

Sparse matrix transform-based linear discriminant analysis for hyperspectral image classification

Jiangtao Peng¹ · Tao Luo²

Received: 2 April 2014 / Revised: 18 July 2015 / Accepted: 26 July 2015 / Published online: 2 August 2015
© Springer-Verlag London 2015

Abstract Due to the high dimensionality of hyperspectral image (HSI), dimension reduction or feature extraction is usually needed before the HSI classification. Traditional linear discriminant analysis (LDA) method for feature extraction usually encounters difficulty because the available training samples in HSI classification are limited, which causes the singularity of data scatter matrix. In this paper, we propose a sparse matrix transform-based LDA (SMT-LDA) algorithm for the HSI classification. By using SMT, the total scatter matrix used in LDA can be constrained to have an eigen-decomposition where the eigenvectors can be sparsely parametrized by a limited number of Givens rotations. In this way, the estimated scatter matrix is always positive definite and well conditioned even in the case of limited training samples. The proposed SMT-LDA method is compared with regularized LDA and PCA-LDA methods on two benchmark hyperspectral data sets. Experimental results indicate that the performance of the proposed method is overall superior to these methods, especially for small-sample-size classification.

Keywords Hyperspectral image · Linear discriminant analysis · Sparse matrix transform · Dimension reduction · Small-sample-size

✉ Jiangtao Peng
pengjt1982@126.com

Tao Luo
luo_tao@tju.edu.cn

¹ Faculty of Mathematics and Statistics, Hubei Key Laboratory of Applied Mathematics, Hubei University, Wuhan 430062, China

² School of Computer Science and Technology, Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin 300072, China

1 Introduction

Hyperspectral remote sensors capture digital images in hundreds of narrow spectral bands spanning the visible-to-infrared spectrum [1]. It can be used to capture high-resolution hyperspectral images (HSIs) for environmental mapping, geological research, plant and mineral identification, crop analysis, and so on. In all of these applications, it usually requires to classify the pixels in the scene, where a pixel (or sample) is represented as a vector whose entries correspond to the reflection or absorption value in different spectral bands. In HSI classification, we usually have few training samples (small samples) coupled with a large number of spectral channels (high dimensionality) [2]. Large number of bands provide rich information for classifying different materials in the scene. However, with few training samples, beyond a certain limit, the classification accuracy decreases as the number of features increases (Hughes phenomenon [3]). In order to obtain good classification performance, it needs more training samples which are rarely feasible in hyperspectral remote sensing applications. Therefore, for high-dimensional small-sample hyperspectral data, the classification is relatively difficult. Moreover, the large amount of features involved in HSI will dramatically increase processing complexity. An HSI data generally consist of thousands of pixels over hundreds of spectral bands. Classification of this tremendous amount of data is time-consuming and requires significant computational effort, which may not be possible in many applications. Therefore, for classification of HSI data, it is common to perform a dimension reduction or feature extraction procedure followed by classification algorithms [4–6].

A basic and commonly used method for feature extraction is the Fisher linear discriminant analysis (LDA) [7,8]. The objective of LDA is to find the most discriminant projection

that maps high-dimensional samples into a low-dimensional space, which maximizes the ratio of between-class scatter against within-class scatter. However, for high-dimensional data such as faces and hyperspectral images, the traditional LDA algorithm encounters several difficulties [9]. First, it is computationally challenging for the eigen-decomposition of dense matrices in high-dimensional space. Second, the scatter matrices are always singular due to the small-sample-size problem. In particular, for hyperspectral data, the scatter matrix is extremely unstable because of the small ratio between the number of available training samples and the number of spectral bands. Many methods have been proposed to solve the ill-posed problem. Shrinkage and regularized covariance estimators are examples of such techniques [10–12]. Regularized LDA (RLDA) regularizes the scatter matrix by shrinking it toward some positive definite target structures, such as the identity matrix or the diagonal of the scatter matrix [10, 13]. Another kind of methods to solve the small-sample-size problem are subspace methods. The Fisherface (also called PCA-LDA) method uses principal component analysis (PCA) as a pre-processing step for dimensionality reduction so as to discard the null space of the within-class scatter matrix and then performs LDA in the lower-dimensional PCA subspace [14]. A potential problem is that PCA step may discard important discriminative information [15].

In this paper, we propose a sparse matrix transform-based LDA (SMT-LDA) algorithm for dimension reduction and classification of high-dimensional hyperspectral data. In the proposed SMT-LDA, the total scatter matrix is constrained to have an eigen-decomposition which can be represented as a sparse matrix transform (SMT) [16]. The SMT is formed by a product of pairwise coordinate Givens rotations. Under this framework, the total scatter matrix can be efficiently estimated using greedy minimization of the negative log likelihood function [16]. The estimated scatter matrix is always positive definite and well conditioned even with limited samples.

2 The algorithm

2.1 Linear discriminant analysis

Consider a set of N samples $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ taking values in an n -dimensional space, and assume that each sample belongs to one of c classes $\{X_1, X_2, \dots, X_c\}$. The between-class scatter matrix is defined as

$$S_b = \sum_{i=1}^c N_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \quad (1)$$

and the within-class scatter matrix is defined as

$$S_w = \sum_{i=1}^c \sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \boldsymbol{\mu}_i)(\mathbf{x}_k - \boldsymbol{\mu}_i)^T \quad (2)$$

where $\boldsymbol{\mu}$ is the total sample mean vector, $\boldsymbol{\mu}_i$ is the mean sample of class X_i , and N_i is the number of samples in class X_i .

