Locality Adaptive Discriminant Analysis for Spectral–Spatial Classification of Hyperspectral Images

Qi Wang, *Senior Member, IEEE*, Zhaotie Meng, and Xuelong Li, *Fellow, IEEE*

***Abstract*— Linear discriminant analysis (LDA) is a popular technique for supervised dimensionality reduction, but with less concern about a local data structure. This makes LDA inapplicable to many real-world situations, such as hyperspectral image (HSI) classification. In this letter, we propose a novel dimensionality reduction algorithm, *locality adaptive discriminant analysis* (LADA) for HSI classification. The proposed algorithm aims to learn a representative subspace of data, and focuses on the data points with close relationship in spectral and spatial domains. An intuitive motivation is that data points of the same class have similar spectral feature and the data points among spatial neighborhood are usually associated with the same class. Compared with traditional LDA and its variants, LADA is able to adaptively exploit the local manifold structure of data. Experiments carried out on several real hyperspectral data sets demonstrate the effectiveness of the proposed method.**

***Index Terms*— Classification, hyperspectral images (HSIs), locality adaptive discriminant analysis (LADA), spectral–spatial.**

I. Introduction

A HYPERSPECTRAL image (HSI) records reflected radiation over a series of spectral bands for each pixel in the image. Hundreds of spectral channels provide much more information than the RGB image [1]. Classification of HSIs is important but challenging due to the high-dimensional feature space and limited training samples. To address this problem, many different techniques have been proposed. Kernel-based methods, such as a support vector machine (SVM), are generally adopted and achieve the state-of-the-art classification performance [2], [3]. The support tensor machine in [4] is an extension to SVM. The effectiveness of these methods comes from learning a higher dimensional feature space, in which features of different classes are linearly separable. Subspace-based methods are another class of algorithms, which project the high-dimensional feature into a lower dimensional subspace while preserving the desired discriminative information.

Numerous dimensionality reduction algorithms are applied to learn a low-dimensional subspace. Principal component analysis [5] and linear discriminant analysis (LDA) [6] are practically and widely used methods because of its good performance in many real-world applications. In [7], hyperspectral data of different rainforest trees were classified via classical LDA. In [8], orthogonal LDA was exploited to extract spectral–spatial feature for HSI classification. Despite its good performance, there are still three drawbacks. First, standard LDA-based methods cannot solve the ill-posed problem. When the number of training samples is less than the dimension of the sample, the *within-class* scatter matrix S_w is irreversible, which results in LDA unsolvable. Second, an intrinsic problem in LDA is overreducing [9]. Suppose that the class number of training samples is C. LDA can reduce the dimension to $C-1$ at most, which restricts its capability when dealing with high dimensional data but with less class number. Third, LDA can get optimal solution when data obey Gaussian distribution, but fails in data with more complicated distribution. The reason is that LDA cannot make full use of the intrinsic data structure in the local area.

To tackle the above-mentioned problems, many techniques have been proposed for HSI classification. Bandos *et al.* [10] apply regularized LDA to mitigate the effects of the ill-posed problem. In [11], semisupervised discriminant analysis uses labeled and unlabeled data, in order to get enough training samples. Wan *et al.* [9] concentrate on individual data instead of summary data alone in learning the transformation matrix, so that its dimension reduction is independent of the class number. These methods alleviate overreducing and ill-posed problem with satisfying performance. Based on the intuitive prior of HSIs [12], data points between a small spatial neighborhood usually belong to the same class. In [13], spectral–spatial LDA incorporates the neighborhood scatter matrix as a regularizer to preserve the local structure. Although plenty of methods [14]–[17] have been developed, it is still difficult to obtain a representative subspace by adaptively exploiting the local manifold structure.

