Methodology

As shown in Eq. (3), Sparse Representation exploits the data relationship in the original data space, and the data noise may affect the graph quality adversely. To reduce this problem, we propose to adjust the data graph during the discriminant analysis, which yields to the following formula

$$\min_{\mathbf{W}, \mathbf{S}} \frac{\sum_{k=1}^c \sum_{x_i, x_j \in k} \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2}{\sum_{k=1}^c \sum_{x_i \in k, x_j \notin k} \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2} + \alpha \|\mathbf{X} - \mathbf{XS}\|_F^2, \quad (4)$$

$$s.t. \mathbf{W}^T \mathbf{W} = \mathbf{I}, \sum_{i=1}^n \mathbf{S}_{ij} = 1, \mathbf{S} \geq 0,$$

where  $\mathbf{I} \in \mathbb{R}^{m \times m}$  is the identity matrix, and  $\alpha$  is a parameter. When the linear transformation  $\mathbf{W}$  is learned, the first term of problem (4) enforces  $\mathbf{S}_{ij}$  to be small/large for the within/between-class samples with large transformed distances. In this way, the data graph is optimized in the subspace. Similarly, when  $\mathbf{S}$  is fixed, the transformed distance  $\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2$  will be small/large for the within/between-class samples with large  $\mathbf{S}_{ij}$ . Consequently, the within-/between-class similar samples are ensured to be close/far away in the transformed subspace. However, it is difficult to optimize problem (4) directly because  $\mathbf{S}$  is involved in both the numerator and denominator of the first term. Supposing the minimum value of the first term is  $\gamma$ , the optimal  $\mathbf{W}$  and  $\mathbf{S}$  should make the value of  $\sum_{k=1}^c \sum_{x_i, x_j \in k} \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2 - \gamma \sum_{k=1}^c \sum_{x_i \in k, x_j \notin k} \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2$  to be close to 0. So problem (4) is equivalent to the following formula

$$\min_{\mathbf{W}, \mathbf{S}} \sum_{k=1}^c \sum_{x_i, x_j \in k} \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2 - \gamma \sum_{k=1}^c \sum_{x_i \in k, x_j \notin k} \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2 + \alpha \|\mathbf{X} - \mathbf{XS}\|_F^2, \quad (5)$$

$$s.t. \mathbf{W}^T \mathbf{W} = \mathbf{I}, \sum_{i=1}^n \mathbf{S}_{ij} = 1, \mathbf{S} \geq 0,$$

where  $\gamma$  can be set as a small value. Denoting a class indicator matrix  $\mathbf{Z} \in \mathbb{R}^{n \times n}$  as

$$\mathbf{Z}_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are from the same class,} \\ -\gamma, & \text{else} \end{cases}. \quad (6)$$

problem (5) can be simplified into

$$\min_{\mathbf{W}, \mathbf{S}} \sum_{i,j=1}^n \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2 \mathbf{Z}_{ij} + \alpha \|\mathbf{X} - \mathbf{XS}\|_F^2, \quad (7)$$

$$s.t. \mathbf{W}^T \mathbf{W} = \mathbf{I}, \sum_{i=1}^n \mathbf{S}_{ij} = 1, \mathbf{S} \geq 0,$$

In HSI data, the pixels within a small region may be highly correlated and belong to the same class. The spatial information is essential for an accurate classification. Given a test sample  $\mathbf{t} \in \mathbb{R}^{d \times 1}$ , we find its surroundings within a  $r \times r$  region, and denote them as  $[\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_{r^2-1}]$ . For these samples, we encourage them to be close to each other in the desired subspace, which yields to the following problem

$$\min_{\mathbf{W}} \sum_{i,j=1}^{r^2-1} \mathbf{W}^T \|\mathbf{t}_i - \mathbf{t}_j\|_2^2. \quad (8)$$

Problem (8) can be further reduced to

$$\min_{\mathbf{W}} \text{Tr}(\mathbf{W}^T \tilde{\mathbf{S}}_t \mathbf{W}), \quad (9)$$

where

$$\tilde{\mathbf{S}}_t = \sum_{i,j=1}^{r^2-1} (\mathbf{t}_i - \mathbf{t}_j)(\mathbf{t}_i - \mathbf{t}_j)^T. \quad (10)$$

Finally, by integrating problem (7) and (9) together, we have the objective function of the proposed DAGL method:

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{S}} \sum_{i,j=1}^n \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2 \mathbf{Z}_{ij} + \alpha \|\mathbf{X} - \mathbf{X}\mathbf{S}\|_F^2 + \beta \text{Tr}(\mathbf{W}^T \tilde{\mathbf{S}}_t \mathbf{W}), \\ \text{s.t. } \mathbf{W}^T \mathbf{W} = \mathbf{I}, \sum_{i=1}^n \mathbf{S}_{ij} = 1, \mathbf{S} \geq 0, \end{aligned} \quad (11)$$

91 where  $\alpha$  and  $\beta$  are parameters. Since DAGL does not need to calculate the inverse matrix of within-class  
 92 scatter, the ill-posed problem is avoided naturally. In addition, the projected dimensionality  $m$  can  
 93 be any value less than  $d$ , so the over-reducing problem does not occur. With the proposed objective  
 94 function, the local data relationship is investigated, and the spatial correlation between the pixels is  
 95 also captured.

### 96 3.3. Optimization Algorithm

97 Problem (11) involves two variables to be optimized, so we consider to fix one and update another  
 98 one iteratively. The data graph  $\mathbf{S}$  is firstly initialized with an efficient method [34].