LDA seeks directions on which data points of different classes are far from each other while requiring data points of the same class to be close to each other. That is, LDA projection maximizes the ratio of between-class scatter against within-class scatter as follows:

$$W^* = \arg \max_W J(W) = \frac{|W^T S_b W|}{|W^T S_w W|} \quad (3)$$

If S_w is nonsingular, the optimal projection is computed by applying an eigen-decomposition on the scatter matrices of the given training data.

2.2 Sparse matrix transform

SMT is originally designed to estimate the covariance matrix [16–19]. Given a set of training samples $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$, and assume \mathbf{x}_k has zero mean. The sample covariance is computed by $S = \frac{1}{N} X X^T$, and S is an unbiased estimate of the true covariance matrix R . The eigen-decomposition of R is: $R = E \Lambda E^T$, where E is the orthogonal eigenvector matrix and Λ is the diagonal matrix of eigenvalues.

Assume the columns of X are independent and identically distributed Gaussian random vectors with mean zero and covariance R . Jointly maximizing the likelihood of X with respect to E and Λ results in [16]

$$\hat{E} = \arg \min_{E \in \Omega} \left\{ |\text{diag}(E^T S E)| \right\} \quad (4)$$

$$\hat{\Lambda} = \text{diag}(\hat{E}^T S \hat{E}) \quad (5)$$

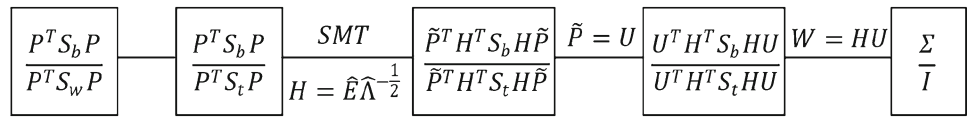
where Ω is the set of allowed orthogonal transforms. Then, $\hat{R} = \hat{E} \hat{\Lambda} \hat{E}^T$ is the maximum-likelihood estimate of the covariance.

Based on the idea that the maximum-likelihood estimate of E can be improved by constraining the feasible set of eigenvectors Ω to a smaller set, Cao et al. [17, 18] proposed to restrict Ω to be the set of all orthonormal transforms that can be represented as the product of K Givens rotations.

In particular, E is approximated by a series of K Givens rotations: $E = E_1 E_2 \dots E_K$, each of which is a simple rotation of angle θ_k about two axes i_k and j_k . Each rotation is given by a matrix of the form $E_k = I + \Theta(i_k, j_k, \theta_k)$ where

$$\Theta(i_k, j_k, \theta_k)_{rs} = \begin{cases} \cos(\theta_k) - 1, & \text{if } r = s = i_k \text{ or } r = s = j_k \\ \sin(\theta_k), & \text{if } r = i_k \text{ and } s = j_k \\ -\sin(\theta_k), & \text{if } r = j_k \text{ and } s = i_k \\ 0, & \text{otherwise.} \end{cases}$$

Fig. 1 Flowchart of SMT-LDA algorithm



The aim is to produce an estimate of the eigenvector matrix that is sparsely parametrized by a limited number of rotations.

A greedy minimization method is used to solve the problem. At each iteration, two coordinates i_k and j_k are first determined by minimizing the cost of (4), which results in

$$(i_k, j_k) \leftarrow \arg \min_{(i,j)} \left(1 - \frac{S_{ij}^2}{S_{ii}S_{jj}} \right)$$

Once i_k and j_k are determined, the Givens rotation E_k^* is given by

$$E_k^* = I + \Theta(i_k, j_k, \theta_k)$$

where $\theta_k = \frac{1}{2} \text{atan}(-2S_{i_k j_k}, S_{i_k i_k} - S_{j_k j_k})$.

2.3 Sparse matrix transform for LDA

Denote: $\bar{X} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N]$, where $\bar{x}_i = \mathbf{x}_i - \boldsymbol{\mu}$. The sample covariance (total scatter matrix) can be computed by $S_t = \bar{X}\bar{X}^T$ if we ignore the constant factor $1/N$, and $S_t = S_w + S_b$.

Based on the SMT technique, we can obtain the eigen-decomposition of S_t as: $S_t = \hat{E}\hat{\Lambda}\hat{E}^T$, where \hat{E} is an SMT of order K and $\hat{\Lambda}$ is a diagonal matrix which are obtained as follows:

$$\hat{E} = \arg \min_{E \in \Omega} \left\{ \left| \text{diag}(E^T S_t E) \right| \right\}$$

$$\hat{\Lambda} = \text{diag}(\hat{E}^T S_t \hat{E})$$

SMT representation brings the eigenvalues $\hat{\Lambda}$ and eigenvectors \hat{E} by using a small number of Givens rotations which avoids eigen-decomposition of dense scatter matrix S_t .