In this letter, a novel dimensionality reduction algorithm, *locality adaptive discriminant analysis* (LADA), is proposed to learn a representative subspace. Similar to locality-aware

Manuscript received August 24, 2017; accepted September 9, 2017. Date of publication September 28, 2017; date of current version October 25, 2017. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002200, in part by the National Natural Science Foundation of China under Grant 61773316, Grant 61379094, and Grant 61761130079, in part by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences under Grant QYZDY-SSW-JSC044, in part by Fundamental Research Funds for Central Universities under Grant 3102017AX010, and in part by the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences. *(Corresponding author: Qi Wang.)*

Q. Wang is with the School of Computer Science, also with the Unmanned System Research Institute, and also with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@gmail.com).

Z. Meng is with the the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: optimalmzt@gmail.com).

X. Li is with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China, and also with The University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: xuelong_li@opt.ac.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2017.2751559

methods, LADA mainly focuses on data points with close relationship. The major difference is that LADA exploits the points' relationship adaptively and adopts a local scatter matrix yielded from a small neighborhood as a regularizer. Furthermore, by deliberately designing the objective function, the *between-class* scatter matrix is nonsingular, which eliminates the overreducing problem naturally. The contributions of LADA are summarized as follows.

1) LADA can deal with the case where data distribution is more complex than Gaussian, and avoid the overreducing problem.
2) LADA has the ability to adaptively capture the relationship between similar points and preserve local manifold structure of spatial neighborhood in the desired subspace.

II. Linear Discriminant Analysis

Given the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n], \mathbf{x}_i \in \mathbb{R}^{d\times 1}$ with C classes, where each column represents a training sample and d is the number of spectral channels. LDA aims to find a transformation matrix $\mathbf{G} \in \mathbb{R}^{d\times m}(m \ll d)$ to convert the d-dimensional data $\mathbf{x}_i$ into an m-dimensional vector

$$\mathbf{y}_i = \mathbf{G}^T\mathbf{x}_i. \tag{1}$$

The intuitive motivation of LDA is to push the data points of different classes far away from each other, and at the same time to pull those within the same class as close as possible. So the objective function of LDA can be written as

$$\max_{\mathbf{G}} \frac{\sum_{i=1}^{C} n_i \|\mathbf{G}^T(\mu^i - \mu)\|_2^2}{\sum_{i=1}^{C}\sum_{j=1}^{n_i} \|\mathbf{G}^T\left(\mathbf{x}_j^i - \mu^i\right)\|_2^2} \tag{2}$$

where n_i is the number of training samples in class i, μ^i is the mean of samples in class i, μ is the mean of all samples, and $\mathbf{x}_j^i$ is the jth sample in class i. Denote the *between-class* scatter matrix $\mathbf{S}_b$ and the *within-class* scatter matrix $\mathbf{S}_w$ as

$$\mathbf{S}_b = \sum_{i=1}^{C} n_i(\mu^i - \mu)(\mu^i - \mu)^T \tag{3}$$

$$\mathbf{S}_w = \sum_{i=1}^{C}\sum_{j=1}^{n_i} n_i\left(\mathbf{x}_j^i - \mu^i\right)\left(\mathbf{x}_j^i - \mu^i\right)^T. \tag{4}$$

Then, the formula (2) can be rewritten into a concise form, which is the Rayleigh coefficient

$$\max_{\mathbf{G}} \frac{\mathrm{tr}(\mathbf{G}^T\mathbf{S}_b\mathbf{G})}{\mathrm{tr}(\mathbf{G}^T\mathbf{S}_w\mathbf{G})} \tag{5}$$

where tr() denotes the trace operator.

From the objective function, it could be clearly seen that LDA mainly emphasizes the global relationship of data, which makes it less sensitive to the local structure. Therefore, locality-aware variants are developed to address this drawback.

III. Locality Adaptive Discriminant Analysis

In this section, the LADA is presented for dimensionality reduction. First, the objective function of LADA is described and analyzed qualitatively. Then, an iterative learning strategy is designed to obtain the optimal solution.

A. Problem Formulation

In real-world applications, such as HSI classification, the data distribution may not be Gaussian. Therefore, it is crucial to capture the local structure of data manifold. Our goal is to learn an optimal transformation matrix $\mathbf{G}$ to pull the similar points together while pushing the dissimilar ones far away from each other.