**When  $\mathbf{S}$  is fixed**, problem (11) becomes

$$\begin{aligned} \min_{\mathbf{W}} \sum_{i,j=1}^n \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2 \mathbf{Z}_{ij} + \beta \text{Tr}(\mathbf{W}^T \tilde{\mathbf{S}}_t \mathbf{W}), \\ \text{s.t. } \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned} \quad (12)$$

Denoting a scatter matrix  $\tilde{\mathbf{S}}_z$  as

$$\tilde{\mathbf{S}}_z = \sum_{i,j=1}^n (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}_{ij}^2 \mathbf{Z}_{ij}, \quad (13)$$

problem (12) is converted into

$$\min_{\mathbf{W}^T \mathbf{W} = \mathbf{I}} \text{Tr}[\mathbf{W}^T (\tilde{\mathbf{S}}_z + \beta \tilde{\mathbf{S}}_t) \mathbf{W}]. \quad (14)$$

99 According to the spectral clustering [38], the optimal  $\mathbf{W}$  for problem (14) is formed by the  $m$   
 100 eigenvectors of matrix  $(\tilde{\mathbf{S}}_z + \beta \tilde{\mathbf{S}}_t)$  corresponding to the  $m$  smallest eigenvalues.

**When  $\mathbf{W}$  is fixed**, by removing the irrelevant terms, problem (11) is transformed into

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{S}} \sum_{i,j=1}^n \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2 \mathbf{Z}_{ij} + \alpha \|\mathbf{X} - \mathbf{X}\mathbf{S}\|_F^2, \\ \text{s.t. } \sum_{i=1}^n \mathbf{S}_{ij} = 1, \mathbf{S} \geq 0. \end{aligned} \quad (15)$$

Fixing the diagonal elements in  $\mathbf{S}$  as 0, the above problem is equivalent to

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{S}} \sum_{i,j=1}^n \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2 \mathbf{Z}_{ij} + \alpha \sum_{j=1}^n \|\mathbf{x}_j - \mathbf{X}\mathbf{s}_j\|_2^2, \\ \text{s.t. } \sum_{i=1}^n \mathbf{S}_{ij} = 1, \mathbf{S} \geq 0, \end{aligned} \quad (16)$$

where  $\mathbf{s}_j \in \mathbb{R}^{n \times 1}$  is the  $j$ -th column of  $\mathbf{S}$ . Since the  $\mathbf{s}_j$  is independent between different  $j$ , we can solve the following problem for each  $j$ :

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{S}} \sum_{i=1}^n \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{S}_{ij}^2 \mathbf{Z}_{ij} + \alpha \|\mathbf{x}_j - \mathbf{X}\mathbf{s}_j\|_2^2, \\ \text{s.t. } \sum_{i=1}^n \mathbf{S}_{ij} = 1, \mathbf{S} \geq 0. \end{aligned} \quad (17)$$

Defining a diagonal matrix  $\mathbf{U} \in \mathbb{R}^{n \times n}$  with  $\mathbf{U}_{ii} = \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{Z}_{ij}$ , we further arrive:

$$\begin{aligned} \min_{\mathbf{s}_j} \mathbf{s}_j^T (\mathbf{U} + \mathbf{X}^T \mathbf{X}) \mathbf{s}_j - 2\alpha \mathbf{s}_j^T \mathbf{X}^T \mathbf{x}_j, \\ \text{s.t. } \mathbf{s}_j^T \mathbf{1} = 1, \mathbf{s}_j \geq 0, \end{aligned} \quad (18)$$

101 where  $\mathbf{1} \in \mathbb{R}^{n \times 1}$  is a column vector with all the elements equal to 1. Because  $(\mathbf{U} + \mathbf{X}^T \mathbf{X})$  is a positive  
102 definite matrix, problem (18) can be readily solved by the Augmented Lagrange Method (ALM) [39].

103 In the above optimization procedure, the original problem (11) is decomposed into two  
104 sub-problems. When solving  $\mathbf{W}$ , a local optimal value is obtained. When solving  $\mathbf{S}$ , the ALM  
105 algorithm is employed, whose convergence is already proved. So the objective value of problem (11)  
106 decreases monotonically in each iteration, and finally converges to a local optimum. The convergence  
107 behaviour of the proposed algorithm will be shown in Section 4.3. The details of the whole framework  
108 is described in Algorithm 1.

---

#### Algorithm 1 Discriminant Analysis with Graph Learning

---

**Input:** training set, testing set, parameter  $K, r, \alpha$  and  $\beta$ .

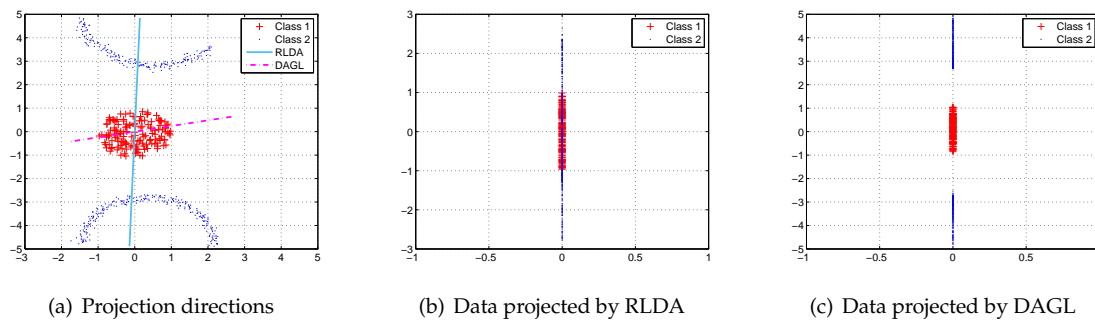
- 1 **For** each test sample:
- 2     Construct the training sub-set  $\mathbf{X}$  by choosing the  $K$  nearest neighbors from the training set.
- 3     Find the surroundings of the test sample within the  $r \times r$  region, and obtain  $\tilde{\mathbf{S}}_t$ .
- 4     Initialize data graph  $\mathbf{S}$ .
- 5     **Repeat:**
- 6         Update  $\mathbf{W}$  by minimizing problem (14).
- 7         Update  $\mathbf{S}$  by solving problem (18).
- 8     **Until** converge
- 9     **End**

**Output:** optimal transformation matrix  $\mathbf{W}^*$  for each test sample.

---

## 109 4. Experiments

110 In this section, experiments are conducted on one toy and two hyperspectral image datasets. The  
111 convergence behavior and parameter sensitivity of the proposed method are also discussed.