Recall that the conventional Fisher LDA criterion function in (3) can be modified as follows [20]:

$$J(W) = \frac{|W^T S_b W|}{|W^T S_t W|} \tag{6}$$

where the within-class scatter matrix S_w in (3) is replaced by the total scatter matrix S_t . Defining a mapping $W = HU$, the criterion (6) is changed to

$$J(W) = \frac{|U^T H^T S_b H U|}{|U^T H^T S_t H U|} = \frac{|U^T \tilde{S}_b U|}{|U^T \tilde{S}_t U|} \tag{7}$$

if the transform matrix H can whiten S_t , that is $\tilde{S}_t = H^T S_t H = I$, then all we need to do is to find the eigenvectors of $\tilde{S}_b = H^T S_b H$, which is just the matrix U .

Figure 1 provides a flowchart of the proposed SMT-LDA algorithm. For the purpose of discriminant analysis, we aim to find a matrix that simultaneously diagonalizes both S_t and S_b . This can be achieved by diagonalizing S_t using SMT first and then diagonalizing S_b . The detailed procedures are outlined below.

(1) Diagonalize S_t

Based on SMT, S_t can be formulated as $S_t = \hat{E}\hat{\Lambda}\hat{E}^T$. Denote $H = \hat{E}\hat{\Lambda}^{-\frac{1}{2}}$, then $H^T S_t H = I$.

(2) Diagonalize S_b

Now we compute orthogonal matrix U and diagonal matrix Σ such that $H^T S_b H = U \Sigma U^T$. Defining $W = HU$, then W diagonalizes S_b .

(3) Projection matrix: $W = HU = \hat{E}\hat{\Lambda}^{-\frac{1}{2}}U$

W diagonalizes S_t and S_b at the same time, that is, $W^T S_t W = I$ and $W^T S_b W = \Sigma$. Moreover,

$$\begin{aligned} S_t^{-1} S_b &= (\hat{E}\hat{\Lambda}\hat{E}^T)^{-1} ((W^{-1})^T \Sigma W^{-1}) \\ &= \hat{E}\hat{\Lambda}^{-1} \hat{E}^T \hat{E} \hat{\Lambda}^{\frac{1}{2}} U \Sigma U^T \hat{\Lambda}^{\frac{1}{2}} \hat{E}^T \\ &= (\hat{E}\hat{\Lambda}^{-\frac{1}{2}} U) \Sigma (U^T \hat{\Lambda}^{\frac{1}{2}} \hat{E}^T) = W \Sigma W^{-1} \end{aligned}$$

That is, W and Σ are the eigenvector and eigenvalue matrices of $S_t^{-1} S_b$, respectively, and the transformed matrix W is the desired discriminant projection matrix.

The SMT-LDA algorithm is shown in Algorithm 1.

2.4 Analysis of SMT-LDA

We provide an analysis of SMT-LDA from the viewpoint of eigenvalues of the scatter matrices. In general, the distance between samples in different classes is bigger than the related distance between samples in the same class [21], so for most of the eigenvalues of S_b and S_w , one can have the inequality $\lambda_{b,i} > \lambda_{w,i}$. As $S_t = S_b + S_w$, we can get $\lambda_{b,i} < \lambda_{t,i} < 2\lambda_{b,i}$.

According to the maximum Rayleigh quotient criterion of LDA in (6), an eigenvector with eigenvalues satisfying $0.5 < \frac{\lambda_{b,i}}{\lambda_{t,i}} < 1$ means that samples in different classes are well separated (on average) in the direction of this eigenvector. In contrast, samples from different classes overlap in the direction of the eigenvectors with $\frac{\lambda_{b,i}}{\lambda_{t,i}} < 0.5$. For SMT-LDA,

Algorithm 1 SMT-LDA

- 1: Compute $S_t = \bar{X} \bar{X}^T$, and perform the SMT:
 - 1) Determine the coordinate set, for $1 \leq i \leq n$,
 $[\text{MaxJ}(i), \text{MaxC}(i)] \leftarrow$
 $[\arg \max_{i < j} \frac{|\bar{X}(i,:) \bar{X}(j,:)^T|}{\|\bar{X}(i,:)\| \|\bar{X}(j,:)\|}, \arg \max_{i < j} \frac{|\bar{X}(i,:) \bar{X}(j,:)^T|}{\|\bar{X}(i,:)\| \|\bar{X}(j,:)\|}]$
 - 2) For $k = 1 : K$
 - (a) $i_k \leftarrow \arg \max_i \text{MaxC}(i), j_k \leftarrow \text{MaxJ}(i_k)$
 - (b) $\theta_k \leftarrow \frac{1}{2} \text{atan}(-2S_{i_k j_k}, S_{i_k i_k} - S_{j_k j_k})$
 - (c) $E_k \leftarrow I + \Theta(i_k, j_k, \theta_k)$
 - (d) $\bar{X} \leftarrow E_k^T \bar{X}$
 - (e) Update the coordinate set
- End For
- 2: SMT estimation of the total scatter matrix: $S_t = E \Lambda E^T$, where $\Lambda = \text{diag}(\bar{X} \bar{X}^T), E = \prod_{k=1}^K E_k$
- 3: Transform for the between-class scatter matrix:
 - 1) Compute $\Phi_b = [\sqrt{N_1}(\mu_1 - \mu), \dots, \sqrt{N_c}(\mu_c - \mu)]$, such that $S_b = \Phi_b \Phi_b^T$
 - 2) Denote $H = E \Lambda^{-\frac{1}{2}}, \tilde{\Phi}_b = H^T \Phi_b$, then $\tilde{S}_b = \tilde{\Phi}_b \tilde{\Phi}_b^T = H^T S_b H$
 - 3) Eigen-decomposition: $\tilde{\Phi}_b^T \tilde{\Phi}_b = V \Sigma^{\frac{1}{2}} V^T$
 - 4) $U = \tilde{\Phi}_b V$ is the eigenvector of \tilde{S}_b , that is $U^T \tilde{S}_b U = (HU)^T S_b HU = \Sigma$
- 4: The optimal discriminant projection is:
 $W = HU = E \Lambda^{-1} E \Phi_b V$