Given the data points $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n], \mathbf{x}_i \in \mathbb{R}^{d\times 1}$, the objective function is designed as

$$\min_{\mathbf{G},\mathbf{s}} \frac{\sum_{i=1}^{C} n_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2 \|\mathbf{G}^T\left(\mathbf{x}_j^i - \mathbf{x}_k^i\right)\|_2^2}{n^{-1}\sum_{j=1}^{n}\sum_{k=1}^{n}\|\mathbf{G}^T\left(\mathbf{x}_j - \mathbf{x}_k\right)\|_2^2} \tag{6}$$

$$\text{s.t.} \sum_{k=1}^{n_i} s_{jk}^i = 1, \quad s_{jk}^i \geq 0 \tag{7}$$

where n is the number of samples, $\mathbf{s}$ is a weighted matrix, s_{jk}^i denotes the weight between the jth and kth samples in class i, and the remaining definitions are the same as those in LDA. Note that $\mathbf{x}_j$ is the jth sample in the whole data set, and it is different from $\mathbf{x}_j^i$.

In formula (7), the weight matrix $\mathbf{s}$ is introduced to capture the local relationship between data points. The constraints on $\mathbf{s}$ avoid the case that some rows of $\mathbf{s}$ are all zeros. Supposing that the transformation matrix $\mathbf{G}$ is obtained, s_{jk}^i will be large if the transformed distance $\|\mathbf{G}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)\|_2^2$ is small, which means $\mathbf{x}_j^i$ and $\mathbf{x}_k^i$ are similar in the learned subspace. If we fix $\mathbf{s}$, which means that the similarity of data points is obtained, and optimize $\mathbf{G}$, then the objective function will consider more similar points in the previously learned subspace. Therefore, the relationship of data points can be acquired in the desired subspace by optimizing $\mathbf{s}$ and $\mathbf{G}$ iteratively.

Similar to basic LDA, we denote the *between-class* scatter matrix $\widetilde{\mathbf{S}}_b$ and the *within-class* scatter matrix $\widetilde{\mathbf{S}}_w$ as

$$\widetilde{\mathbf{S}}_b = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(\mathbf{x}_j - \mathbf{x}_k)(\mathbf{x}_j - \mathbf{x}_k)^T \tag{8}$$

$$\widetilde{\mathbf{S}}_w = \sum_{i=1}^{C} n_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2\left(\mathbf{x}_j^i - \mathbf{x}_k^i\right)\left(\mathbf{x}_j^i - \mathbf{x}_k^i\right)^T. \tag{9}$$

Then, formula (7) becomes

$$\max_{\mathbf{G}} \frac{\mathrm{tr}(\mathbf{G}^T\widetilde{\mathbf{S}}_b\mathbf{G})}{\mathrm{tr}(\mathbf{G}^T\widetilde{\mathbf{S}}_w\mathbf{G})}. \tag{10}$$

In HSI classification, integrating spectral and spatial information is popular and can effectively improve classification performance. It is intuitive that data points within a small spatial neighborhood usually belong to the same class. Therefore, we exploit the spatial information by making full use of neighborhood for each test sample. Given $\mathbf{z}_1 \in \mathbb{R}^{d\times 1}$ is a test sample, we denote the K neighbor points of $\mathbf{z}_1$ as $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K]$. After transformation, these points still

keep similar in the low-dimensional space

$$\begin{aligned}\min_G \sum_{j=1}^{K} \|\mathbf{G}^T\mathbf{z}_i - \mathbf{G}^T\overline{\mathbf{z}}\|_2^2 &= \sum_{j=1}^{K} \|\mathbf{G}^T(\mathbf{z}_i - \overline{\mathbf{z}})\|_2^2 \\ &= \|\mathbf{G}^T(\mathbf{Z} - \overline{\mathbf{z}}\mathbf{1}^T)\|_2^2 \\ &= \mathrm{tr}(\mathbf{G}^T(\mathbf{Z} - \overline{\mathbf{z}}\mathbf{1}^T)(\mathbf{Z} - \overline{\mathbf{z}}\mathbf{1}^T)^T\mathbf{G}) \\ &= \mathrm{tr}(\mathbf{G}^T\mathbf{S}_z\mathbf{G}) \end{aligned} \tag{11}$$