**Figure 1.** (a) Projection directions found by RLDA and DAGL. (b) One-dimensional data projected by RLDA. (c) One-dimensional data projected by DAGL. For a better illustration, the projected data is plotted in the plane coordinate system, and the horizon coordinate of the projected data is set as zero.

#### 112 4.1. Performance on Toy Dataset

113 A toy dataset is constructed to demonstrate that the proposed DAGL can captures the local data  
114 structure.

115 **Dataset:** as visualized in Fig. 1 (a), the toy dataset consists of two-dimensional samples from two  
116 classes. Samples from the first class obey the Gaussian distribution, and those from the second class  
117 are distributed in the two-moon shape. The coordinates of the samples are taken as the features.

118 **Performance:** We transform the samples into the one-dimensional subspace with regularized  
119 Linear Discriminant Analysis (RLDA) [18] and the proposed DAGL. And for DAGL,  $\beta$  is set as 0 since  
120 spatial distance is equivalent to the feature distance. Fig. 1 (a) shows the learned projection directions.  
121 It is manifest that DAGL finds the correct projection direction successfully, while LDA fails. On this  
122 dataset, the local data structure is inconsistent with the global structure, and the mean values of the  
123 samples cannot reflect their real relationship. So RLDA is unable to project the data correctly, as shown  
124 in Fig. 1 (b). On the other hand, the proposed DAGL does not rely on any assumption on the data  
125 distribution, and learns the local data manifold adaptively, so it finds discriminative subspace, where  
126 the samples are linearly separable, as shown in Fig. 1 (c).

#### 127 4.2. Performance on Hyperspectral Image Datasets

128 In this part, experiments are conducted on hyperspectral image datasets. The data samples are  
129 projected into the subspace, and then classified by the SVM classifier. The parameters of SVM are  
130 selected by grid search within  $\{2^0, 2^1, \dots, 2^{10}\}$  and  $\{2^0, 2^1, \dots, 2^{20}\}$ . Three widely-used measurements,  
131 overall accuracy (OA), average accuracy (OA) and kappa statics ( $\kappa$ ) are adopted as evaluation criteria.

132 **Datasets:** two hyperspectral image datasets are employed in the experiments, including *Indian*  
133 *Pines* and *KSC* [16] datasets.

134 *Indian Pines* dataset was captured by AVIRIS sensor over the northwestern Indiana, and annotates  
135 10249 pixels from 16 classes. Each pixel is with 220 spectral bands. In the experiments, only 200 bands  
136 are used because the other 20 bands are affected by water absorption. The spatial resolution of this  
137 dataset is 20 m.

138 *KSC* dataset was captured by AVIRIS sensor over the Kennedy Space Center (KSC), Florida. After  
139 removing the water absorption and low SNR bands, there remains 176 bands. 5211 pixels from 13  
140 classes, which represent the various land cover types, are used for classification.

141 For each dataset, we randomly select 5% samples as training set and all the remaining samples as  
142 test set. To alleviate the random error caused by the dataset partition, we repeating the experiments for  
143 five times and report the average results. The sizes of the training and test sets for the two datasets are  
144 exhibited in Table 1 and 2. Through experiments, we have found that a small portion of the training

145 set is enough for a good performance. When classifying a test sample, we just select its 50 nearest  
 146 neighbors (in feature space) from the training set, and use them to train the proposed DAGL model.

**Table 1.** Number of training and test samples for each class on Indian Pines dataset.

No.	Class	Training	Test	No.	Class	Training	Test
1	Alfalfa	3	51	9	Oats	1	19
2	Corn-notill	72	1362	10	Soybeans-notill	49	914
3	Corn-mintill	40	741	11	Soybeans-mintill	122	2304
4	Corn	12	222	12	Soybeans-clean	31	582
5	Grass-pasture	24	451	13	Wheat	11	201
6	Grass-tree	38	709	14	Woods	65	1229
7	Grass-pasture-mowed	2	24	15	Bldg-grass-tree-drives	17	315
8	Hay-windrowed	25	464	16	Stone-steel-towers	5	90

**Table 2.** Number of training and test samples for each class on KSC dataset.

No.	Class	Training	Test	No.	Class	Training	Test
1	Scurb	38	719	8	Graminoid-marsh	22	405
2	Willow-swamp	13	230	9	Spartina-marsh	26	494
3	Cabbage-palm-hammock	13	243	10	Cattail-marsh	21	383
4	Cabbage-palm/oak-hammock	13	239	11	Salt-marsh	21	398
5	Slash-pine	9	152	12	Mud-flats	26	477
6	Oak/broadleaf-hammock	12	217	13	Water	47	880
7	Hardwood-swamp	6	99				

147 **Competitors:** for a quantitative comparison, six dimensionality reduction algorithms are taken  
 148 as competitors, including regularized LDA (RLDA) [18], Semi-supervised Discriminant Analysis  
 149 (SDA) [21], Block Collaborative Graph-based Discriminant Analysis (BCGDA) [37], Spectral-Spatial  
 150 LDA (SSLDA) [22], and Locality Adaptive Discriminant Analysis (LADA) [23]. To demonstrate the  
 151 usefulness of dimensionality reduction, the classification result with all features is taken as the baseline,  
 152 termed as RAW.