based on the criterion (7) and Algorithm 1, an eigenvector with eigenvalue satisfying $0.5 < \sigma_i < 1$ means that samples in different classes are well separated in the direction of this eigenvector, where σ_i is the diagonal element of Σ .

Taking Salinas data set (see in next Section) as an example, we compare the ratio of eigenvalues of S_b and S_t in LDA and SMT-LDA methods. To evaluate the performance of the algorithms in small-sized training samples situation, the number of training samples for each class is set to 5. We compute the between-class scatter matrix S_b , total scatter matrix S_t , and SMT transformed between-class scatter matrix \tilde{S}_b . Then, we find the corresponding eigenvalues and show the ratio of eigenvalues in Fig. 2, where the number of eigenvalues is 15, which equals to the number of classes minus one. From the figure, we can see only four eigenvectors with eigenvalues satisfying $0.5 < \frac{\lambda_{b,i}}{\lambda_{t,i}} < 1$ in traditional LDA, while eight eigenvectors with eigenvalues satisfying $0.5 < \sigma_i < 1$ in SMT-LDA. Moreover, the ratio of eigenvalues is relatively small for LDA. By using the SMT, the eigenvalues σ_i s are larger than the corresponding parts in LDA. Based on the maximum Rayleigh quotient criterion, SMT-LDA is more discriminative than LDA, especially in this small-sample-size case.

In the following, we show the eigenvalues of S_t before and after SMT in Fig. 3, where S_t has only 71 eigenvalues in this case. To better display the results, the first several eigenvalues are truncated. The last 50 eigenvalues of S_t in LDA are close to 0, while the corresponding eigenvalues in SMT-LDA are much larger. As the null space of S_t contains discriminative information, SMT-LDA is much effective in

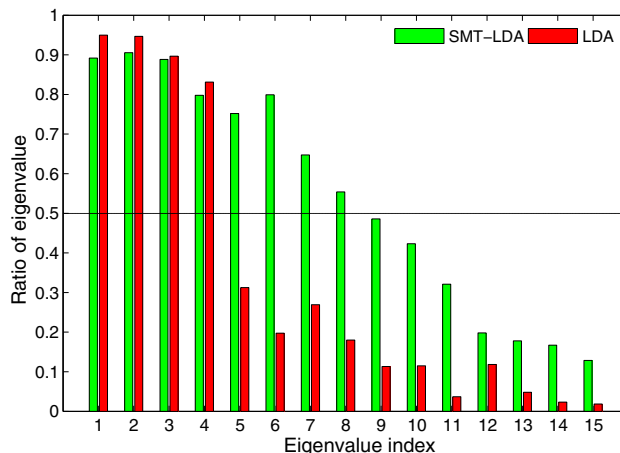


Fig. 2 Ratio of eigenvalues in LDA and SMT-LDA

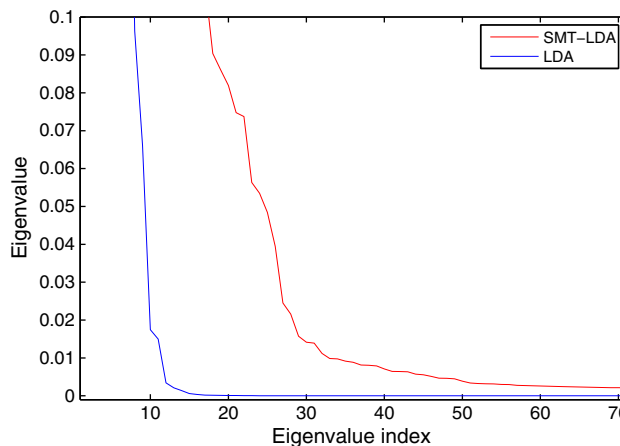


Fig. 3 Eigenvalues of S_t in LDA and SMT-LDA

keeping the intrinsic information and more discriminative than the original LDA.

3 Experimental results and discussion

In this section, we demonstrate the effectiveness of the proposed SMT-LDA algorithm for classification of two hyperspectral data sets. The 1-nearest neighbor (NN) and SVM classifiers are used. The results for classification are compared to those obtained by the RLDA [10] and PCA-LDA [14]. The results on the original data are also included. All data used in this paper are normalized to have a range of [0, 1].