where $\mathbf{1} = [1, \ldots, 1]^K$, $\overline{\mathbf{z}} = \frac{1}{K}\sum_{j=1}^{K}\mathbf{z}_j$ and $\mathbf{S}_z = (\mathbf{Z} - \overline{\mathbf{z}}\mathbf{1}^T)(\mathbf{Z} - \overline{\mathbf{z}}\mathbf{1}^T)^T$. We add the local scatter matrix $\mathbf{S}_z$ to the objective function of (10) as a regularizer with a tradeoff parameter λ. Then, the optimization objective becomes

$$\min_{\mathbf{G},\mathbf{s}} \frac{\mathrm{tr}(\mathbf{G}^T(\widetilde{\mathbf{S}}_w + \lambda\mathbf{S}_z)\mathbf{G})}{\mathrm{tr}(\mathbf{G}^T\widetilde{\mathbf{S}}_b\mathbf{G})} \tag{12}$$

$$\text{s.t.} \sum_{k=1}^{n_i} s_{jk}^i = 1, \quad s_{jk}^i \geq 0. \tag{13}$$

B. Optimization Strategy

In this section, an adaptive optimization strategy is presented to solve formula (13). First, the weight of the points in class i is initialized as $1/n_i$, and the weight of points from different classes is set to 0. Then, we optimize $\mathbf{G}$ and $\mathbf{s}$ iteratively.

When $\mathbf{s}$ is fixed, formula (13) can be solved via eigendecomposition. To obtain a stable solution, the matrix $(\widetilde{\mathbf{S}}_w + \lambda\mathbf{S}_z)$ should be nonsingular. However, when $n < d$, it is singular. We address this by referring Tikhonov regularization [18], for matrix $(\widetilde{\mathbf{S}}_w + \lambda\mathbf{S}_z + \gamma\mathbf{I})$ is certainly nonsingular if $\gamma > 0$.

When $\mathbf{G}$ is fixed, the objective function (13) can be reduced to

$$\min_{\mathbf{s}} \sum_{i=1}^{C}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2 \|\mathbf{G}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)\|_2^2 \tag{14}$$

$$\text{s.t.} \sum_{k=1}^{n_i} s_{jk}^i = 1, s_{jk}^i \geq 0 \tag{15}$$

which is equivalent to the following problem:

$$\min_{\mathbf{s}_j^i} \sum_{k=1}^{n_i} {s_{jk}^i}^2 \|\mathbf{G}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)\|_2^2 \tag{16}$$

$$\text{s.t.} \sum_{k=1}^{n_i} s_{jk}^i = 1, s_{jk}^i \geq 0 \tag{17}$$

where $\mathbf{s}_j^i$ is a column vector with its kth element equal to s_{jk}^i. Denoting that υ_k and a vector $\boldsymbol{\alpha}$ are equal to $\|\mathbf{G}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)\|_2^2, \mathbf{s}_j^i$, respectively, formula (17) is simplified to

$$\min_{\boldsymbol{\alpha}^T\mathbf{1}=1, \boldsymbol{\alpha}\geq 0} \sum_{k=1}^{n_i} \alpha_k^2\upsilon_k. \tag{18}$$

Defining a diagonal matrix $\mathbf{V}$ with V_{kk} equal to υ_k, formula (18) becomes

$$\min_{\boldsymbol{\alpha}^T\mathbf{1}=1, \boldsymbol{\alpha}\geq 0} \sum_{k=1}^{n_i} \boldsymbol{\alpha}^T\mathbf{V}\boldsymbol{\alpha}. \tag{19}$$