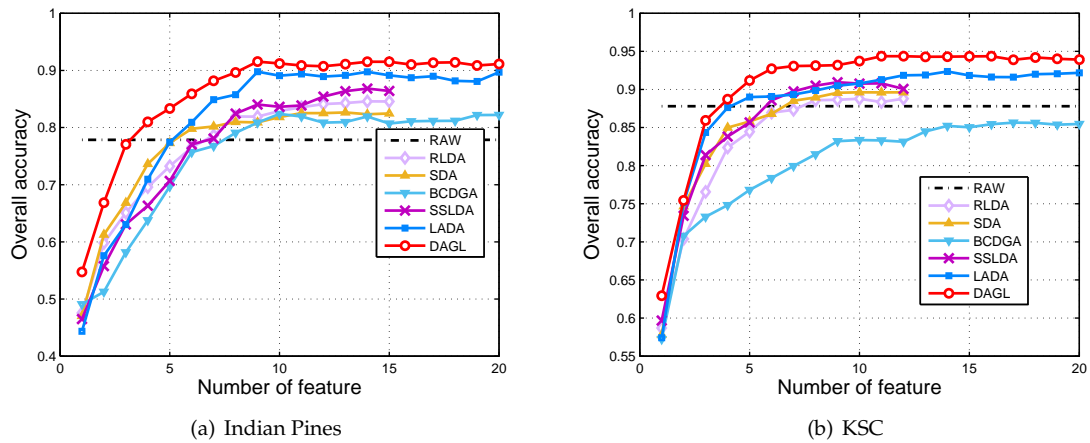
153 The parameter of RLDA is searched in the range of  $\{10^{-3}, 10^{-2}, \dots, 1\}$ . For SDA, the  
 154 parameter is searched in  $\{0.1, 0.5, 2.5, 12.5\}$ . The parameters of BCGDA, SSLDA and LADA are  
 155 selected in  $\{10^{-3}, 10^{-2}, \dots, 10^3\}$ . For DAGL,  $\alpha$  and  $\beta$  are searched in  $\{10^{-3}, 10^{-2}, \dots, 10^3\}$  and  
 156  $\{10^{-4}, 10^{-2}, \dots, 10^4\}$  respectively, and the size of the neighborhood  $r$  is set as 5 empirically.

157 **Performance:** each method is performed with different reduced dimensionality. The reduced  
 158 dimensionality of BCGDA, LADA and DAGL varies within the range of  $\{1, 2, \dots, 20\}$ . Because  
 159 LDA, RLDA, SDA and SSLDA have the over-reducing problem, the dimensionality varies within  
 160  $\{1, 2, \dots, 15\}$  and  $\{1, 2, \dots, 12\}$  on Indian Pines and KSC respectively.

161 The curves of OA versus the reduced dimensionality on different datasets are shown in Fig. 2. The  
 162 proposed DAGL achieves the highest OA constantly. Especially, on Indian Pines dataset, DAGL  
 163 exceeds the second best one to a large extent when the reduced dimensionality is less than 4.  
 164 In Fig. 2, the performance becomes stable when the dimensionality increases to a certain value.  
 165 This phenomenon implies that a low-dimensional subspace is sufficient for sustaining the valuable  
 166 information. Compared with RAW, the performance with projected data is better in most cases, which  
 167 validates that dimensionality reduction does improve the classification accuracy.

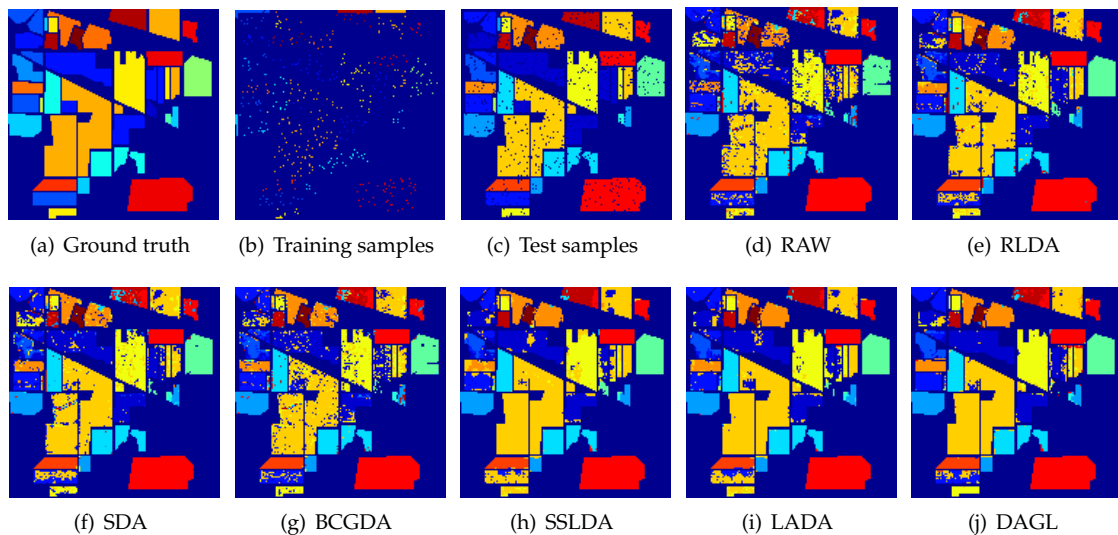
168 The quantitative results of the methods are given in Table 3 and 4. Each method uses its optimal  
 169 reduced dimensionality. It can be seen that DAGL outperforms all the competitors in terms of OA, AA  
 170 and  $\kappa$ . RLDA neglects the local data relationship, so it cannot captures the manifold structure. SDA





**Figure 2.** OA versus the reduced dimensionality of different methods on (a) Indian Pines and (b) KSC datasets.

171 and SSLDA preserve the local data relationship with a predefined data graph. But their performance  
 172 may be adversely affected by the graph quality. BCDGA learns the affinity graph with the original  
 173 data by sparse representation. Because the data graph is fixed during the discriminant analysis, the  
 174 data relationship in the desired subspace cannot be exploited. LADA does not have this problem since  
 175 it integrates graph learning and discriminant analysis jointly. However, it just learns the within-class  
 176 correlation and fails to discover the similar samples from different classes. The proposed DAGL  
 177 investigates the local data relationship adaptively, and pushes the between-class similar samples apart.  
 178 Therefore, it achieves the best performance on all occasions.



**Figure 3.** Classification maps for the Indian Pines dataset with different dimensionality reduction methods.

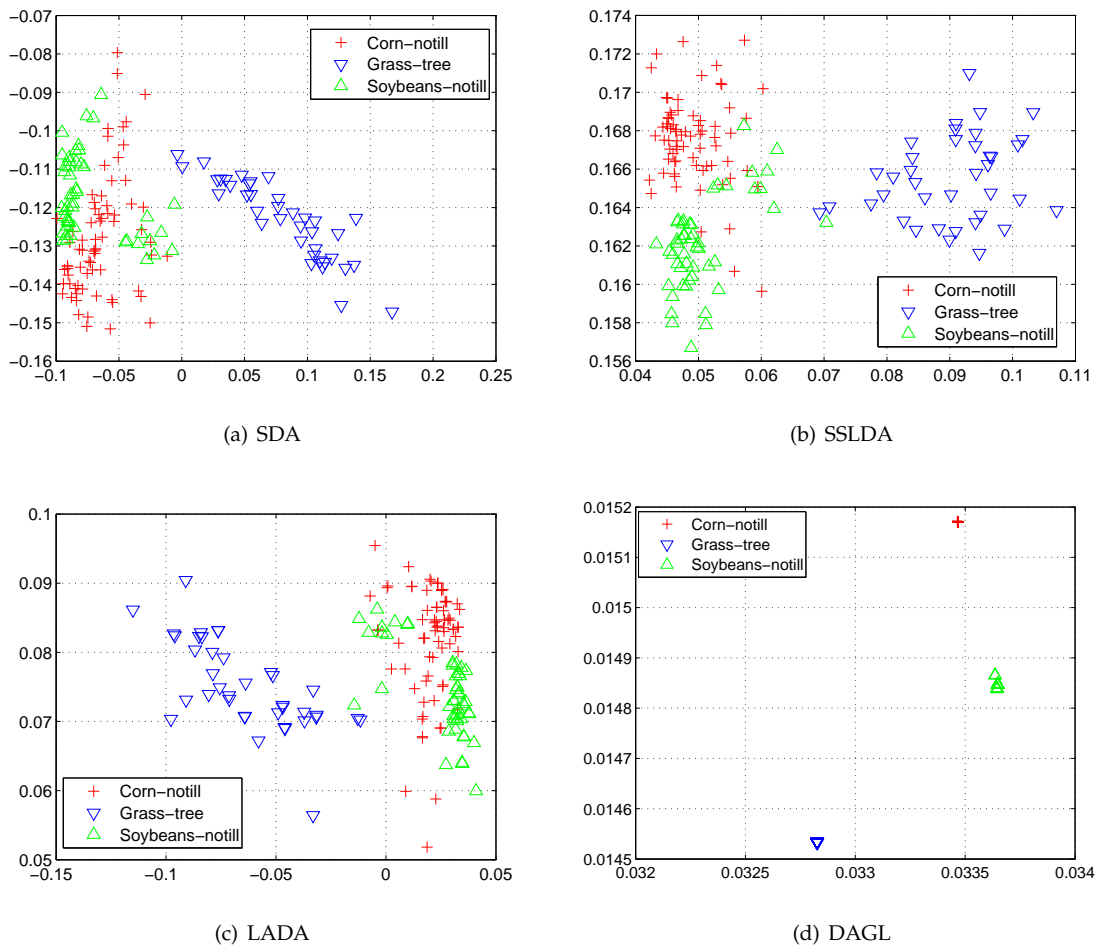
179 Furthermore, the classification maps of different methods on Indian Pines are also visualized  
 180 in Fig. 3. SSLDA, LADA and DAGL, which enforce the spatial smoothness within a small region,  
 181 show better visualization quality than the others. Thus, the utilization of spatial information improves  
 182 the classification performance. It is worthwhile to mention that the methods with spatial constraints  
 183 are time-consuming, as shown in Table 3 and 4, since they need to find the surroundings and train

**Table 3.** Performance of different methods on Indian Pines image (with the best reduced dimensionality in brackets). Best results are in bold face.

Class	RAW(200)	RLDA(14)	SDA(13)	BCDGA(10)	SSLDA(14)	LADA(9)	DAGL(9)
1	0.5678	0.6471	0.5686	0.5963	0.5490	0.7059	<b>0.7182</b>
2	0.7089	0.7880	0.7819	0.7010	0.8047	0.8333	<b>0.8624</b>
3	0.6699	0.7760	0.6802	0.7092	0.6802	0.7395	<b>0.8062</b>
4	0.4240	0.5315	0.4550	0.5460	0.6441	0.6937	<b>0.8668</b>
5	0.8514	0.8847	0.9135	0.8947	0.9246	0.9290	<b>0.9379</b>
6	0.9163	0.9506	0.9661	0.9463	<b>0.9803</b>	0.9746	0.9790
7	0.5017	<b>0.6250</b>	0.4583	0.5767	0.4167	0.5417	0.5616
8	0.9384	0.9397	0.9784	0.9340	0.9921	0.9978	<b>0.9987</b>
9	0.2337	0.3158	0.3682	0.3258	0.2632	<b>0.3684</b>	0.2778
10	0.8003	0.8077	0.7713	0.7682	0.7954	0.8884	<b>0.9163</b>
11	0.8003	0.8720	0.8342	0.8578	0.9353	0.9280	<b>0.9722</b>
12	0.7246	0.8176	0.7984	0.8024	0.8608	0.9038	<b>0.9224</b>
13	0.9451	0.9712	0.9795	0.9300	0.9453	0.9403	<b>0.9917</b>
14	0.9231	0.987	0.9756	0.9813	0.987	0.9837	0.9894
15	0.4357	0.6222	0.4921	0.5556	0.7746	<b>0.8889</b>	0.7111
16	0.9056	0.8444	<b>0.9667</b>	0.9256	0.8222	0.7667	0.6889
OA	0.7785	0.8458	0.8265	0.8239	0.8683	0.8978	<b>0.9153</b>
AA	0.7092	0.7738	0.7493	0.7532	0.7735	0.8177	<b>0.8250</b>
kappa	0.7538	0.8203	0.7999	0.7933	0.8663	0.8830	<b>0.9088</b>
training time	0s	0.10s	0.25s	0.72s	1105.23s	3703.62s	553.08s

**Table 4.** Performance of different methods on KSC image (with the best reduced dimensionality in brackets). Best results are in bold face.