3.1 Hyperspectral data

(1) Salinas: The data were acquired by the 224-band AVIRIS sensor over Salinas Valley in Southern California, USA, at

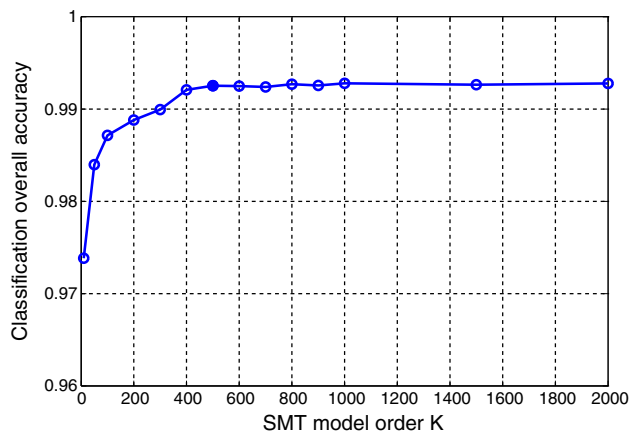


Fig. 4 Effect of the SMT model order K

low altitudes, resulting in an improved pixel resolution of 3.7 meter per pixel. The area covered comprises 512 lines by 217 samples. Twenty spectral bands are removed due to water absorption and noise, resulting in a corrected image containing 204 spectral bands over the range of 0.4 to 2.5m. The Salinas scene consists of the 16 ground-truth classes and 54,129 samples.

(2) Indian Pines: The data were acquired by the AVIRIS sensor in 1992. The image contains 145×145 pixels and 220 bands, where 20 channels were discarded because of atmospheric affection. Sixteen different land-cover classes are available in the original ground truth. The number of samples is 10,249 ranging from 20 to 2455 in each class.

3.2 Investigation on SMT model order

In this subsection, we investigate the effect of SMT model order K (i.e., the number of Givens rotations) on the final results. For this purpose, we select a small subscene of Salinas image, denoted Salinas-A, which comprises 86×83 pixels located within the scene at [samples, lines] = [591–676, 158–240] and includes six classes. We show the SMT-LDA classification overall accuracy on different model orders K in Fig. 4, where 30 samples in each of the six classes are randomly chosen for training and the rest samples are used for validating. The experiment is repeated ten times, and the averaged overall accuracy is reported. It can be seen that SMT-LDA algorithm is stable when K is not less than 500. In the following, we set K to be 500.

3.3 Comparison results

To evaluate the performance of different algorithms in the challenging situations with high dimensionality and small-sized training samples, the number of training samples for

each class is set to 5, 10, 15, 20, 25, and 30, respectively. The remaining samples form the testing set. In each case, the experiment is repeated ten times with randomly chosen training samples. Finally, the ten times results are averaged. The proposed SMT-LDA method is compared with other traditional LDA methods, including regularized LDA (RLDA) and subspace LDA (PCA-LDA). In the case of limited training samples, LDA does not work well as the problem of an unstable matrix inversion. RLDA alleviates the ill-posed problem by shrinking the scatter matrix toward the identity matrix or the diagonal of scatter matrix [10, 13]. In the experiments, the regularization form in RLDA is: $\hat{S} = S + \eta \cdot \text{diag}\{S\}$, where $\eta = 0.1$. In order to overcome the singularity of within-class scatter matrix S_w , PCA-LDA [14] first employs PCA to discard the null space of S_w and then applies LDA in the lower-dimensional PCA subspace. When S_w is nonsingular, PCA step is not performed and PCA-LDA reverts back to LDA. The classification results on the original data without dimensionality reduction are also included for comparison. The threefold cross-validation is used to select the optimal penalty parameters C and RBF kernel parameter γ in SVM.

The comparison results of the proposed method with the traditional LDA methods on the two HSI data sets are shown in Figs. 5 and 6. These two figures show the overall accuracy versus the number of training samples in each class. As expected, the classification accuracy increases as the training samples increase except for PCA-LDA. When the number of training samples N is smaller than the number of features d , PCA-LDA first employs PCA to reduce the dimension of the feature space to $N - c$ such that S_w is no longer degenerate and then applies the standard LDA to reduce the dimension to $c - 1$, where c is the number of classes. Take Salinas data set for example, the number of feature is $d = 204$ and the number of class is $c = 16$. When the number of samples in each class N_c is equal or greater than 15, then the number of total training samples N is larger than the number of features d ($N = c \times N_c \geq 16 \times 15 > d = 204$), so S_w is nonsingular and no PCA step is used in PCA-LDA. In this case, PCA-LDA is the same as LDA. In the cases of five and ten training samples per class, PCA-LDA first employs PCA to reduce the dimension of original space to 64 ($N - c = 16 \times 5 - 16 = 64$) and 144 ($N - c = 16 \times 10 - 16 = 144$), respectively, and then applies the standard LDA to reduce the dimension of PCA subspace to 15 ($c - 1 = 15$). The PCA step removes the redundant information, but it discards discriminative information at the same time. It is difficult to choose an optimal reduced dimension in PCA step.

Compared with RLDA, SMT-LDA achieves better classification results on the Salinas data set, especially in the small sample situations. On Indian Pines data set, SMT-LDA outperforms traditional approaches significantly even with limited training samples.