Without the second constraint $\boldsymbol{\alpha} \geq 0$, the Lagrangian function of formula (19) is

$$\mathcal{L}(\alpha, \eta) = \boldsymbol{\alpha}^T\mathbf{V}\boldsymbol{\alpha} - \eta(\boldsymbol{\alpha}^T\mathbf{1} - 1) \tag{20}$$

where η is the Lagrangian multiplier. Taking the derivative of (20) with respect to $\boldsymbol{\alpha}$ and setting it to zero, we get

$$2\mathbf{V}\boldsymbol{\alpha} - \eta\mathbf{1} = 0. \tag{21}$$

Together with the constraint $\boldsymbol{\alpha}^T\mathbf{1} = 1$, $\boldsymbol{\alpha}$ can be computed as

$$\alpha_k = \frac{1}{\upsilon_k} \times \left(\sum_{t=1}^{n_i} \frac{1}{\upsilon_t}\right)^{-1}. \tag{22}$$

Fortunately, the above α_k satisfies the constraint $\boldsymbol{\alpha} \geq 0$, so it is also the optimal solution to formula (19). Accordingly, the optimal solution to the formula (17) is

$$s_{jk}^i = \frac{1}{\|\mathbf{G}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)\|_2^2} \times \left(\sum_{t=1}^{n_i} \frac{1}{\|\mathbf{G}^T(\mathbf{x}_j^i - \mathbf{x}_t^i)\|_2^2}\right)^{-1}. \tag{23}$$

By optimizing $\mathbf{G}$ and $\mathbf{s}$ iteratively, our method is capable of exploiting the data points' local relationship in the desired low-dimensional subspace. The implementation details of LADA are shown in Algorithm 1.

Algorithm 1 LADA

Require: Data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ and $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K]$, where $\mathbf{z}_1$ is the target test sample, the desired dimension m, hyper-parameters λ, γ.

1: Initialize weight matrix $\mathbf{s}$, scatter matrix $\widetilde{S}_w, \widetilde{S}_b, \mathbf{S}_z$.
2: **repeat**
3: Compute transformation matrix $\mathbf{G}$ by finding the m largest eigenvectors of matrix $(\widetilde{\mathbf{S}}_w + \lambda\mathbf{S}_z + \gamma\mathbf{I})^{-1}\widetilde{S}_b$.
4: Update $\mathbf{s}$ with Eq. 23.
5: Compute current $\mathrm{tr}(\mathbf{G}^T(\widetilde{\mathbf{S}}_w + \lambda\mathbf{S}_z)\mathbf{G})$
6: Compute the difference between current trace and last trace.
7: **until** converge, which means the value in step 6 is less than ϵ, which is set manually.
8: $\mathbf{Y} = \mathbf{G}^T\mathbf{X}$

Ensure: $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n], \mathbf{y}_i \in \mathbb{R}^{m\times 1}$, $\mathbf{G}$

IV. Experiments

A. Data Set Description

1) Indian Pines Image [19]: This image is gathered by the AVIRIS sensor over the Indian Pines test site in northwestern Indiana. The original image consists of 145×145 pixels with 224 spectral bands. After removing the 24 bands (104–108, 150–163, and 220) affected by water absorption, the remaining 200 bands are used for classification. Each pixel has a specific class label, and there are in total 16 classes.

2) Pavia University Image [19]: This image is gathered by the ROSIS sensor over Pavia, northern Italy. The original image consists of 610×610 pixels with 103 spectral bands. Some pixels containing no information have to be discarded, and the remaining 610×340 pixels with nine classes are used for classification.