Class	RAW(200)	RLDA(14)	SDA(13)	BCDGA(10)	SSLDA(14)	LADA(9)	DAGL(9)
1	0.9316	0.9179	0.9082	0.9224	0.9972	0.9861	<b>0.9986</b>
2	0.8204	0.8043	<b>0.8913</b>	0.8187	0.9130	0.8870	0.8739
3	0.8995	0.7860	0.8971	0.7466	0.9712	0.9424	<b>0.9918</b>
4	0.7933	0.7448	0.7573	0.5037	0.7397	0.7448	<b>0.9833</b>
5	0.4768	0.5855	0.6447	0.6153	<b>0.7303</b>	0.7237	0.7239
6	0.5753	0.8065	0.7558	0.5814	<b>0.9078</b>	0.8894	0.8295
7	0.7476	0.6364	0.6768	0.6363	0.9192	0.8586	<b>0.9293</b>
8	0.8542	<b>0.8963</b>	0.8593	0.8520	0.8148	0.8148	0.8347
9	0.9515	0.9737	0.9757	0.9047	0.9774	0.9974	<b>0.9981</b>
10	0.9143	0.9556	0.9269	0.9291	<b>0.9869</b>	0.9661	0.9765
11	0.9397	0.9749	0.9347	0.9798	0.9917	0.9935	<b>0.9975</b>
12	0.8412	0.8050	<b>0.8616</b>	0.7857	0.7894	0.8470	0.8423
13	0.9707	0.9773	0.9886	0.9986	0.8966	0.9989	<b>0.9994</b>
OA	0.8780	0.8880	0.8963	0.8564	0.9094	0.9236	<b>0.9437</b>
AA	0.8243	0.8357	0.8522	0.7903	0.8950	0.8961	<b>0.9214</b>
kappa	0.8651	0.8754	0.8845	0.8291	0.9027	0.9148	<b>0.9354</b>
training time	0	0.06s	0.17s	0.23s	571.44s	1121.28s	216.86s



**Figure 4.** Two dimensional subspace found by (a) SDA, (b) SSLDA, (c) LADA and (d) DAGL on Indian Pines dataset.

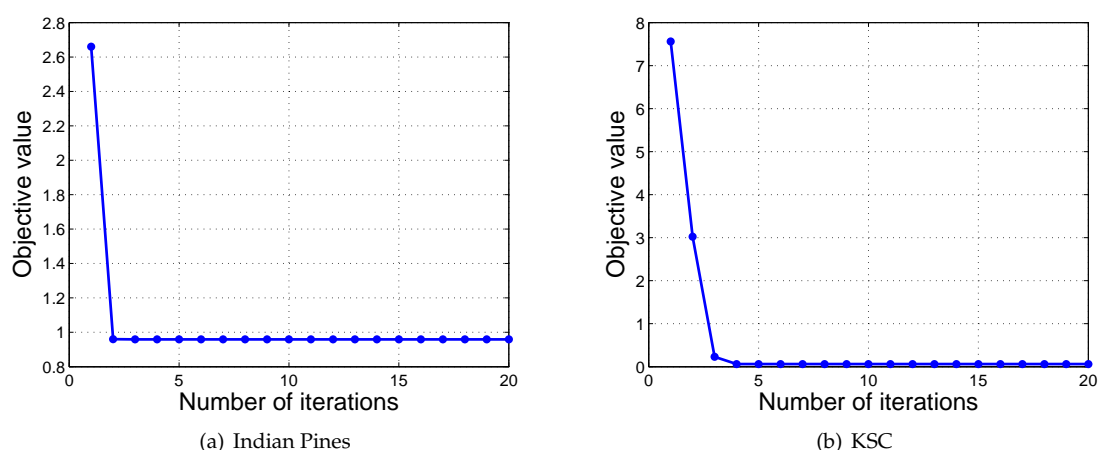
184 the model for each sample. Compared to SSLDA and LADA, DAGL is more efficient because the  
 185 optimization method converges fast.

186 Similar to the experiments on toy dataset, we also visualize the two-dimensional subspace  
 187 learned from the Indian Pines dataset. Taking the 5% samples from the Corn-notill, Grass-tree and  
 188 Soybeans-notill classes, we project the data into two-dimensional subspace with SDA, SSLDA, LADA  
 189 and the proposed DAGL. In this experiment, the spatial-smoothness terms of SSLDA, LADA and  
 190 DAGL are removed so that we do not need to train the models for each sample separately. Figure 4  
 191 shows the projected data, the subspace found by DAGL separates the samples from different classes  
 192 far away. This result explains the good performance of DAGL on the Indian Pines dataset when the  
 193 reduced dimensionality is low.

#### 194 4.3. Convergence and Parameter Sensitivity

195 The convergence behavior of the proposed optimization algorithm is studied experimentally. We  
 196 randomly choose two test samples from the Indian Pines and KSC datasets, and plot the changes  
 197 of the objective values during the optimization. From Fig. 5, we can see that the objective values of  
 198 problem (11) converge within five iterations, which verifies that the optimization algorithm is effective  
 199 and efficient.

200 In addition, the parameter sensitivity of DAGL is also discussed. The objective function (11)  
 201 contains two parameters, i.e.  $\alpha$  and  $\beta$ .  $\alpha$  affects the learning of the data graph, while  $\beta$  controls



**Figure 5.** Convergence curves of the proposed optimization algorithm on (a) Indian Pines and (b) KSC datasets.

202 the weight of the spatial smoothness term. With varying  $\alpha$  and  $\beta$ , the variance of OA is shown in  
 203 Fig. 6. We can see that DAGL is robust to  $\alpha$  and  $\beta$  in a wide range. When  $\alpha$  and  $\beta$  become very small,  
 204 the performance drops because the graph quality decreases and the spatial smoothness cannot be  
 205 guaranteed.

## 206 5. Conclusion

207 In this paper, we propose a new supervised dimensionality reduction method, termed as  
 208 Discriminant Analysis with Graph Learning (DAGL). DAGL learns the data graph automatically  
 209 during the discriminant analysis. It pulls the within-class similar samples together while pushing the  
 210 between-class similar samples far away. Compared with LDA and its graph-based variants, DAGL  
 211 is able to learn the data relationship within the desired subspace, which contains more valuable features  
 212 and less noise. In addition, DAGL ensures the smoothness within the neighborhood, so it can discover  
 213 the spatial correlation within hyperspectral images. Through the experiments on Indian Pines and  
 214 KSC datasets, DAGL provides better classification results than the state-of-the-art competitors.

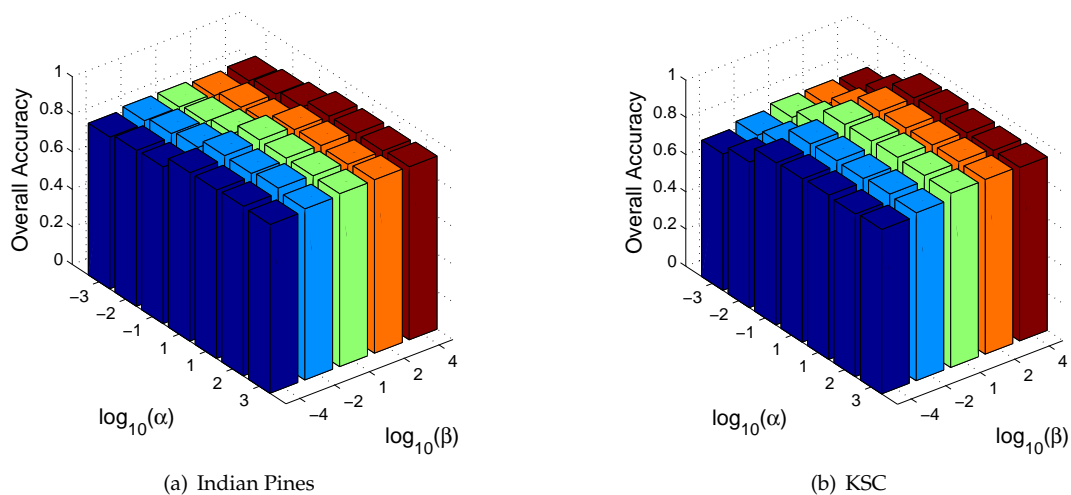
215 In future work, we would like to generalize the proposed method to the kernel version, and learn  
 216 the non-linear transformation of HSI data. It is also desirable to improve the optimization algorithm to  
 217 increase the computation efficiency.

218 **Author Contributions:** All authors conceived and designed the study. Mulin Chen carried out the experiments.  
 219 All authors discussed the basic structure of the manuscript, and Mulin Chen finished the first draft. Qi Wang,  
 220 Xuelong Li reviewed and edited the draft

221 **Conflicts of Interest:** The authors declare no conflict of interest.

## 222 References

- 223 1. Wang, Q.; Zhang, F.; Li, X. Optimal Clustering Framework for Hyperspectral Band Selection. *IEEE Trans.*  
 224 *Geosci. Remote Sens.* **2018**, DOI: 10.1109/TGRS.2018.2828161.
- 225 2. Mitra, P.; Murthy, C.; Pal, S. Unsupervised Feature Selection using Feature Similarity. *IEEE Trans. Pattern*  
 226 *Anal. Mach. Intell.* **2002**, *24*, 301-312.
- 227 3. He, X.; Cai, D.; Niyogi, P. Laplacian Score for Feature Selection. In *Advances in Neural Information*  
 228 *Processing Systems*, British Columbia, Canada, December 2005; pp. 507-514.
- 229 4. Nie, F.; Xu, D.; Li, X.; Xiang, S. Semisupervised Dimensionality Reduction and Classification through Virtual  
 230 Label Regression. *IEEE Trans. Syst. Man. Cybern.* **2011**, *41*, 675-685.



**Figure 6.** OA with varying  $\alpha$  and  $\beta$  on **(a)** Indian Pines and **(b)** KSC datasets.

- 231 5. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When Deep Learning Meets Metric Learning: Remote Sensing  
232 Image Scene Classification via Learning Discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*,  
233 2811-2821.
- 234 6. Liu, W.; Yang, X.; Tao, D.; Cheng, J.; Tang, Y. Multiview Dimension Reduction via Hessian Multiset Canonical  
235 Correlations. *Inf. Fus.* **2018**, *41*, 119-128.
- 236 7. Wang, Q.; Wan, J.; Yuan, Y. Locality Constraint Distance Metric Learning for Traffic Congestion Detection.  
237 *Pattern Recognit.* **2018**, *75*, 272-281.
- 238 8. Wang, Q.; Wan, J.; Yuan, Y. Deep Metric Learning for Crowdedness Regression. *IEEE Transactions on Circuits  
239 and Systems for Video Technology* **2017**, DOI: 10.1109/TCSVT.2017.2703920.
- 240 9. Li, X.; Chen, M.; Wang, Q. Locality Adaptive Discriminant Analysis. In International Joint Conference on  
241 Artificial Intelligence, Melbourne, Victoria, Australia, August 2017; pp. 2201-2207.
- 242 10. Nie, F.; Xiang, S.; Zhang, C. Neighborhood MinMax Projections. In International Joint Conference on  
243 Artificial Intelligence, Hyderabad, India, January 2007; pp. 993-998.
- 244 11. Fan, Zi.; Xu, Y.; Zhang, D. Local Linear Discriminant Analysis Framework using Sample Neighbors. *IEEE  
245 Trans. Neural Netw.* **2011**, *22*, 1119-1132.
- 246 12. Lu, J.; Plataniotis, K.; Venetsanopoulos, A. Regularization Studies of Linear Discriminant Analysis in Small  
247 Sample Size Scenarios with Application to Face Recognition. *Pattern Recognit. Lett.* **2005**, *26*, 181-191.
- 248 13. Wan, H.; Guo, G.; Wang, H.; Wei, X. A New Linear Discriminant Analysis Method to Address the  
249 Over-Reducing Problem. In International Conference on Pattern Recognition and Machine Intelligence,  
250 Warsaw, Poland, July 2015; pp. 65-72.
- 251 14. Prasad, S.; Bruce, L.M. Limitations of Principal Component Analysis for Hyperspectral Target Recognition.  
252 *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 625-629.
- 253 15. Fukunaga, K. *Introduction to statistical pattern recognition*, 2nd ed.; Academy Press: San Francisco, CA, USA,  
254 1972; pp. 2133-2143.
- 255 16. Hyperspectral Remote Sensing Scenes. Available online: [http://www.ehu.es/ccwintco/index.php/Hyperspectral\\_Remote\\_Sensing\\_Scenes](http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes).
- 257 17. He, X.; Niyogi, P. Locality Preserving Projections. In Advances in Neural Information Processing Systems,  
258 Vancouver, Canada, December 2004; pp. 153-160.
- 259 18. Bandos, T. V.; Bruzzone, L.; Camps-Valls, G. Classification of hyperspectral images with regularized linear  
260 discriminant analysis. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 862-873.
- 261 19. Kumar, N.; Agrawal, R. K. Two-Dimensional Exponential Discriminant Analysis for Small Sample Size in  
262 Face Recognition. *IJAISC* **2016**, *5*, 194-208.
- 263 20. Dong, Y.; Du, B.; Zhang, L.; Zhang, L. Dimensionality Reduction and Classification of Hyperspectral Images  
264 using Ensemble Discriminative Local Metric Learning. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2509-2524.

- 265 21. Cai, D.; He, X.; Han, J. Semi-Supervised Discriminant Analysis. In IEEE International Conference on  
266 Computer Vision, Rio de Janeiro, Brazil, October 2007; pp. 1-7.
- 267 22. Yuan, H.; Tang, Y.; Lu, Y.; Yang, L.; Luo, H. Spectral-Spatial Classification of Hyperspectral Image based on  
268 Discriminant Analysis. *Int. J. Artif. Intell. Soft Comput.* **2014**, *7*, 2035-2043.
- 269 23. Wang, Q.; Meng, Z.; Li, X. Locality Adaptive Discriminant Analysis for Spectral-Spatial Classification of  
270 Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens. Lett.* **2017**, *14*, 2077-2081.
- 271 24. Bressan, M.; Vitria, J. Nonparametric Discriminant Analysis and Nearest Neighbor Classification. *Pattern  
272 Recognit. Lett.* **2003**, *24*, 2743-2749.
- 273 25. Luxburg, U. A Tutorial on Spectral Clustering. *Stat. Comput.* **2007**, *17*, 395-416.
- 274 26. Shi, J.; Malik, J. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*,  
275 888-905.
- 276 27. Yao, X.; Han, J.; Zhang, D.; Nie, F. Revisiting Co-Saliency Detection: A Novel Approach Based on Two-Stage  
277 Multi-View Spectral Rotation Co-Clustering. *IEEE Trans. Image Process.* **2017**, *26*, 3496-3209.
- 278 28. Huang, J.; Nie, F.; Huang, H.; Ding, C. Robust Manifold Nonnegative Matrix Factorization. *ACM Trans.  
279 Knowl. Discov. Data* **2013**, *8*, 11:1-11:21.
- 280 29. Cai, D.; He, X.; Zhou, K.; Han, J.; Bao, H. Locality Sensitive Discriminant Analysis. In International Joint  
281 Conference on Artificial Intelligence, Hyderabad, India, January 2007; pp. 708-713.
- 282 30. Wang, Q.; Chen, M.; Li, X. Quantifying and Detecting Collective Motion by Manifold Learning. In AAAI  
283 Conference on Artificial Intelligence, San Francisco, California, USA, February 2017; pp. 4292-4298.
- 284 31. Wu, T.; Zhou, Y.; Zhang, R.; Xiao, Y.; Nie, F. Self-Weighted Discriminative Feature Selection via Adaptive  
285 Redundancy Minimization. *Neurocomputing* **2018**, *275*, 2824-2830.
- 286 32. Huang, J.; Nie, F.; Huang, H. A New Simplex Sparse Learning Model to Measure Data Similarity for  
287 Clustering. In International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, July 2015;  
288 pp. 3569-3575.
- 289 33. Liu, W.; Zha, Z.; Wang, Y.; Lu, K.; Tao, D. p-Laplacian Regularized Sparse Coding for Human Activity  
290 Recognition. *IEEE Trans. Ind. Electron.* **2016**, *63*, 5120-5129.
- 291 34. Nie, F.; Zhu, W.; Li, X. Unsupervised Feature Selection with Structured Graph Optimization. In AAAI  
292 Conference on Artificial Intelligence, Phoenix, Arizona, USA, February 2016; pp. 1302-1308.
- 293 35. Li, X.; Chen, M.; Nie, F.; Wang, Q. A Multiview-Based Parameter Free Framework for Group Detection. In  
294 AAAI Conference on Artificial Intelligence, San Francisco, California, USA, February 2017; pp. 4147-4153.
- 295 36. Nie, F.; Wang, X.; Jordan, M.; Huang, H. The Constrained Laplacian Rank Algorithm for Graph-Based  
296 Clustering. In AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, February 2016; pp.  
297 1969-1976.
- 298 37. Ly, N. H.; Du, Q.; Fowler, J. E. Collaborative Graph-Based Discriminant Analysis for Hyperspectral Imagery.  
299 *IEEE J. Sel. Top. Appl. Earth Obs. and Remote Sens.* **2014**, *7*, 2688-2696.
- 300 38. Zhang, R.; Nie, F.; Li, X. Self-weighted spectral clustering with parameter-free constraint. *Neurocomputing*  
301 **2017**, *241*, 164-170.
- 302 39. Nie, F.; Wang, H.; Huang, H.; Ding, C. Joint Schatten  $p$ -norm and  $\ell_p$ -norm Robust Matrix Completion for  
303 Missing Value Recovery. *Knowl. Inf. Syst.* **2015**, *42*, 525-544.