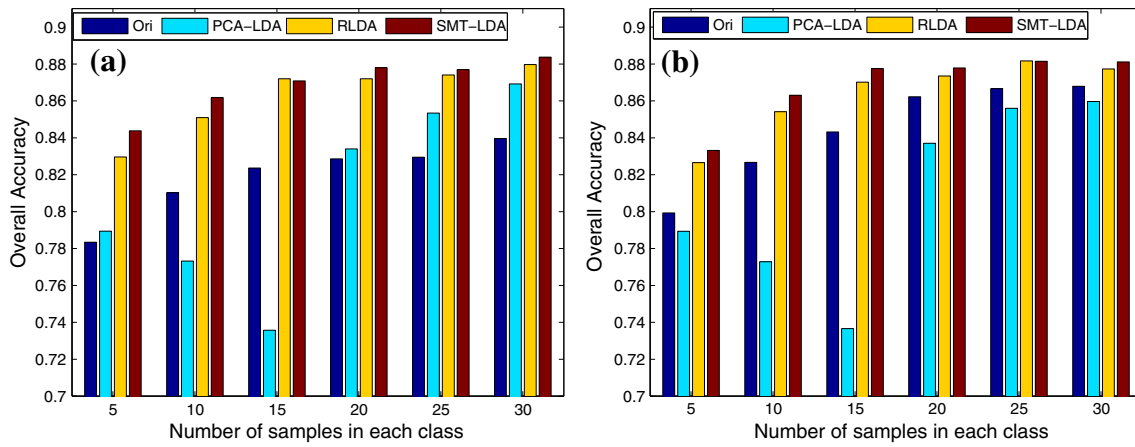


Fig. 5 Classification overall accuracy on Salinas data set. **a** NN, **b** SVM

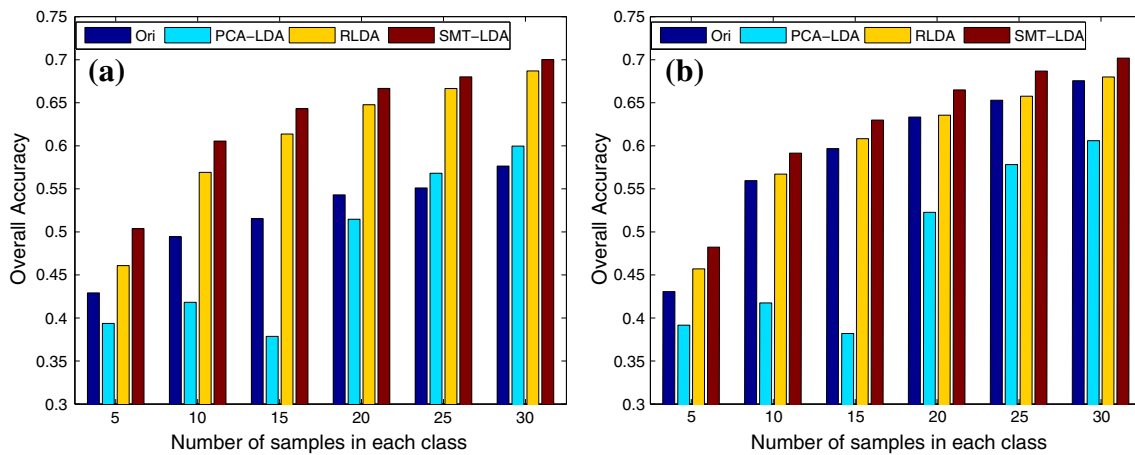


Fig. 6 Classification overall accuracy on Indian Pines data set. **a** NN, **b** SVM

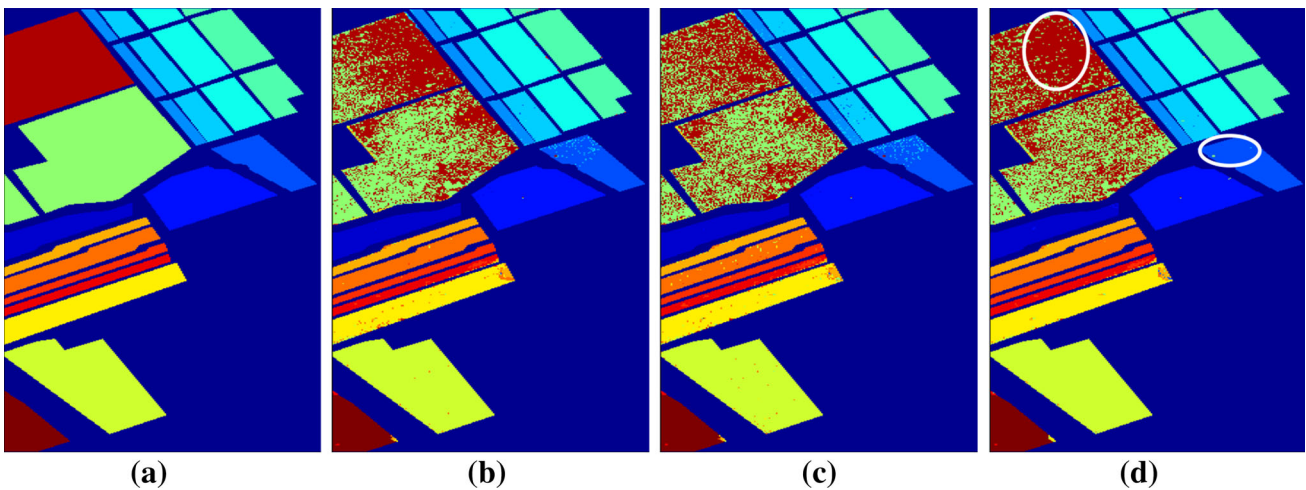


Fig. 7 Classification maps on Salinas data set. **a** Ground truth, **b** PCA-LDA, **c** RLDA, **d** SMT-LDA

The visual classification maps of PCA-LDA, RLDA, and SMT-LDA on the Salinas and Indian Pines data sets with 30 training samples per class are shown in Figs. 7 and 8. It can

be seen that SMT-LDA results in a more accurate map than traditional LDA methods, seeing the circled areas in Figs. 7d and 8d.

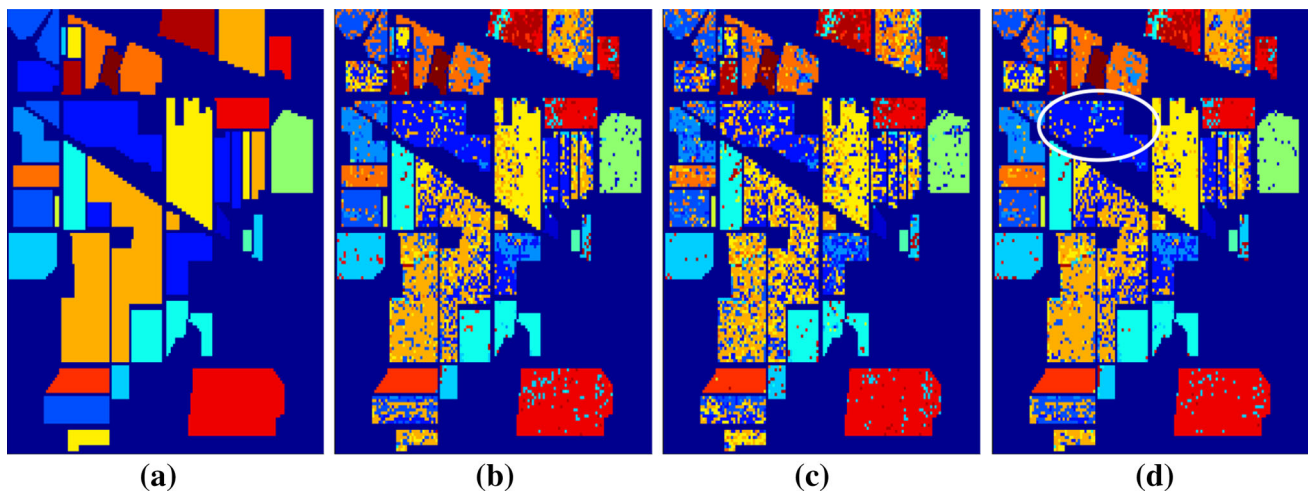


Fig. 8 Classification maps on Indian Pines data set. **a** Ground truth, **b** PCA-LDA, **c** RLDA, **d** SMT-LDA

Table 1 McNemar’s test on the Salinas data set to validate whether the difference between SMT-LDA and each of other LDA methods is statistical significant

N_c	NN			SVM		
	Ori	PCA-LDA	RLDA	Ori	PCA-LDA	RLDA
5	−34.75	−31.22	−9.90	−19.91	−24.96	−3.65
10	−29.87	−46.70	−7.93	−22.77	−48.64	−6.51
15	−28.59	−63.52	0.81	−22.55	−67.13	−5.72
20	−29.97	−26.66	−4.56	−10.97	−26.48	−3.54
25	−27.75	−15.29	−2.10	−10.49	−17.50	−0.15
30	−26.12	−9.35	−2.83	−9.27	−14.83	−3.46

Table 2 McNemar’s test on the Indian Pines data set to validate whether the difference between SMT-LDA and each of other LDA methods is statistical significant

N_c	NN			SVM		
	Ori	PCA-LDA	RLDA	Ori	PCA-LDA	RLDA
5	−14.75	−22.83	−11.11	−10.96	−19.05	−6.23
10	−21.08	−31.82	−9.17	−6.93	−30.29	−6.47
15	−23.67	−40.94	−7.23	−6.89	−39.27	−5.70
20	−22.70	−26.42	−4.70	−6.66	−26.02	−7.79
25	−23.62	−20.18	−3.47	−7.46	−20.61	−7.47
30	−22.71	−18.34	−3.32	−5.81	−18.75	−5.63

In order to validate whether the differences in accuracy between SMT-LDA and other LDA-based methods are statistically significant, we perform the McNemar’s test for each algorithm. The McNemar’s test [22] computes the Z-statistics as follows:

$$Z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \tag{8}$$

where f_{12} is the number of test samples that are erroneously classified by SMT-LDA and not by the comparison method and f_{21} has a dual meaning [22]. Accepting the common 5% level of significance, the difference between the results of SMT-LDA and of each compared method is

statistically significant if $|Z| > 1.96$ [22]. When this condition is met, a negative or positive value of Z indicates that SMT-LDA or the compared method is more accurate, respectively.

Tables 1 and 2 show the statistical test results using McNemar’s test on the Salinas and Indian Pines data sets, respectively. On Salinas, $Z < -1.96$ is obtained when comparing SMT-LDA with each previous LDA technique except for the RLDA, where the difference in accuracy between SMT-LDA and RLDA is not significant in the cases of 15 labeled samples per class for training with an NN classifier and 25 labeled samples per class for training with an SVM classifier. On Indian Pines, the differences between the

accuracies of SMT-LDA and of other LDA methods are statistically significant ($Z < -3$) in all cases.

From the above results, it can be seen that SMT-LDA provides more accurate predictions and outperforms traditional LDA methods in terms of overall classification performance. In fact, SMT-LDA can be also considered as a regularized method for estimating the covariance or scatter matrix [16]. The improvement over LDA methods indicates the benefits of using a sparse matrix transform technique for estimating the scatter matrix in solving the optimal discriminant vector.

4 Conclusions

In this paper, we have proposed an SMT-LDA method for dimension reduction and classification of hyperspectral remote sensing image. Because the available training samples in HSI classification are usually very limited, traditional LDA method is typically unstable. By representing the eigen-decomposition of total scatter matrix as an SMT, the total scatter matrix can be efficiently estimated and the estimator is always positive definite, which overcomes the singularity problem in traditional LDA algorithm. Experimental results demonstrate that SMT-LDA is usually more accurate, especially in small-sample-size cases.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China under Grants Nos. 61306070 and 11371007, by the Natural Science Foundation of Hubei Province under Grant No. 2015CFB327, and by the special fund from the State Key Joint Laboratory of Environment Simulation and Pollution Control (Research Center for Eco-environmental Sciences, Chinese Academy of Sciences) (Project No. 15K02ESPCR). The authors would like to thank Prof. D. Landgrebe for providing the Indian Pines data set, Prof. C.A. Bouman for sharing SMT codes, and Prof. C. Lin for providing LIBSVM toolbox.

References

- Borengasser, M., Hungate, W.S., Watkins, R.: *Hyperspectral Remote Sensing-Principles and Applications*. CRC Press, Boca Raton (2008)
- Plaza, A., Benediktsson, J.A., Boardman, J.W., et al.: Recent advances in techniques for hyperspectral image processing. *Remote Sens. Environ.* **113**, S110–S122 (2009)
- Hughes, G.F.: On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inform. Theory* **14**, 55–63 (1968)
- Harsanyi, J.C., Chang, C.I.: Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. *IEEE Trans. Geosci. Remote Sens.* **32**(4), 779–785 (1994)
- Benediktsson, J.A., Pesaresi, M., Arnason, K.: Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *IEEE Trans. Geosci. Remote Sens.* **41**, 1940–1949 (2003)
- Zhou, Y., Peng, J., Chen, C.L.: Dimension reduction using spatial and spectral regularized local discriminant embedding for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **53**(2), 1082–1095 (2015)
- Lee, Y.P.: Palm vein recognition based on a modified (2D)²LDA. *Signal Image Video Process.* **9**, 229–242 (2015)
- Wang, Y., Ni, H., Liu, P., Li, W.: Improved SDA based on mixed weighted Mahalanobis distance. *Signal Image Video Process.* doi:10.1007/s11760-014-0703-y
- Yu, H., Yang, J.: A direct LDA algorithm for high-dimensional data—with application to face recognition. *Pattern Recogn.* **34**(10), 2067–2070 (2001)
- Friedman, J.: Regularized discriminant analysis. *J. Am. Stat. Assoc.* **84**(405), 165–175 (1989)
- Hoffbeck, J.P., Landgrebe, D.A.: Covariance matrix estimation and classification with limited training data. *IEEE Trans. Pattern Anal. Mach. Intell.* **18**(7), 763–767 (1996)
- Chen, H., Pan, Z., Li, L., Tang, Y.: Learning rates of coefficient-based regularized classifier for density level detection. *Neural Comput.* **25**(4), 1107–1121 (2013)
- Bandos, T.V., Bruzzone, L., Camps-Valls, G.: Classification of hyperspectral images with regularized linear discriminant analysis. *IEEE Trans. Geosci. Remote Sens.* **47**(3), 862–873 (2009)
- Belhumeur, P.N., Hespanha, J., Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 711–720 (1997)
- Yektaei, M., Bhattacharya, P.: A criterion for measuring the separability of clusters and its applications to principal component analysis. *Signal Image Video Process.* **5**, 93–104 (2011)
- Cao, G.Z., Bouman, C.A.: Covariance estimation for high dimensional data vectors using the sparse matrix transform. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) *Advances in Neural Information Processing Systems*, vol. 21, pp. 225–232. MIT Press, Cambridge (2009)
- Cao, G.Z., Bachega, L.R., Bouman, C.A.: The sparse matrix transform for covariance estimation and analysis of high dimensional signals. *IEEE Trans. Image Process.* **20**(3), 625–640 (2011)
- Theiler, J., Cao, G.Z., Bachega, L.R., Bouman, C.A.: Sparse matrix transform for hyperspectral image processing. *IEEE J. Sel. Top. Signal Process.* **5**(3), 424–437 (2011)
- Peng, J.: Sparse matrix transform based weight updating in partial least squares regression. *J. Math. Chem.* **52**(8), 2197–2209 (2014)
- Liu, K., Cheng, Y.Q., Yang, J.Y., Liu, X.: An efficient algorithm for Foley–Sammon optimal set of discriminant vectors by algebraic method. *Int. J. Pattern Recogn. Artif. Intell.* **6**(5), 817–829 (1992)
- Dornaika, F., Bosaghzadeh, A.: Exponential local discriminant embedding and its application to face recognition. *IEEE Trans. Cybern.* **43**(3), 921–934 (2013)
- Foody, G.M.: Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogramm. Eng. Remote Sens.* **70**(5), 627–633 (2004)