TABLE I
CLASSIFICATION PERFORMANCE OF DIFFERENT ALGORITHMS FOR THE INDIAN PINES IMAGE (WITH BEST FEATURE NUMBER IN BRACKETS)

Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	OA	AA	kappa
RAW (200)	12.5	66.26	62.33	16.42	84.56	92.02	**98.64**	97.14	33.33	71.64	71.53	51.43	98.36	90.79	28.42	**99.58**	72.62±1.18	69.13±3.04	68.67±1.22
RLDA (15)	25.00	57.16	44.04	20.30	84.41	92.49	10.00	99.43	0.00	46.11	67.20	56.11	**99.02**	93.77	54.74	85.93	67.75±1.54	58.48±1.34	62.91±1.69
SDA (9)	63.75	67.43	61.97	35.82	85.88	94.46	77.50	95.71	**63.33**	72.07	77.23	56.00	98.03	93.71	34.32	88.15	75.65±1.52	72.84±1.50	72.15±1.72
LPP (21)	60.00	64.79	60.90	37.01	84.71	93.62	87.50	96.29	56.67	72.51	76.73	55.31	98.03	93.22	37.05	88.15	75.03±1.06	72.66±3.12	71.43±1.20
SSLDA (10)	**81.25**	86.55	85.92	**78.81**	92.94	97.09	82.50	**100**	53.33	82.40	93.32	85.60	91.80	96.48	88.21	82.96	90.48±1.24	86.20±2.63	89.10±1.43
LADA (24)	75.00	**88.70**	**86.73**	74.63	**93.97**	**98.12**	60.00	**100**	53.33	**86.47**	**94.77**	**87.31**	93.11	**96.96**	**90.11**	82.22	**91.75**±1.50	85.09±3.42	90.56±1.72

B. Experimental Setup

The proposed LADA is compared with four dimensionality reduction methods, i.e., regularized LDA (RLDA) [10], locality preserving projections (LPP) [14], SDA [11], and spectral-spatial LDA (SSLDA) [13]. The 1-nearest neighbor (1NN) is applied to classify the test samples. Each algorithm is conducted five times, and the average result and variance are calculated to avoid the random error. Overall accuracy (OA), average accuracy (AA), and kappa statics (κ) are used to evaluate the classification performance. The classification maps corresponding to each compared algorithm are displayed. Especially, RAW is also compared with other methods. It takes initial data as features, which are classified by the following 1NN classifier.

The data points used for training and test are randomly sampled from each class. In our experiments, we sample 5% points as training set, and the remaining 30% data as test set. For SDA and LPP, additional 20% samples are applied for training. In RLDA, the hyperparameter γ is selected in $\{10^{-3}, \ldots, 10^{0}\}$. The parameter α in SDA is selected in $\{0.1, 0.5, 2.5, 12.5\}$. The parameters λ and γ in both SSLDA and LADA are selected in $\{10^{-3}, \ldots, 10^{3}\}$. All these parameters are determined by fivefold cross validation, and finally, α, λ, and γ are 0.1, 100, and 10^{-3}, respectively. The neighborhood point number K is 9 in our implementation. The smallest number of K is 9, for the 3×3 neighborhood is the smallest region to keep spatial information. With increasing K, more neighbor points will be included, and it is beneficial to data set with large neighborhood within the same class. K should be set corresponding to the prior of the data set.

C. Experimental Results and Analysis

We have conducted the experiment five times for Indian Pines and Pavia University, respectively. The OAs and standard deviations with different feature numbers in each compared algorithm are shown in Fig. 1. RLDA, SDA, and SSLDA are all LDA-based methods, so the maximal feature number is $C-1$, where C is the class number of training samples. LPP is an unsupervised method, with no constraints on class number. It is clear that in Fig. 1(a), LADA outperforms the compared algorithms in each dimension. Compared with SSLDA, an LDA-based method with spatial neighborhood constraint, LADA can learn more representative subspace by adaptively exploiting data points with a close relationship in the spectral domain. Moreover, compared with other LDA-based methods without spatial neighborhood constraint, LADA still has a great superiority. As shown in Fig. 1(b), LADA performs worse with feature numbers less than 4. However, benefiting from the formulation (8) where $\widetilde{\mathbf{S}}_b$ is full rank, LADA can reduce the feature number within d, where d is the dimension of sample, so it can keep more features and avoid the problem of small feature numbers.

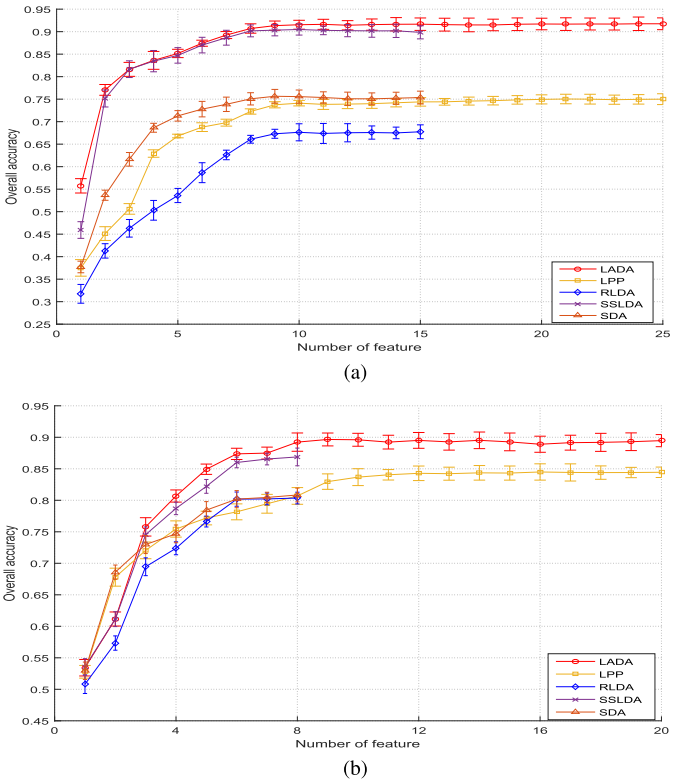

Fig. 1. OA and standard deviation of compared algorithms with features increasing. (a) Indian Pines image. (b) Pavia University image.

The classification performance of different dimensionality reduction algorithms is shown in Tables I and II for Indian Pines and Pavia University, respectively. OA, AA, κ, and their standard deviations (%) are computed to evaluate each algorithm. The optimal feature numbers obtained in Fig. 1 are written in brackets. There are 16 classes in the Indian Pines data set, and the OA of each class is calculated. LADA achieves the best OA (91.75%) with 24 features in all compared methods. Among all the 16 classes, it also has the highest accuracy of ten classes. In the Pavia University image, LADA also achieves the best performance (89.66% OA) with nine features. It surpasses other compared algorithms in eight classes. The results of the Pavia data set are worse than that of the India Pines data set. As there are many small buildings in the University of Pavia, most homogenous regions are small than Indian Pines. Besides, the region in Indian Pines is more regular. The large and regular homogenous regions ensure the better performance in the Indian Pines data set. SSLDA, slightly worse than LADA, is the second place method. The

TABLE II

CLASSIFICATION PERFORMANCE OF DIFFERENT ALGORITHMS FOR THE PAVIA UNIVERSITY IMAGE (WITH BEST FEATURE NUMBER IN BRACKETS)

Class	1	2	3	4	5	6	7	8	9	OA	AA	kappa
RAW (103)	74.48	94.50	68.61	82.24	99.22	49.69	80.24	74.71	96.67	81.93±0.12	80.04±0.83	75.76±0.16
RLDA (8)	75.92	91.56	60.76	83.79	99.48	52.41	73.12	72.0	90.0	80.36±0.04	77.73±1.34	73.78±1.69
SDA (8)	78.56	90.90	63.80	75.34	99.42	52.82	81.42	76.43	96.67	80.84±1.22	79.04±1.73	74.43±1.53
LPP (16)	79.68	94.90	69.11	83.45	99.61	58.79	82.61	78.0	**97.22**	84.50±0.82	82.60±1.74	79.26±1.12
SSLDA (8)	79.44	95.82	80.0	84.66	**100**	63.91	92.89	84.86	91.11	86.88±1.18	85.85±2.54	82.47±1.31
LADA (9)	**82.0**	**98.16**	**81.27**	**91.55**	**100**	**70.08**	**93.68**	**85.14**	92.22	**89.66**±1.35	88.23±2.42	86.19±1.62

(a) (b) (c) (d) (e) (f) (g)

Fig. 2. Classification maps for the Indian Pines image by 1NN. (a) Ground truth. (b) LADA (OA = 0.9175). (c) LPP (OA = 0.7503). (d) RAW (OA = 0.7262). (e) RLDA (OA = 0.6775). (f) SDA (OA = 0.7565). (g) SSLDA (OA = 0.9048).

different best feature numbers of LADA and SSLDA indicate that without overreducing, our method can keep more useful information. All the results manifest that our method can better capture the local manifold structure of the original data and project them to a representative subspace.

To further compare the results, we display the classification maps of each method for Indian Pines in Fig. 2. The best feature numbers are selected and the highest OA is attached in brackets. (Considering the constraint of page numbers, we do not show the classification maps for Pavia University.) The best visualization effect of LADA indicates that our method can learn a more representative subspace by simultaneously exploiting spectral and spatial information.

Although the performance of LADA is satisfying, the computing demand is higher than other LDA-based methods. The main reason is that the convergence of LADA is time-consuming and the computing complexity is proportional to the number of test samples.

V. CONCLUSION

In this letter, we propose a novel dimensionality reduction method LADA and apply it to HSI classification. LADA focuses on data points with a close relationship both in spectral and spatial domains. First, we design a locality adaptive objective function, which aims to adaptively learn a representative subspace of high-dimensional spectral feature. Second, we impose a local scatter matrix, generating from a small neighborhood, as a regularizer of the above objective function, so that the local spatial structure can be preserved after transformation. Finally, we design an optimization strategy to iteratively learn the transformation matrix $\mathbf{G}$. Experimental results on the real-world HSI data set demonstrate that the proposed method is applicable in data with complex distribution than Gaussian. The best classification performance shows that it is more powerful to the preserve local manifold structure in the desired subspace.

REFERENCES

[1] A. Zare, J. Bolton, J. Chanussot, and P. Gader, "Foreword to the special issue on hyperspectral image and signal processing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 1841–1843, 2014.

[2] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[3] W. Gao and Y. Peng, "Ideal kernel-based multiple kernel learning for spectral-spatial classification of hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 1051–1055, Jul. 2017.

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

[5] I. T. Jolliffe, *Principal Component Analysis*. Hoboken, NJ, USA: Wiley, 2002.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Francisco, CA, USA: Academic, 2013.

[7] M. L. Clark, D. A. Roberts, and D. B. Clark, "Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales," *Remote Sens. Environ.*, vol. 96, no. 3, pp. 375–398, 2005.

[8] H. R. Shahdoosti and F. Mirzapour, "Spectral–spatial feature extraction using orthogonal linear discriminant analysis for classification of hyperspectral data," *Eur. J. Remote Sens.*, vol. 50, no. 1, pp. 111–124, 2017.

[9] H. Wan, G. Guo, H. Wang, and X. Wei, "A new linear discriminant analysis method to address the over-reducing problem," in *Proc. Int. Conf. Pattern Recognit. Mach. Intell.*, 2015, pp. 65–72.

[10] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.

[11] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–7.

[12] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.

[13] H. Yuan, Y. Y. Tang, Y. Lu, L. Yang, and H. Luo, "Spectral-spatial classification of hyperspectral image based on discriminant analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2035–2043, Jun. 2014.

[14] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160.

[15] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1185–1198, Apr. 2012.

[16] X. Huang, Q. Lu, and L. Zhang, "A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas," *ISPRS J. Photogramm. Remote Sens.*, vol. 90, no. 4, pp. 36–48, Apr. 2014.

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

[18] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[19] *Hyperspectral Remote Sensing Scenes*. Accessed: Mar. 5, 2017. [Online]. Available: http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes