

Locality-Preserving Dimensionality Reduction and Classification for Hyperspectral Image Analysis

Wei Li, *Student Member, IEEE*, Saurabh Prasad, *Member, IEEE*, James E. Fowler, *Senior Member, IEEE*, and Lori Mann Bruce, *Senior Member, IEEE*

Abstract—Hyperspectral imagery typically provides a wealth of information captured in a wide range of the electromagnetic spectrum for each pixel in the image; however, when used in statistical pattern-classification tasks, the resulting high-dimensional feature spaces often tend to result in ill-conditioned formulations. Popular dimensionality-reduction techniques such as principal component analysis, linear discriminant analysis, and their variants typically assume a Gaussian distribution. The quadratic maximum-likelihood classifier commonly employed for hyperspectral analysis also assumes single-Gaussian class-conditional distributions. Departing from this single-Gaussian assumption, a classification paradigm designed to exploit the rich statistical structure of the data is proposed. The proposed framework employs local Fisher's discriminant analysis to reduce the dimensionality of the data while preserving its multimodal structure, while a subsequent Gaussian mixture model or support vector machine provides effective classification of the reduced-dimension multimodal data. Experimental results on several different multiple-class hyperspectral-classification tasks demonstrate that the proposed approach significantly outperforms several traditional alternatives.

Index Terms—Dimensionality reduction, Gaussian-mixture-model (GMM), hyperspectral data, local discriminant analysis, support vector machine.

I. INTRODUCTION

STATISTICAL PATTERN-CLASSIFICATION systems for the analysis of hyperspectral imagery (HSI) typically employ dimensionality reduction followed by classification in order to learn statistical models for each class in the reduced-dimension feature space, subsequently using that information to classify unlabeled HSI pixels/samples. Dimensionality-reduction algorithms [1], as the name suggests, are typically designed to reduce the dimensionality of the feature space without losing desirable information. HSI data typically have hundreds (even thousands) of spectral bands per pixel, and these bands are often highly correlated. Dimensionality reduction seeks to decrease computational complexity and ameliorate statistical ill-conditioning by discarding redundant features that can potentially deteriorate classification performance [2], [3]. Popu-

lar dimensionality-reduction techniques include unsupervised approaches, such as principal component analysis (PCA) and independent component analysis (ICA), as well as supervised approaches, such as Fisher's linear discriminant analysis (LDA) [4], [5]. There are numerous variants of these techniques. For example, in [6], segmented PCA is applied to group original bands of the HSI data into subsets of highly correlated adjacent bands which are, however, suboptimal at best for a general pattern-classification problem [7]. After dimensionality reduction, classification is applied. A popular parametric classification strategy typically employed after dimensionality reduction is based on the maximum-likelihood estimation (MLE) [8] of posterior probabilities.

A key limitation to techniques such as PCA, LDA, MLE, and their variants is that they assume that the class-conditional distributions are Gaussian [9]. However, real-life observational data are often not Gaussian and, in extreme cases, are actually strongly multimodal. PCA and LDA are likely to fail as dimensionality-reduction techniques under such conditions. In this paper, we propose a classification paradigm that is designed to exploit the rich statistical structure of the data. It does not make the simplifying single-Gaussian assumption, and performs effective dimensionality reduction and classification of highly non-Gaussian multimodal high-dimensional HSI data. Toward that goal, we adopt a recently proposed local Fisher's discriminant analysis (LFDA) [10] to reduce the dimensionality of HSI data before employing a Gaussian-mixture-model (GMM) classifier or a support-vector-machine (SVM) classifier. Unlike LDA, LFDA is designed to handle multimodal non-Gaussian class distributions, and preserves the underlying structure of such distributions in the projection.

GMM [11] classifiers have proved beneficial for a variety of classification tasks, such as speech and speaker recognition, biometrics, etc. Although some preliminary studies of GMMs have been undertaken for HSI analysis [12], [13], GMMs are not a popular tool within the hyperspectral-classification community. The fundamental hesitation most researchers have when employing a technique such as GMMs for HSI analysis is the impractical size of the resulting parameter space. Learning such high-dimensional parameter vectors using limited (and costly) ground-truth/training data is highly impractical. On the other hand, employing conventional linear dimensionality-reduction techniques such as PCA and LDA as preprocessing often destroys the underlying multimodal structure of the data, rendering GMMs ineffective. In contrast, in this paper, we adopt locality-preserving dimensionality reduction—specifically, LFDA—for GMM classification.

Manuscript received April 13, 2011; revised July 19, 2011; accepted August 7, 2011. Date of publication October 3, 2011; date of current version March 28, 2012. This material is based upon the work supported in part by the National Science Foundation under Grant CCF-0915307 and in part by the National Geospatial-Intelligence Agency under Grant HM1582-10-1-0001.

The authors are with the Department of Electrical and Computer Engineering and the Geosystems Research Institute, Mississippi State University, Starkville, MS 39762 USA (e-mail: liwei089@ieee.org; saurabh.prasad@ieee.org; fowler@ece.msstate.edu; bruce@ece.msstate.edu).

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

Digital Object Identifier 10.1109/TGRS.2011.2165957

Furthermore, we demonstrate that an appropriately optimized LFDA preprocessing ensures that the GMM models learned in the HSI feature space do not have an unreasonably large number of free parameters to be estimated from training data.

LFDA significantly outperforms traditional supervised dimensionality-reduction tools, preserving the multimodal structure of the data in the reduced-dimension projected space. With that observation, we also test the benefit of LFDA for SVM [14] classifiers. SVMs seek to separate classes by learning an optimal decision hyperplane that best separates the training samples in a kernel-induced feature space. Nonlinear kernel projections within the SVM framework often help convert nonlinear separation in the input space to a linear separation in the kernel-induced space, wherein a margin-maximizing hyperplane classifier is employed. In recent work [15]–[17], SVMs have been shown to be effective for remote-sensing image classification, especially when the training-data-set size is small. In this paper, we also demonstrate that the locality-preserving property of LFDA yields very effective dimensionality reduction for SVM classifiers as well. We demonstrate the various practical aspects of the proposed algorithms pertinent to HSI classification, including optimizing system parameters (such as the dimensionality of the LFDA projected subspace, GMM initialization, number of Gaussian mixtures, kernel parameter for the SVM classifier, etc).

The remainder of this paper is organized as follows. In Section II, we discuss conventional dimensionality-reduction techniques and provide a motivation for LFDA-based dimensionality reduction for hyperspectral classification. We also provide a description of the LFDA algorithm as well as empirical evidence of its benefits with a synthetic data set. In Section III, we describe in detail the GMM and SVM classifiers used in this paper as well as the motivation for employing LFDA as a dimensionality-reduction preprocessing for classification of HSI data. In Section IV, we provide a description of the experimental hyperspectral data set used to validate the proposed approach, describe the experimental setup, and show how to optimize the proposed system. We validate the proposed approach with several popular hyperspectral remote-sensing data sets, studying the classification performance by comparing to current state-of-the-art parametric classification methods over a wide range of practical operating conditions, such as pixel mixing and reduced training-sample size. We conclude by summarizing our results and suggesting future directions in Section V.

II. DIMENSIONALITY REDUCTION

Dimensionality reduction is a critical preprocessing step for HSI analysis. Owing to the dense spectral sampling of HSI data, the associated spectral information in the hyperspectral bands is typically highly correlated and of very high dimension. Hence, dimensionality reduction is commonly applied as a preprocessing step to reduce the dimensionality of the data to ensure a well-conditioned representation of the class-conditional statistics. Common dimensionality-reduction methods include PCA, LDA, and their many variants, such as subspace LDA [7], stepwise LDA [3], etc. PCA seeks to find a linear transformation which projects the data from a high-dimensional space to a

lower dimensional subspace by maximizing the variance of the data in the projected subspace. The optimal projection in this sense is determined by the eigenvectors corresponding to the largest eigenvalues of the covariance matrix of the original data. PCA constitutes unsupervised dimensionality reduction and is commonly employed by researchers for classification and representation tasks. However, PCA provides, at best, suboptimal dimensionality reduction for classification tasks—that is, it is well understood that PCA can potentially discard information useful to the classification task at hand, particularly if such information is contained along the low-energy directions [7].

On the other hand, LDA is also commonly employed to project high-dimensional data onto a smaller dimensional subspace. However, LDA maximizes the between-class scatter while minimizing the within-class scatter (more specifically, it seeks to find a transformation that maximizes Fisher's ratio in the projected subspace). In that respect, under the assumption of homoscedastic Gaussian class-conditional distributions, LDA is optimized for classification tasks.

The LDA transformation is obtained by solving a generalized eigenvalue problem

$$S_b \Phi = \Lambda S_w \Phi \quad (1)$$

where Λ is the diagonal eigenvalue matrix, Φ is the transformation matrix, S_b is the between-class scatter matrix, and S_w is the within-class scatter matrix. Although it is designed to maximize class separation in the projected subspace (as measured by Fisher's ratio), it is still suboptimal in that it assumes that class-conditional distributions are Gaussian with a homoscedastic covariance structure. Such an approach will not perform well when the data are heteroscedastic and can completely break down if the data are multimodal. Finally, the dimensionality of the projected subspace after an LDA transformation is upper bounded by $c - 1$ by design (c is the number of classes in the classification task), which is another major drawback of LDA, particularly when the dimensionality of the input space is very high, and c is small. A significant amount of potentially useful information can be lost when the final dimensionality is drastically smaller than the dimensionality of the input space. A detailed analysis of PCA, LDA, and their variants can be found in [1] and [7].

A. LFDA

LFDA [10] has been recently proposed as an extension to LDA, which, by not restricting the class distributions to be unimodal Gaussian, is expected to outperform LDA significantly for many practical classification situations. LFDA combines the properties of LDA and locality-preserving projections (LPP) [18]. Unlike LDA or PCA, LPP is a linear manifold-learning technique that seeks to find a linear map that preserves the local structure of neighboring samples in the input space. In other words, after an LPP mapping, neighborhood points in the original input space remain neighbors in the LPP-embedded space, and vice versa. In [10], a detailed description of LPP and LFDA is provided. By invoking a similar idea exploited in LPP, LFDA obtains good between-class separation in the

projection while preserving the within-class local structure (i.e., neighboring data pairs in the original space remain close under the projection) at the same time. It is hence expected that LFDA will surpass LDA and LPP as a dimensionality-reduction projection when the data are significantly non-Gaussian or even severely multimodal.

Consider a data set with training samples $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ in \mathbb{R}^d (d -dimensional feature space) and class labels $y_i \in \{1, 2, \dots, c\}$, where c is the number of classes and n is the total number of training samples. Let n_l be the number of available training samples for the l th class, $\sum_{l=1}^c n_l = n$. Define $A_{i,j} \in [0, 1]$ as the affinity between \mathbf{x}_i and \mathbf{x}_j

$$A_{i,j} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\gamma_i \gamma_j}\right) \quad (2)$$

where $\gamma_i = \|\mathbf{x}_i - \mathbf{x}_i^{(k_{nn})}\|$ denotes the local scaling of data samples in the neighborhood of \mathbf{x}_i , and $\mathbf{x}_i^{(k_{nn})}$ is the k_{nn} -nearest neighbor of \mathbf{x}_i . $A_{i,j}$ is then a symmetric matrix (referred to as the affinity matrix) of size $n \times n$ which measures the distance among data samples. Note that there are clearly many different ways to define an affinity matrix, but the *heat kernel*, as defined in (2), has been shown to result in very effective locality-preserving properties.

In LFDA, the *local* between-class $S^{(lb)}$ and within-class $S^{(lw)}$ scatter matrices are defined as

$$S^{(lb)} = \frac{1}{2} \sum_{i,j=1}^n W_{i,j}^{(lb)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \quad (3)$$

$$S^{(lw)} = \frac{1}{2} \sum_{i,j=1}^n W_{i,j}^{(lw)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \quad (4)$$

where $W^{(lb)}$ and $W^{(lw)}$ are $n \times n$ matrices defined as

$$W_{i,j}^{(lb)} = \begin{cases} A_{i,j}(1/n - 1/n_l), & \text{if } y_i = y_j = l \\ 1/n, & \text{if } y_i \neq y_j \end{cases} \quad (5)$$

$$W_{i,j}^{(lw)} = \begin{cases} A_{i,j}/n_l, & \text{if } y_i = y_j = l \\ 0, & \text{if } y_i \neq y_j. \end{cases} \quad (6)$$

Maximizing Fisher's ratio as defined using the local scatter matrices, we have that

$$\Phi_{\text{LFDA}} = \arg \max_{\Phi_{\text{LFDA}}} \text{tr}$$

$$\left[\left(\Phi_{\text{LFDA}}^\top S^{(lw)} \Phi_{\text{LFDA}} \right)^{-1} \Phi_{\text{LFDA}}^\top S^{(lb)} \Phi_{\text{LFDA}} \right] \quad (7)$$

is given by $S^{(lb)} \Phi_{\text{LFDA}} = \Lambda S^{(lw)} \Phi_{\text{LFDA}}$, where Λ is the diagonal eigenvalue matrix and $\Phi_{\text{LFDA}} \in \mathbb{R}^d$ is the transformation matrix.

It is readily seen that, in LFDA, the global between- and within-class scatter matrices in the original expression for Fisher's ratio are replaced by their local versions defined in (3) and (4). LFDA can thus be viewed as a localized variant of LDA because it does not force far-apart data pairs of the same class to be close. Another way to picture this is as follows. By design,

the contribution of samples within a class that are far apart to the scatter matrix is very small, while that of samples that are close to each other is significantly higher. On the other hand, in traditional LDA, all samples within a class contribute equally to the scatter matrices. When class-conditional distributions are significantly multimodal, traditional LDA fails because in estimating *global* scatter across all samples within a class, the local structure of the samples distributed over the various modes is lost. LFDA, on the other hand, treats samples of a class within each cluster/mode independently when estimating the scatter matrices, thereby preserving local neighborhoods even when the data distributions are complex. Hence, the linear projection learned using LFDA can be expected to maximize Fisher's ratio (from between- to within-class scatter) even when input class-conditional distributions are multimodal. Clearly, when $A_{i,j} = 1 \forall i, j$, LFDA degenerates to traditional LDA. Due to the data-dependent weighting by W in the estimation of the scatter matrices, unlike LDA, the final dimensionality after an LFDA projection is not upper bounded by $c - 1$.

Examples of dimensionality reduction for a 2-D two-class multimodal synthetic data set using LFDA and LDA are shown in Figs. 1–3. Here, k_{nn} is chosen as seven, and the dimensionality of the projected subspace is chosen as one for both LDA and LFDA. Fig. 1 shows the original two-class multimodal classification problem, along with the projecting directions learned using LDA and LFDA. Figs. 2 and 3 show the histograms of the data in the LFDA and LDA projected subspaces, respectively. Note that LFDA preserves the multimodal structure of the data in the projected subspace. Another measure that quantifies this fact is the Kullback–Leibler (KL) distance [19]. The KL distance between classes 1 and 2 using LFDA (Fig. 2) is 4.5, while this distance using LDA (Fig. 3) is 1.4. These figures show that LFDA preserves the multimodal structure of the data in the projected subspace. Unlike LFDA, LDA can distort the information contained in multimodal distributions, often projecting them onto subspaces wherein a unimodal statistical structure is imposed. Hence, when using LFDA, in going from Figs. 1 and 2 (LFDA), the inherent Bayes error is approximately retained, but it increases significantly when going from Figs. 1–3 (traditional LDA).

Motivated by these properties, in this paper, we use LFDA as a dimensionality-reduction step for hyperspectral image classification. We propose that, combined with a classifier that can handle multimodal distributions, such as a GMM, the resulting classification is expected to accurately capture the class-conditional statistics in a reduced-dimension subspace, especially when classes are multimodal. The spectral response of remotely sensed data can be affected by many factors, such as differences in illumination conditions, geometric features of material surfaces, and atmospheric effects [20]. It is hence reasonable to expect that statistical distributions of classes/objects in a remotely sensed image will possess a complicated multimodal structure. Classifiers such as those based on GMMs are hence a natural fit for remotely sensed data [12]. Another common scenario where such multimodal structures would exist is when the spatial resolution of the acquired imagery is not fine enough to resolve the objects on ground, resulting in *mixed pixels*.

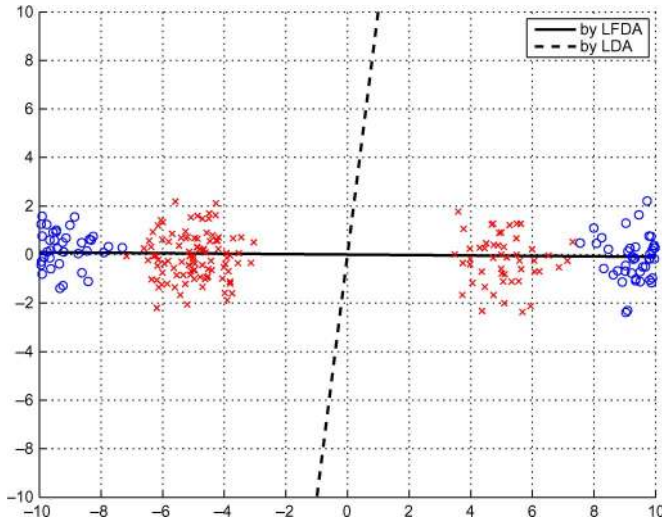


Fig. 1. Synthetic 2-D multimodal data and the directions of linear dimensionality-reduction projection as estimated using LDA and LFDA.

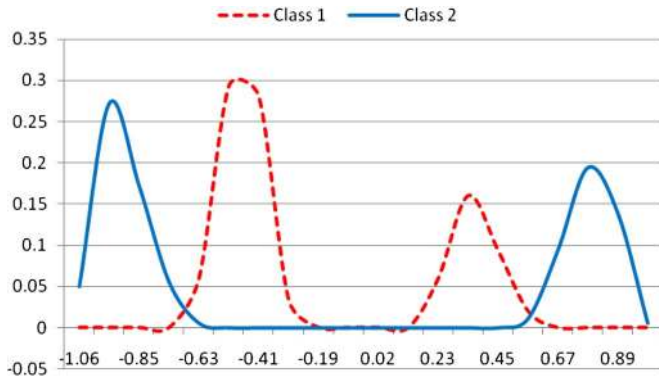


Fig. 2. Histogram of the synthetic data in the previous figure when projected onto a 1-D subspace using LFDA.

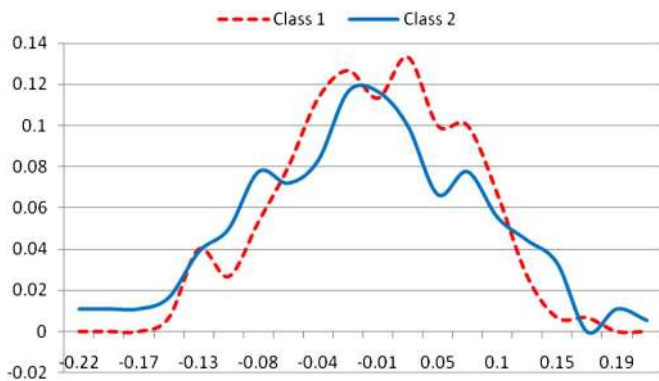


Fig. 3. Histogram of the synthetic data in the previous figure when projected onto a 1-D subspace using LDA.

One of the principal reasons why the HSI research community has shied away from using GMMs is the significant number of parameters that must be learned, necessitating large training-data set sizes. To the best of our knowledge, LFDA, as a preprocessing to GMMs, has not been explored in the literature, and in this paper, we demonstrate the benefit of doing so. We demonstrate how the LFDA transformation can be optimized for the HSI-classification task at hand, and how,

when combined with GMM classifiers, the resulting system outperforms traditional approaches. Furthermore, we demonstrate that the benefits of LFDA do not simply stop at GMMs, and in fact, LFDA is a very effective dimensionality-reduction tool for SVMs as well. We explore these ideas next.

III. PARAMETRIC CLASSIFICATION

The Gaussian maximum-likelihood classifier [8]—arguably one of the most commonly employed parametric classifiers for remote-sensing tasks—assumes Gaussian class-conditional statistics and relies on the first- and second-order statistics of the data. The discriminant (class-membership) function is given by

$$g_l(\mathbf{x}) = p(\mathbf{x}|C_l)P(C_l), \quad l = 1, 2, \dots, c \quad (8)$$

where $P(C_l)$ is the prior probability for the l th class label C_l and c is the number of classes. The likelihood function $p(\mathbf{x}|C_l)$ is assumed to take the parametric form $p(\mathbf{x}|C_l) \propto \mathcal{N}(\mu_l, \Sigma_l)$, where the mean vector μ_l and covariance matrix Σ_l are estimated from the training data. Gaussian MLE-based classification is simple (i.e., a small number of parameters need to be estimated) and has attractive convergence properties as the amount of training data increases [11]. However, a fundamental limitation of MLE classifiers is the assumption that class-conditional likelihoods are Gaussian—for a variety of classification tasks, this assumption can be inaccurate, resulting in suboptimal classification performance.

A. GMM

A GMM [11]–[13] can be viewed as a combination of two or more normal Gaussian distributions. In a typical GMM representation, a probability density function for $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ in \mathbb{R}^d is written as the sum of K Gaussian components (modes), i.e.,

$$p(\mathbf{x}) = \sum_{k=1}^K \alpha_k \mathcal{N}(\mathbf{x}, \mu_k, \Sigma_k) \quad (9)$$

where

$$\mathcal{N}(\mathbf{x}, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \times \exp \left[-\frac{1}{2} (\mathbf{x} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x} - \mu_k) \right]. \quad (10)$$

In (9), K is the number of mixture components, while α_k , μ_k , and Σ_k are the mixing weight, mean, and covariance matrix, respectively, of the k th component. These last three quantities can be expressed by the parameter vector $\Theta = \{\alpha_k, \mu_k, \Sigma_k\}$.

Once the optimal number of components (K) per GMM has been determined, the parameters for the mixture model can be estimated by the expectation–maximization (EM) algorithm [21], an iterative optimization strategy. The EM algorithm finds a (local) maximum-likelihood or maximum *a posteriori* (MAP) estimation of the parameters. Specifically, given a data set $\mathbf{X}_0 = \{\mathbf{x}_i\}_{i=1}^{n_l}$, \mathbf{x}_i is one data vector in the k th component

subset \mathcal{Q}_k , and n_l is the number of samples in \mathcal{Q}_k . The resulting complete data log-likelihood is

$$\mathcal{L}(\Theta, \mathbf{X}_0) = \sum_{k=1}^K \sum_{i \in \mathcal{Q}_k} p(k|\mathbf{x}_i, \Theta) \log [\alpha_k \mathcal{N}(\mathbf{x}_i, \mu_k, \Sigma_k)] \quad (11)$$

where $p(k|\mathbf{x}_i, \Theta)$ is the posterior probability for the k th component of the GMM and can be written as

$$p(k|\mathbf{x}_i, \Theta) = \frac{\alpha_k \mathcal{N}(\mathbf{x}_i, \mu_k, \Sigma_k)}{\sum_{k=1}^K \alpha_k \mathcal{N}(\mathbf{x}_i, \mu_k, \Sigma_k)}. \quad (12)$$

At each iteration, the parameter Θ is obtained by maximizing the likelihood function $\mathcal{L}(\Theta, \mathbf{X}_0)$ described in (11) (the M-step). The parameter set is then updated (12) with an expected value of these parameters for the next iteration (the E-step)

$$\hat{\alpha}_k = \frac{1}{n_l} \sum_{i=1}^{n_l} p(k|\mathbf{x}_i, \Theta) \quad (13)$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n_l} p(k|\mathbf{x}_i, \Theta) \mathbf{x}_i}{\sum_{i=1}^{n_l} p(k|\mathbf{x}_i, \Theta)} \quad (14)$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n_l} p(k|\mathbf{x}_i, \Theta) (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^\top}{\sum_{i=1}^{n_l} p(k|\mathbf{x}_i, \Theta)}. \quad (15)$$

The aforementioned procedure is iterated until the relative difference between successive values of the complete data log-likelihood provided by (11) reaches some predetermined convergence threshold [22].

Estimating an appropriate number of components/modes (K) is important to successful learning and deployment of GMMs for classification tasks. The Akaike information criterion (AIC) [23] is a commonly employed metric to estimate an optimal value for K . For the parameter vector Θ , the AIC is expressed in terms of the likelihood function as

$$\text{AIC}(\Theta) = -2\mathcal{L}_{\max}(\Theta, \mathbf{X}_0) + 2K \quad (16)$$

where $\mathcal{L}_{\max}(\Theta, \mathbf{X}_0)$ is the maximum log-likelihood function according to each model and K is the number of clusters to be estimated. The preferred model is the one with the minimum $\text{AIC}(\Theta)$ value.

The Bayes information criterion (BIC) [24] is another metric commonly used for estimating an optimal value of K in GMM models, and it is given as

$$\text{BIC} = -2\mathcal{L}_{\max}(\Theta, \mathbf{X}_0) + K \log(n) \quad (17)$$

where n is the total number of samples. It has been reported in the pattern-recognition community (i.e., [25]) that, for certain applications, the AIC tends to overestimate the value of K , while the BIC often yields a much smaller K and is hence more effective. In this paper, we study the efficacy of both the AIC and BIC for GMM-based classification of high-dimensional HSI.

As we mentioned previously, the size of the resulting parameter space often makes GMMs impractical in HSI-analysis tasks. For example, if the HSI data are d dimensional, then the resulting dimensionality of the parameter space for a K -component GMM, assuming full covariance matrices, is

$K(1 + d(d-1)/2) + Kd$. For $d = 100$ and $K = 10$ (a reasonable choice for the HSI dimensionality and number of mixture components, respectively), the resulting parameter space has dimension 50 510. Learning such high-dimensional parameter vectors using limited (and costly) ground-truth/training data is highly impractical.

We have demonstrated in Section II how LFDA can significantly outperform LDA as a dimensionality-reduction tool, preserving the multimodal structure of the data in the reduced-dimension projected space. Consider the hypothetical example provided earlier—for an LFDA projection of a 100-D space onto a 10-D subspace, the resulting parameter space for a ten-mixture GMM reduces from 50 510 down to 560. We hence argue that LFDA serves as an ideal dimensionality-reduction projection for GMM classifiers; we test this hypothesis for a challenging hyperspectral-classification task in the experimental results to follow in Section IV. First, we consider SVMs as an alternative to GMMs for classification.

B. SVM

For a training-data set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ in \mathbb{R}^d with class labels $y_i \in \{+1, -1\}$ and a nonlinear kernel mapping $\phi(\cdot)$, an SVM [16] classifies binary data by determining an optimal hyperplane in the kernel-induced space by solving

$$\min_{\omega, \xi_i, b} \left\{ \frac{1}{2} \|\omega\|^2 + \varsigma \sum_{i=1}^n \xi_i \right\} \quad (18)$$

subject to the constraints

$$y_i (\langle \phi(\omega, \mathbf{x}_i) \rangle + b) \geq 1 - \xi_i \quad (19)$$

for $\xi_i \geq 0$ and $i = 1, \dots, n$, where ω is normal to the optimal decision hyperplane (i.e., $\langle \omega, \phi(\mathbf{x}) \rangle + b = 0$), n denotes the number of samples, b is the bias term, ς is the regularization parameter which controls the generalization capacity of the machine, and ξ_i is the positive slack variable allowing one to accommodate permitted errors appropriately. The aforementioned problem is solved by maximizing its Lagrangian dual form [17]

$$\max_{\alpha} \left\{ \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right\} \quad (20)$$

where $\alpha_1, \alpha_2, \dots, \alpha_n$ are nonzero Lagrange multipliers constrained to $0 \leq \alpha_i \leq \varsigma$, and $\sum_i \alpha_i y_i = 0$, for $i = 1, \dots, n$. Some commonly implemented kernel functions are the linear kernel, the polynomial kernel, and the radial-basis-function (RBF) kernel [16]. In this paper, RBF is considered

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right) \quad (21)$$

where σ is a width parameter characterizing the RBF. Finally, the decision function is represented as

$$f(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^n y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \right). \quad (22)$$

As mentioned before, traditional SVMs are binary classifiers by design. Various approaches exist to extend the binary SVMs to tasks involving more than two classes [26], [27]. The most popular approach is one-against-all [28], which trains the SVM for every possible class pair. It is common to employ a backward feature-reduction algorithm—recursive feature elimination (RFE) [28]—to eliminate redundant features that do not contribute positively to the SVM classifier. The RFE approach focuses on retaining features that maximize the separation margin while minimizing the generalization error. RFE-SVM is therefore becoming increasingly popular in high-dimensional classification applications, such as HSI analysis [15]–[17]. Hence, it serves as a powerful baseline against which to compare our proposed methods.

In this paper, we demonstrate the use of LFDA as an alternate dimensionality-reduction tool for SVM classifiers. The locality-preserving quality of LFDA is the key motivation behind studying its benefits with SVM classifiers for HSI classification. By preserving locality and neighborhood relations, complex nonlinear decision surfaces in the input space are expected to be preserved in the low-dimensional LFDA projected subspace, wherein an SVM classifier can operate to attain linear separation in a kernel-induced space. Furthermore, since the LFDA embedding optimizes Fisher's ratio, one can be assured that the projected subspace will possess good class separation.

IV. EXPERIMENTAL RESULTS

In this section, we validate our approach with several popular HSI data sets and present experimental results demonstrating the benefits of LFDA-based dimensionality reduction for nonlinear classifiers such as GMMs and SVMs. We report the performance of classification systems as measured by the overall classification accuracy, along with the 95% confidence intervals for these estimates. The primary objectives of the experimental results reported in the next two subsections are as follows: 1) tuning the parameters of the classification system (dimensionality reduction and classification) for the HSI task at hand and 2) quantifying the efficacy of LFDA-based dimensionality reduction for HSI classification and comparing it to that of traditional state-of-the-art methods commonly employed by researchers in the HSI community, over a wide range of operating conditions (i.e., studying the sensitivity of the classification algorithm to the amount of training data used, as well as the extent of pixel mixing in the data set). All data used in this paper are normalized to have a range of [0, 1].

A. Experimental Hyperspectral Data

The first experimental HSI data set employed was acquired using the National Aeronautics and Space Administration's Airborne Visible/Infrared Imaging Spectrometer sensor and was collected over northwest Indiana's Indian Pines test site in June 1992.¹ The image represents a vegetation-classification scenario with 145×145 pixels and 220 bands in the 0.4- to

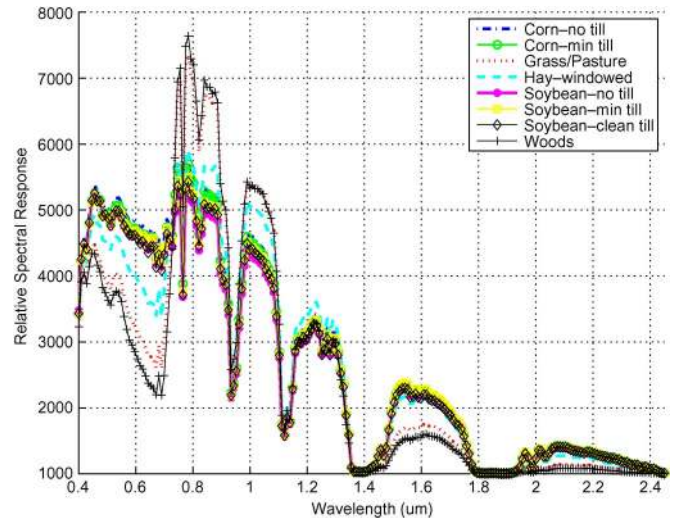


Fig. 4. Spectral signatures of eight classes from the Indian Pines data set.

2.45- μm region of the visible and infrared spectrum with a spatial resolution of 20 m. The main crops of soybean and corn in the image are in their early growth stage. The *no till*, *min till*, and *clean till* indicate the amounts of previous crop residue remaining. Fig. 4 shows the spectral signatures for the eight classes extracted from this imagery. Approximately 8600 labeled pixels are employed to train and validate/quantify the efficacy of the proposed system. This data set is partitioned into approximately 1496 training pixels and 7102 test pixels.

The other two data sets used in this paper were collected by the Reflective Optics System Imaging Spectrometer sensor [29]. The image, covering the city of Pavia, Italy, was collected under the HySens project managed by DLR (the German Aerospace Agency). The images have 115 spectral bands with a spectral coverage from 0.43 to 0.86 μm and a spatial resolution of 1.3 m. Two scenes are used in our experiment. The first one of these is the university area which has 103 spectral bands with a spatial coverage of 610×340 pixels. The second one is the Pavia city center which has 102 spectral bands with 1096×715 pixels formed by combining two separate images representing different areas of the Pavia city. Figs. 5 and 6 show the spectral signatures of the nine classes in this data set. The numbers of training and testing samples used for the University of Pavia data set are 1476 and 7380, respectively. The numbers of training and testing samples used for the Pavia Centre data set are 1477 and 8862, respectively.

B. Optimizing LFDA-SVM and LFDA-GMM

In this section, we report experiments demonstrating the sensitivity of the proposed LFDA-GMM and LFDA-SVM approaches over a wide range of the parameter space and show how this information can be used to optimize the system for any HSI-classification task. System parameters—such as dimensionality of the projected subspace, k_{nn} in the affinity matrix, and σ for the RBF kernel in SVMs—are optimized using training data. Development data are derived from the available training data by further dividing them into training and testing samples for tuning these parameters. The testing

¹ftp://ftp.ecn.purdue.edu/biehl/MultiSpec.

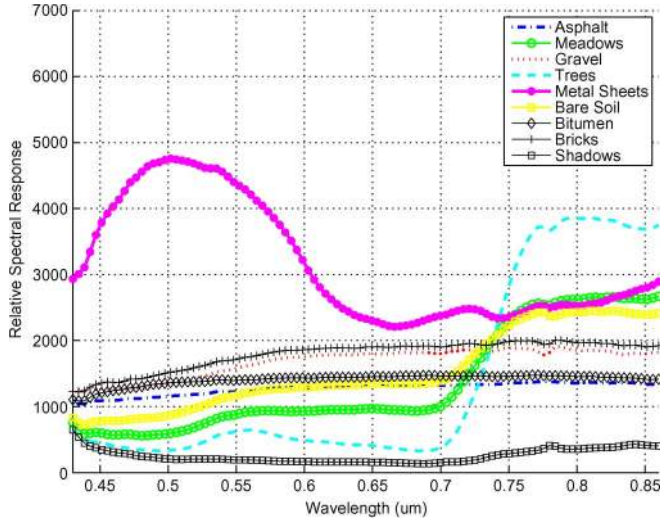


Fig. 5. Spectral signatures of nine classes from the University of Pavia data set.

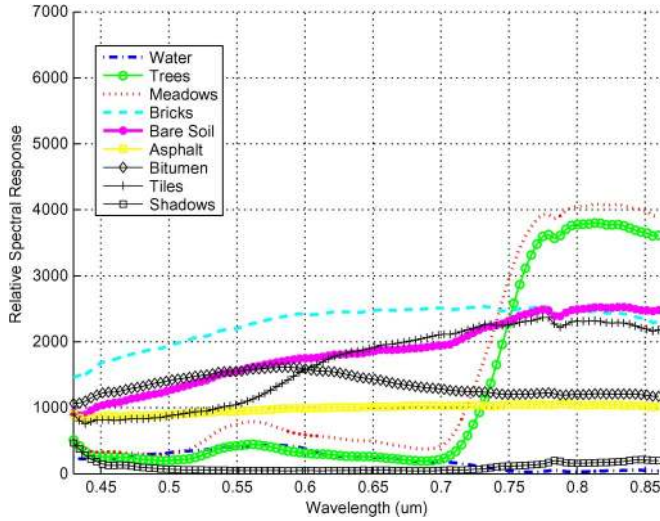


Fig. 6. Spectral signatures of nine classes from the Pavia Centre data set.

accuracy obtained using this development data set is used to gauge an effective range for the system parameters to ensure a reliable classification performance.

As we mentioned previously, unlike traditional LDA, the dimensionality of the LFDA projected subspace is not restricted to $c - 1$. As a result, when LFDA is applied for dimensionality reduction of hyperspectral data, we need to find an optimal dimensionality for the LFDA projection. It is also expected that the k_{nn} term used to estimate γ_i in (2) will affect the affinity matrix in LFDA. Fig. 7 shows the overall development-data accuracy of LFDA-GMM as a function of k_{nn} and reduced dimension for the University of Pavia HSI data set. It is clear from Fig. 7 that the system is sensitive to the choice of system parameters, but these follow a systematic trend, consistently seen across all the data sets. In particular, it can be seen that the performance peaks at small values of both the reduced dimension and k_{nn} . The optimal dimensionality of the LFDA projected subspace is expected to vary with the data set at hand. However, as can be seen in Fig. 7, an optimal value of

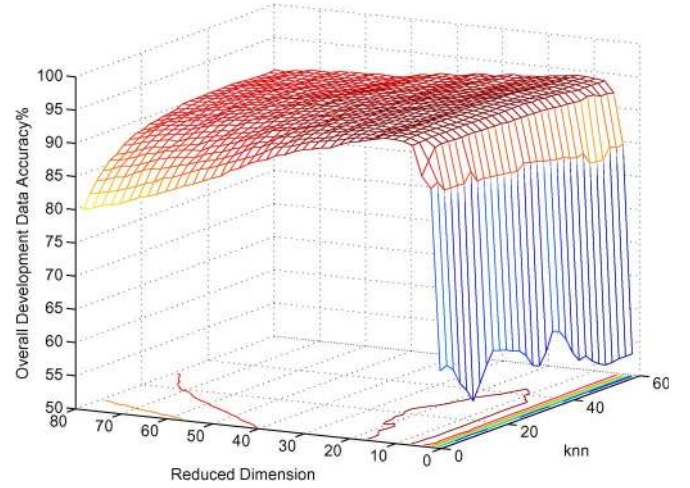


Fig. 7. LFDA-GMM for the University of Pavia data set: Overall development accuracy versus reduced dimension and parameter k_{nn} .

the reduced-dimension parameter for all data sets is obviously smaller than the true dimensionality of the input space. For example, a reduced dimension of ten appears to be optimal for the University of Pavia data set. Although the extent of dimensionality reduction is significant in such a projection, a GMM classifier trained on the LFDA projected subspace is expected to capture the underlying statistical structure accurately without the accompanying excessive overdimensionality of the resulting parameter vector that would have been experienced if it were trained on the original input space.

Fig. 8 shows the overall development-data accuracy of LFDA-SVM for the University of Pavia data set. The SVM performance is known to be sensitive to the width parameter (σ) for the RBF kernel in (21), as is LFDA to the reduced dimension and k_{nn} . We performed a grid search over a wide range of all these system parameters, studying the development-data accuracy as a function of these parameters. In a practical setting, one can study these performance curves to optimally tune the system. Fig. 9 shows the overall development-data accuracy as a function of reduced dimension for the RFE-SVM baseline algorithm with several values of σ . Table I summarizes the optimal parameters that we chose for LFDA-GMM, LFDA-SVM, and RFE-SVM. We note that the optimal reduced dimension for LFDA-SVM is obviously smaller than that for RFE-SVM. This further suggests that LFDA is able to find a transformation that can significantly reduce the dimensionality while preserving the rich statistical structure of the data—something that a feature-elimination strategy cannot achieve.

The quality (i.e., the accuracy or goodness of fit) of the GMM model learned from the training data is sensitive to the initialization of the parameter space, because the convergence of the EM algorithm employed in GMM training would depend on the initial seed selection of the parameter space. In this paper, as is commonly done, a K -means clustering is used for initializing the parameter vector [30]. To determine an appropriate value of K , we incrementally increase the number of mixture components. Following this, a metric such as BIC or AIC is employed to determine the optimal number of

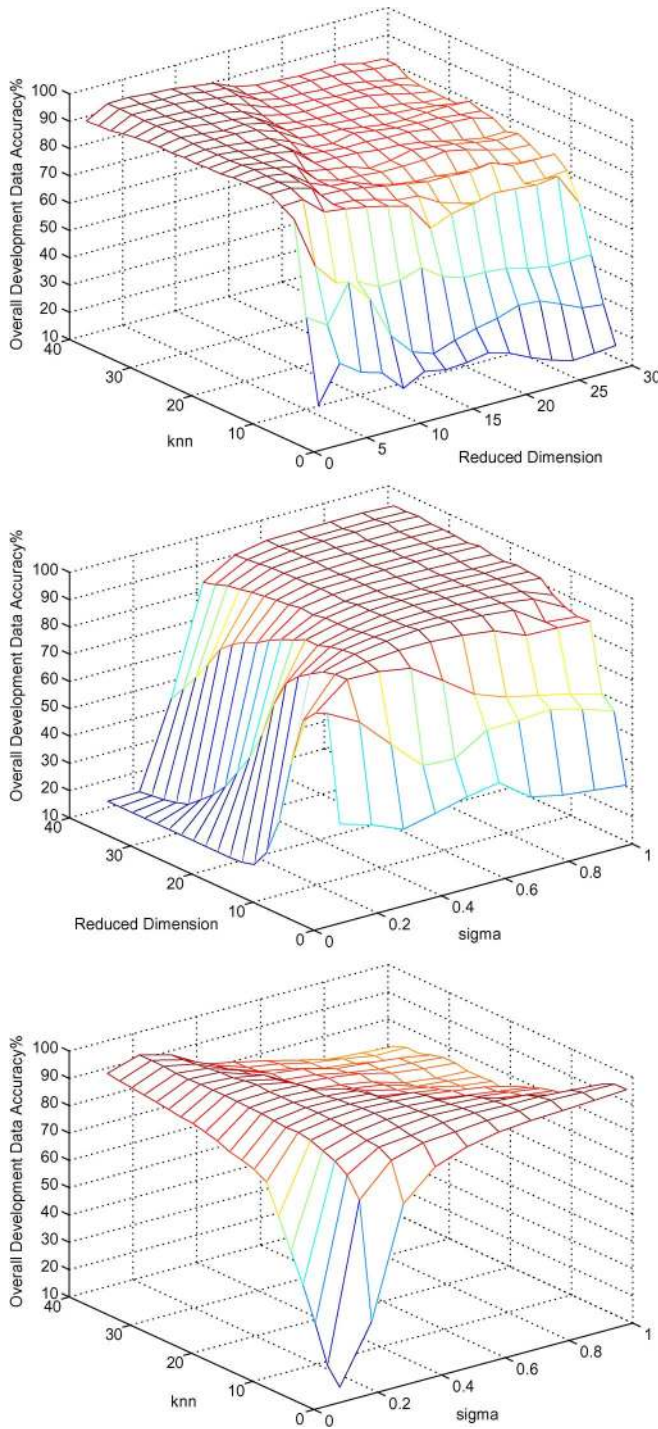


Fig. 8. LFDA-SVM for the University of Pavia data set: Overall development accuracy versus reduced dimension, k_{nn} , and σ .

components for the GMM. Table II shows the number of mixtures estimated using these two metrics for the Pavia Centre data set. The maximum number of components per class is set to five, and each column in the table corresponds to a unique class in the data set. From Table II, we observe that the number of components estimated using AIC is consistently greater than that estimated using BIC. We found the overall accuracy with AIC to be almost the same as that with BIC. Hence, we conclude that, for such HSI-classification tasks, BIC is better

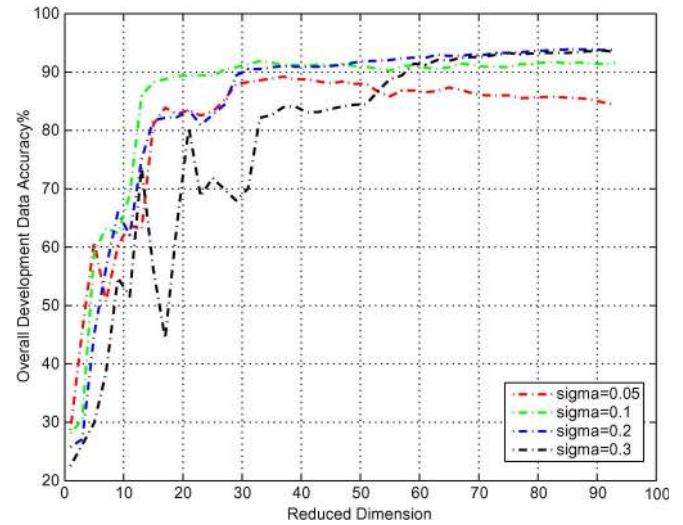


Fig. 9. RFE-SVM for the University of Pavia data set: Overall development accuracy versus reduced dimension and σ .

TABLE I
OPTIMAL PARAMETERS FOR THE VARIOUS ALGORITHMS AFTER TUNING FOR EXPERIMENTAL HYPERSPECTRAL DATA

		Reduced Dimension	k_{nn}	σ
Indian Pines	LFDA-GMM	11	7	—
	LFDA-SVM	11	7	0.4
	RFE-SVM	80	—	0.3
University of Pavia	LFDA-GMM	10	7	—
	LFDA-SVM	10	7	0.5
	RFE-SVM	60	—	0.2
Pavia Centre	LFDA-GMM	10	7	—
	LFDA-SVM	10	7	0.5
	RFE-SVM	60	—	0.2

TABLE II
NUMBER OF COMPONENTS ESTIMATED USING BIC AND AIC FOR THE PAVIA CENTRE DATA SET, WITH TA VARYING FROM 100% TO 50%

TAs	BIC									
100%	1	1	1	2	1	2	2	2	2	(14)
90%	1	1	2	2	1	1	2	2	1	(13)
80%	1	1	1	2	1	2	2	2	4	(16)
70%	1	1	1	2	3	2	2	2	2	(16)
60%	1	1	2	2	3	1	3	2	2	(17)
50%	1	1	3	4	1	1	3	4	2	(20)
TAs	AIC									
100%	1	1	3	4	1	3	4	3	2	(22)
90%	1	1	3	2	3	3	4	4	2	(23)
80%	1	1	4	3	1	3	4	4	3	(24)
70%	1	1	4	2	3	4	5	2	3	(26)
60%	1	3	4	3	3	4	5	2	3	(28)
50%	1	1	4	5	4	4	4	4	3	(30)

suited to determine the number of mixtures in the GMM model; we thus employ BIC in all experiments that follow. It is also important to point out that a full covariance matrix is used and learned for each Gaussian mode in the GMM models. To further reduce complexity, a diagonal covariance matrix can be used, but that is not necessary in this paper because of the significant dimensionality reduction already attained by LFDA.

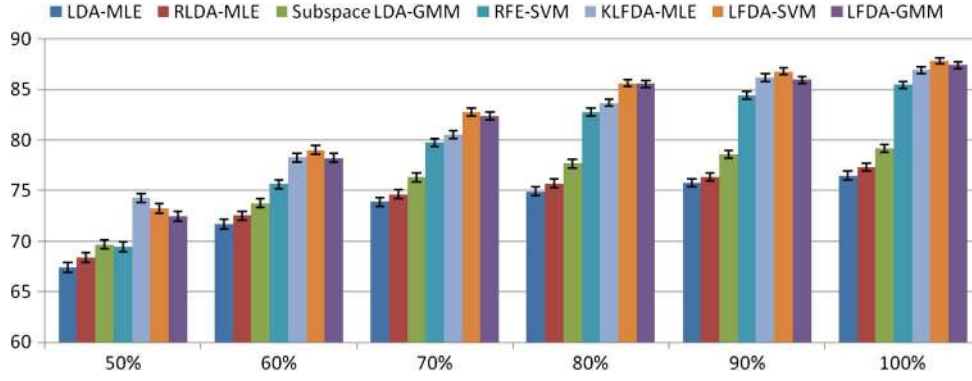


Fig. 10. Indian Pines: Overall accuracy versus pixel-mixing abundance, both expressed in percentage, for several different classification methods.

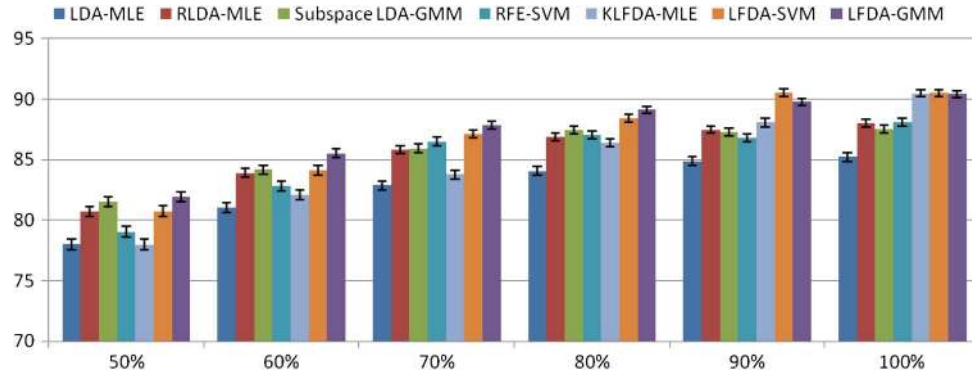


Fig. 11. University of Pavia: Overall accuracy versus pixel-mixing abundance, both expressed in percentage, for several different classification methods.

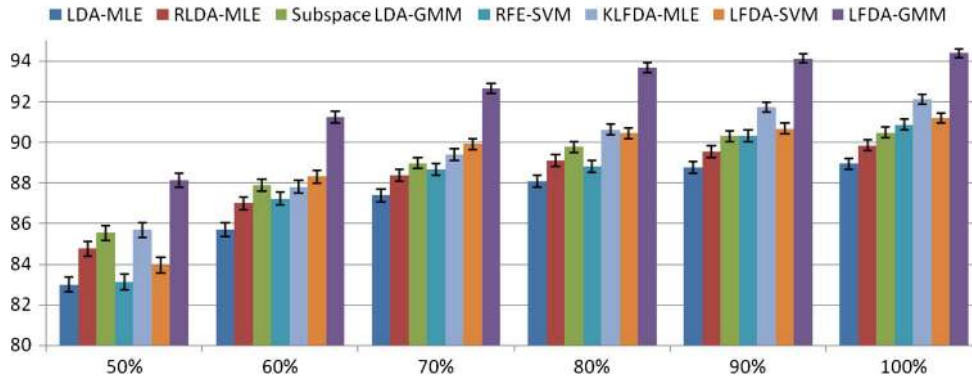


Fig. 12. Pavia Centre: Overall accuracy versus pixel-mixing abundance, both expressed in percentage, for several different classification methods.

C. Comparison Against Current State-of-the-Art Parametric Classification Techniques

To demonstrate the benefits of LFDA as a powerful dimensionality-reduction tool for HSI classification, we compare its performance using GMM and SVM classifiers with that of other traditional dimensionality-reduction methods, including LDA, regularized LDA (RLDA), and subspace LDA. Data distribution in LDA-projected subspaces tends to be Gaussian, which is hence followed by the conventional quadratic Gaussian MLE classifier. RLDA [31] alleviates the problem of an unstable inverse commonly encountered in traditional LDA under small-sample-size and high-dimensionality situations. The resulting algorithms are thus referred to as LDA-MLE and RLDA-MLE in this paper.

In subspace LDA [7], an intermediate PCA transformation is employed to discard the null space of the within-class scatter matrix, following which LDA is applied. This is an alternate mechanism to alleviate ill-conditioning in LDA formulations. Additionally, subspace LDA is an interesting algorithm to which to compare LFDA, since LFDA essentially combines LPP (an unsupervised linear manifold learning) and LDA (a supervised dimensionality reduction) to exploit the benefits of LPP within the LDA setup. Subspace LDA followed by GMM (subspace LDA-GMM) is hence another algorithm we study to highlight the benefits of LFDA-GMM.

By design, LDA, RLDA, and subspace LDA result in a $(c - 1)$ -dimensional feature subspace after the dimensionality-reduction projection. The extent of dimensionality reduction after LFDA, as discussed previously, is determined by studying

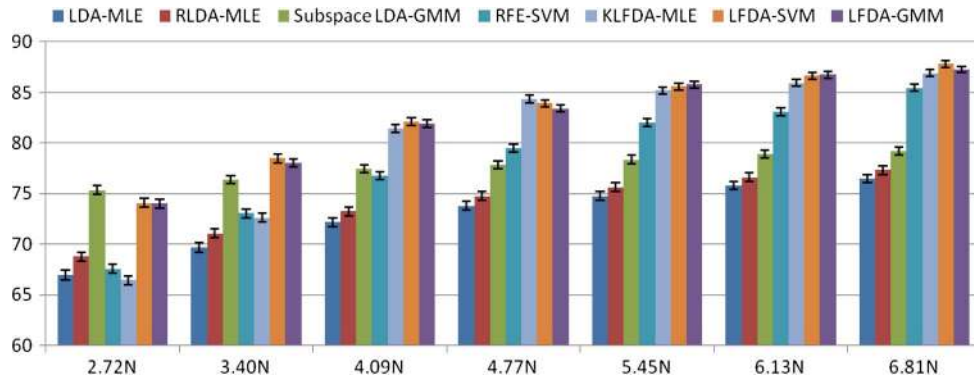


Fig. 13. Indian Pines: Overall accuracy (expressed in percentage) versus the training-data set size.

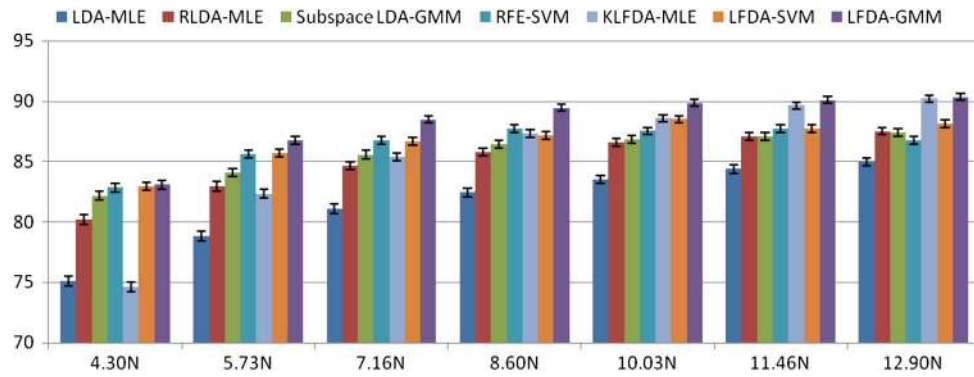


Fig. 14. University of Pavia: Overall accuracy (expressed in percentage) versus the training-data set size.

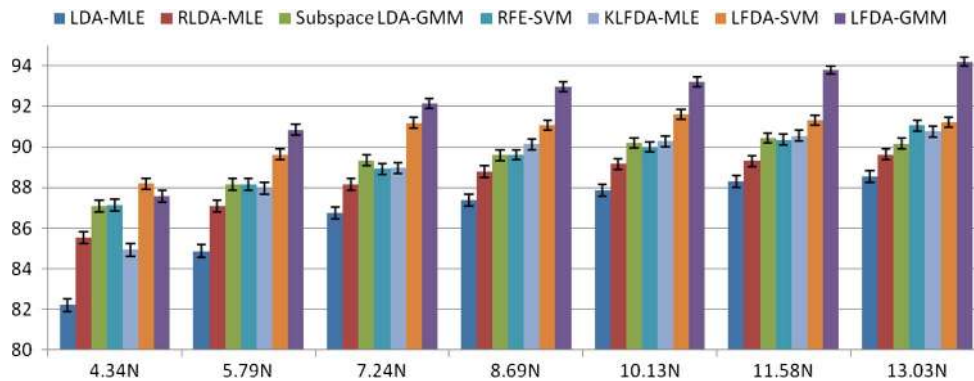


Fig. 15. Pavia Centre: Overall accuracy (expressed in percentage) versus the training-data set size.

the performance as a function of different subspace dimensions and choosing a value that maximizes the development-data accuracy.

Another algorithm that we will employ as a baseline is RFE-SVM, which is an established and powerful dimensionality-reduction and classification approach. Finally, we recently studied a kernel extension of the LFDA algorithm—kernel LFDA with MLE classifier (KLFDA-MLE) for HSI classification [32]. We found KLFDA-MLE to outperform powerful approaches such as RFE-SVM in [28]. In this paper, we use this as an additional baseline algorithm to which to compare the performance of LFDA-SVM and LFDA-GMM.

A comparison of the proposed methods (LFDA-GMM and LFDA-SVM) with these state-of-the-art parametric classification techniques is shown in Figs. 10–18. To simulate real-

life challenging operating conditions, we provide results for a wide range of pixel-mixing conditions. In many situations, the spatial resolution may not be fine enough to resolve the object of interest on ground, and inadvertent mixing between multiple classes may occur [33]. In this experiment, we use the data sets described previously and linearly mix signatures from background classes with the signature of the class being classified. We report results over a range of percentage target-abundance (TA) values. For example, a TA of 70% indicates that 30% of background signatures were mixed linearly with 70% of the target class. An abundance of 100% implies that pure pixels are employed without any mixing. Here, “target class” simply refers to the true class of the pixel currently being classified/tested. The background signatures/pixels used for mixing the target class are gathered (with uniform weights) from across all the other classes.

Figs. 10–12 show the overall accuracy as a function of TA for all four HSI data sets. The methods proposed in this paper—both LFDA-GMM and LFDA-SVM—are indeed very powerful classification approaches, outperforming traditional state-of-the-art approaches significantly, even under adverse TA conditions.

We also conducted an experiment wherein we varied the amount of training data and studied the sensitivity of the proposed methods relative to conventional methods over a range of training-data set sizes [34]. In practical situations, the number of training samples available is often insufficient to estimate models for each class. We report the overall accuracy of these classification systems as a function of the relative training-sample size in Figs. 13–15. This relative training-sample size (on the horizontal axis) is expressed relative to the spectral dimensionality of the HSI data. Hence, an abundance of $6N$ implies that the amount of training data used is six times the dimensionality of the feature space (here, N denotes the dimensionality of the original feature space, or the number of spectral bands for this classification task). To avoid any spatial biases, we randomly chose a subset of training samples for each sample-size value and repeated the experiment 20 times, reporting the average classification accuracy. Note that, with decreasing training-data set size, the overall accuracy for all systems decreases, as expected. However, the overall accuracy of LFDA-GMM and LFDA-SVM is always higher than that of the other baseline methods. Even at a very low training-data set size (e.g., $3.4N$ in Fig. 10), the performance of the proposed methods is impressive.

We also report visual ground-cover classification maps for the Indian Pines and Pavia data sets, since they come with labeled ground truth for training and visual comparison. Figs. 16–18 show the thematic maps resulting from the classification of these hyperspectral scenes using LDA-MLE, LFDA-GMM, RFE-SVM, and LFDA-SVM. We produced ground-cover maps of the entire HSI scene for these images (including unlabeled pixels). However, to facilitate easy comparison between methods, only areas for which we have ground truth are shown in these maps. Clearly, LFDA-GMM and LFDA-SVM consistently result in maps that are less noisy and more accurate compared to traditional state-of-the-art methods. For example, see the circled area of Soybeans-min till in Fig. 18 or the circled region of Asphalt in Fig. 16. It can be observed that LFDA-GMM and LFDA-SVM result in maps having a significantly reduced misclassification noise.

V. CONCLUSION

We have presented a locality-preserving discriminant analysis for hyperspectral dimensionality reduction. This approach was previously proposed and tested with simple binary-classification data sets having a much smaller dimensionality. Additionally, previous work has not studied LFDA as a dimensionality-reduction tool for complex nonlinear classifiers such as GMMs and nonlinear kernel-based SVMs. In this paper, we argued that LFDA serves as a very effective dimensionality-reduction tool for such classifiers, for very high-dimensional and challenging classification tasks such as HSI ground-cover

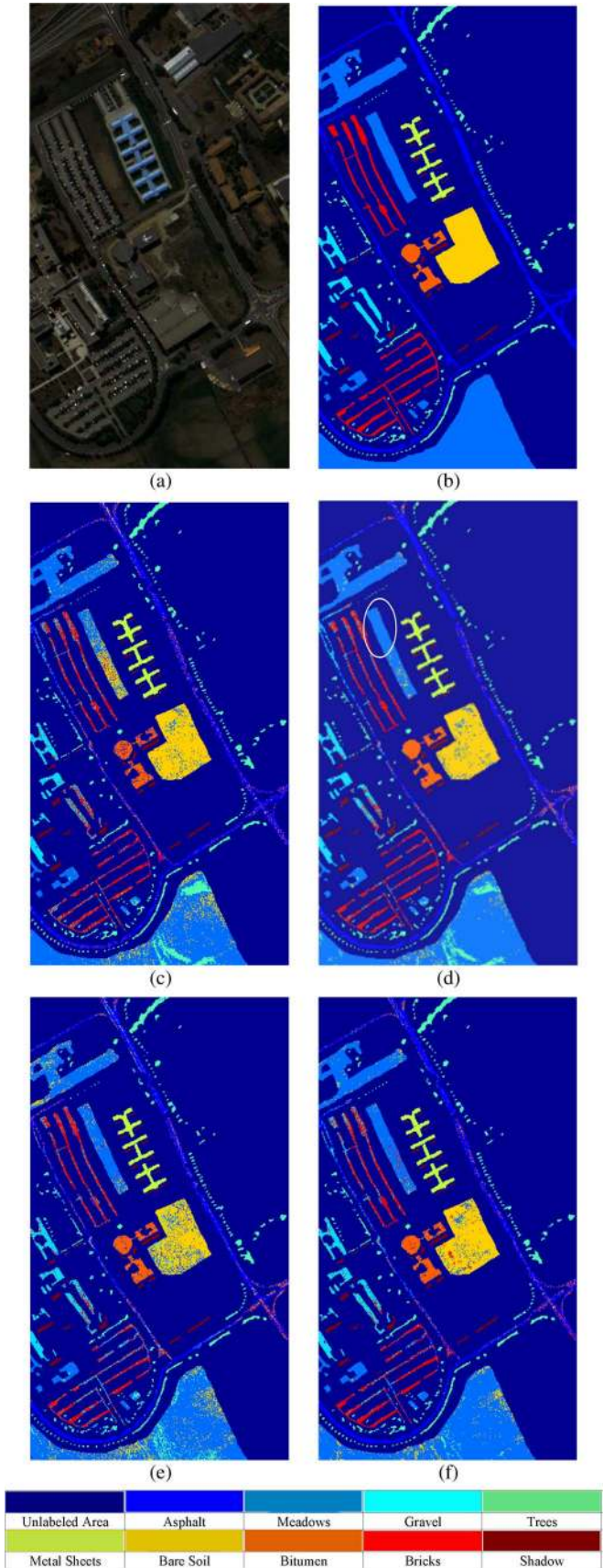


Fig. 16. (a) False color image of the University of Pavia (using bands 60, 30, and 2), (b) ground truth of the labeled area with nine classes, and (c) thematic maps resulting from classification of “LDA-MLE,” (d) “LFDA-GMM,” (e) “RFE-SVM,” and (f) “LFDA-SVM.”

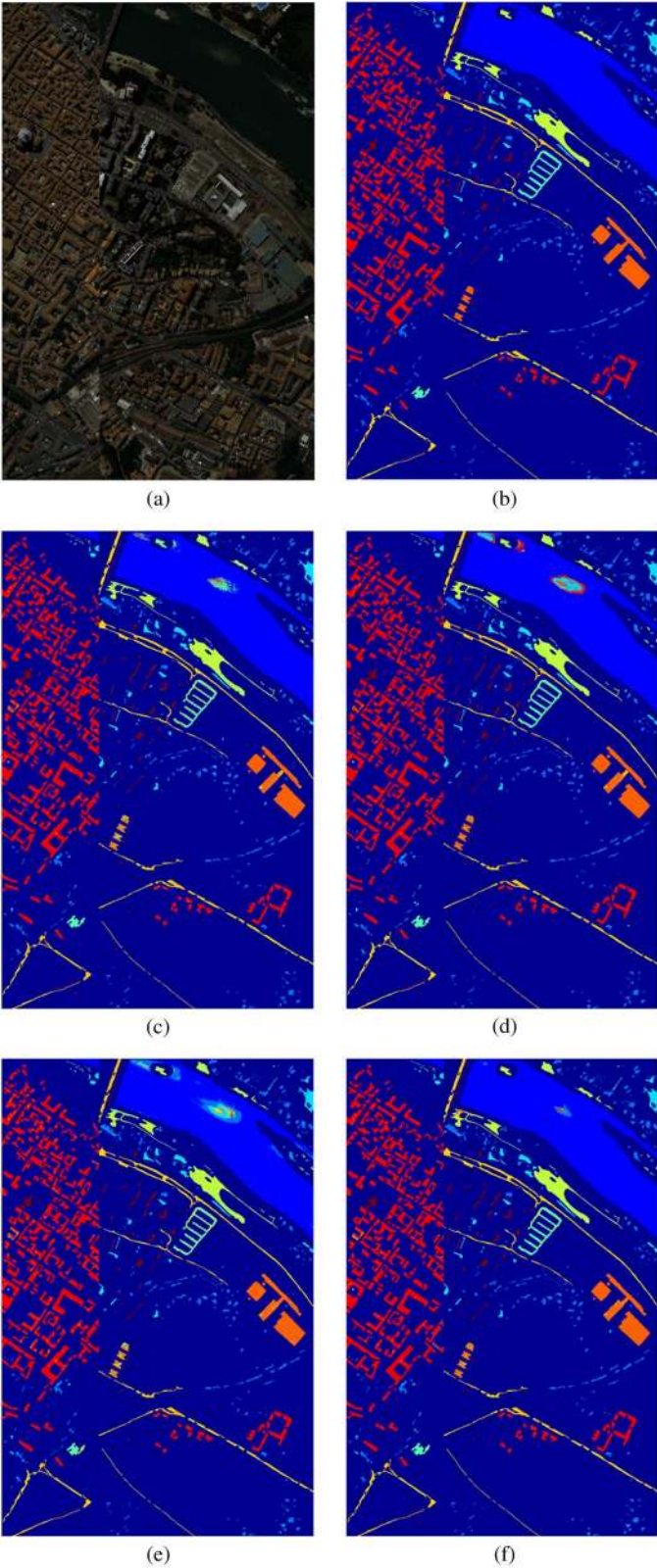


Fig. 17. (a) False color image of the Pavia Centre, (b) ground truth of the labeled area with nine classes, and (c) thematic maps resulting from classification of “LDA-MLE,” (d) “LFDA-GMM,” (e) “RFE-SVM,” and (f) “LFDA-SVM.”

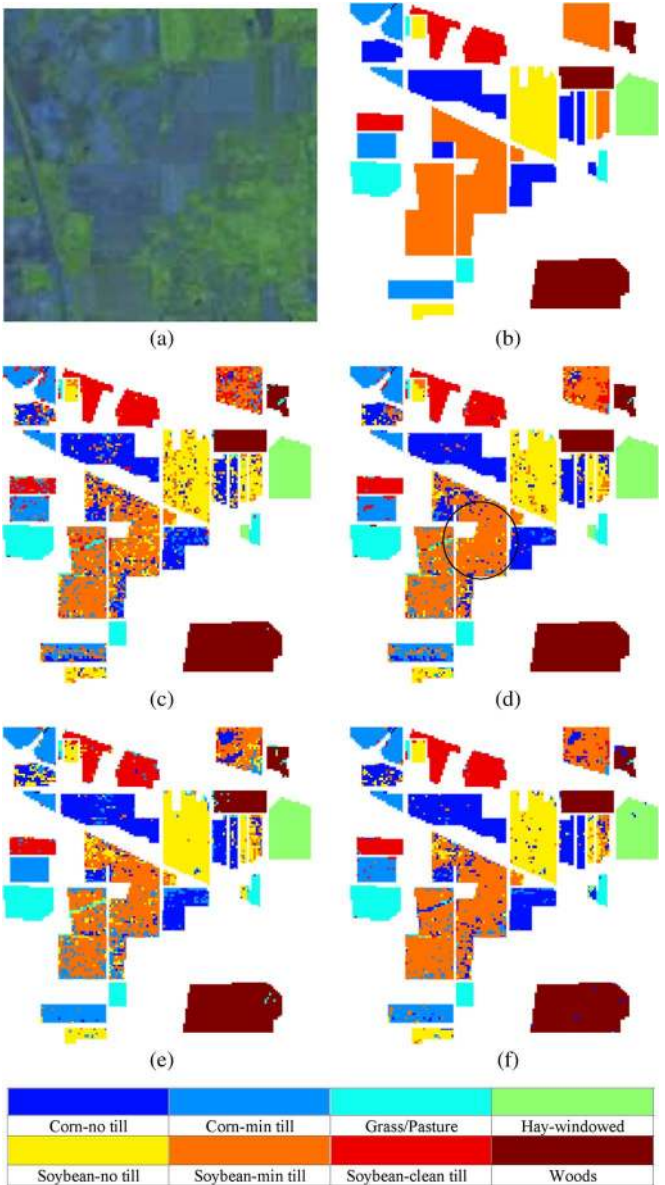


Fig. 18. (a) False color image of Indian Pines (using bands 80, 30 and 20), (b) ground truth of the labeled area with eight classes, and (c) thematic maps resulting from classification of “LDA-MLE,” (d) “LFDA-GMM,” (e) “RFE-SVM,” and (f) “LFDA-SVM.”

classification. Furthermore, since LFDA is a linear projection, its complexity is not significantly higher than that of traditional LDA. While there is some additional overhead in terms of memory when estimating the affinity matrix, LFDA does not add any additional complexity to the classification backend (e.g., GMM or SVM).

We provided experimental results with HSI data demonstrating system performance over the parameter space for the proposed approach, a procedure that can be employed for optimizing the system for any classification task at hand. We also studied the performance of the proposed system under two real-life adverse operating scenarios commonly encountered in remote-sensing analysis—i.e., under very limited training data and under severe pixel mixing—and we showed that the proposed approach significantly outperforms conventional techniques under both conditions.

We note that, although we have not included results of GMM classification without any dimensionality reduction in this paper, we performed such classification as well and found that the overall classification performance of basic GMMs with HSI data is dismally low—for example, with the University of Pavia data set, when all of the training data were used, the accuracy was a low 53%, while when the amount of training data was dropped to $4.3N$, the accuracy dropped to 20%, indicating that GMM training effectively broke down due to an immensely high-dimensional parameter space. These observations further corroborate our arguments as to the benefits of LFDA as a dimensionality-reduction tool for classifiers such as GMMs.

We note also that, in this paper, we have considered classification tasks involving up to nine classes in a relatively well-structured environment. However, this data complexity can indeed be scaled up to complicated scenes involving many more classes—in fact, we expect the LFDA-GMM/SVM approach to be even more effective at capturing subtle statistical differences for classification in such complex environments.

REFERENCES

- [1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
- [2] C. Lee and D. A. Landgrebe, "Analyzing high-dimensional multispectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 31, no. 4, pp. 792–800, Jul. 1993.
- [3] S. Prasad and L. M. Bruce, "Decision fusion with confidence-based weight assignment for hyperspectral target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1448–1456, May 2008.
- [4] L. Zhang, Y. Zhong, B. Huang, J. Gong, and P. Li, "Dimensionality reduction based on clonal selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4172–4186, Dec. 2007.
- [5] M. D. Farrell and R. M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 192–195, Apr. 2005.
- [6] X. Jia and J. A. Richards, "Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 1, pp. 538–542, Jan. 1999.
- [7] S. Prasad and L. M. Bruce, "Limitations of principal component analysis for hyperspectral target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 625–629, Oct. 2008.
- [8] S. Di Zenzo, R. Bernstein, S. D. Degloria, and H. C. Kolsky, "Gaussian maximum likelihood and contextual classification algorithms for multitrop classification," *IEEE Trans. Geosci. Remote Sens.*, vol. GRS-25, no. 6, pp. 805–814, Nov. 1987.
- [9] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
- [10] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, no. 5, pp. 1027–1061, May 2007.
- [11] S. Tadjudin and D. A. Landgrebe, "Robust parameter estimation for mixture model," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 1, pp. 439–445, Jan. 2000.
- [12] M. M. Dundar and D. A. Landgrebe, "A model-based mixture-supervised classification approach in hyperspectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 12, pp. 2692–2699, Dec. 2002.
- [13] A. Berge and A. H. S. Solberg, "Structured Gaussian components for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3386–3396, Nov. 2006.
- [14] C. Cortes and V. N. Vapnik, "Support vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
- [15] 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.
- [16] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.
- [17] D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski, "Learning relevant image features with multiple-kernel classifications," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3780–3791, Oct. 2010.
- [18] X. He and P. Niyogo, "Locality preserving projections," in *Advances in Neural Information Processing System*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.
- [19] M. N. Do, "Fast approximation of Kullback–Leibler distance for dependence trees and hidden Markov models," *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 115–118, Apr. 2003.
- [20] G. Shaw and D. Manolakis, "Signal processing for hyperspectral image exploitation," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 12–16, Jan. 2002.
- [21] N. Vlassis and A. Likas, "A greedy EM algorithm for Gaussian mixture learning," *Neural Process. Lett.*, vol. 15, no. 1, pp. 77–87, Feb. 2002.
- [22] A. P. Benavent, F. E. Ruiz, and J. M. S. Martínez, "EBEM: An entropy-based EM algorithm for Gaussian mixture models," in *Proc. IEEE Int. Conf. Pattern Recognit.*, New York, Jun. 2006, vol. 2, pp. 451–455.
- [23] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.
- [24] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.
- [25] R. J. Steele and A. E. Raftery, "Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models," Dept. Stat., Univ. Washington, Washington, DC, Tech. Rep. 559, Sep. 2009.
- [26] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
- [27] I. Guyon, J. Weston, S. Barhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, Jan. 2002.
- [28] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 674–677, Oct. 2007.
- [29] P. Gamba, "A collection of data for urban area characterization," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Anchorage, AK, Sep. 2004, vol. 1, pp. 69–72.
- [30] T. Su and J. G. Dy, "In search of deterministic methods for initializing k -means and Gaussian mixture clustering," *Intell. Data Anal.*, vol. 11, no. 4, pp. 319–338, Sep. 2007.
- [31] 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.
- [32] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving discriminant analysis in kernel-induced spaces for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 894–898, Sep. 2011.
- [33] N. Dobigeon, S. Moussaoui, M. Coulon, J. Y. Tourneret, and A. O. Hero, "Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4355–4368, Nov. 2009.
- [34] F. A. Mianji and Y. Zhang, "Robust hyperspectral classification using relevance vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2100–2112, Jun. 2011.



Wei Li (S'11) received the B.E. degree in communications engineering from Xidian University, Xi'an, China, in 2007 and the M.S. degree in electrical engineering from Sun Yat-Sen University, Guangzhou, China, in 2009. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, Mississippi State University, Starkville. His supervisor is Dr. James E. Fowler, and his co-advisor is Dr. Saurabh Prasad.

Since 2009, he has been a Research Assistant with the Geosystems Research Institute, Mississippi State University. His research interests include hyperspectral image compression and statistical pattern recognition.



Saurabh Prasad (S'05–M'09) received the B.S. degree in electrical engineering from Jamia Millia Islamia, New Delhi, India, in 2003, the M.S. degree in electrical engineering from Old Dominion University, Norfolk, VA, in 2005, and the Ph.D. degree in electrical engineering from Mississippi State University, Starkville, in 2008.

He is currently an Assistant Research Professor with the Geosystems Research Institute (GRI) and an Adjunct Assistant Professor with the Department of Electrical and Computer Engineering, Mississippi State University. He has been the Lead Author of several successful grant proposals to agencies such as the National Geospatial-Intelligence Agency, the National Aeronautics and Space Administration, and the Department of Homeland Security, and serves as the Principal Investigator/Technical Lead for these projects at Mississippi State University. His research interests include statistical pattern recognition, adaptive signal processing and kernel methods for medical imaging, and optical and SAR remote sensing. In particular, his current research work involves the use of information fusion techniques for designing robust statistical pattern-classification algorithms for hyperspectral remote-sensing systems operating under low-signal-to-noise-ratio, mixed-pixel, and small-training-sample-size conditions.

Dr. Prasad is an active Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and the *Elsevier Pattern Recognition Letters*. He was awarded the GRI's Graduate Research Assistant of the Year Award in May 2007 and the Office-of-Research Outstanding Graduate Student Research Award in April 2008 at Mississippi State University. In July 2008, he received the Best Student Paper Award at IGARSS'2008 held in Boston, MA. In October 2010, he received the State Pride Faculty Award at Mississippi State University for his academic and research contributions. He was the Lead Editor of the book entitled *Optical Remote Sensing: Advances in Signal Processing and Exploitation Techniques*, published in March 2011.



James E. Fowler (S'91–M'96–SM'02) received the B.S. degree in computer and information science engineering and the M.S. and Ph.D. degrees in electrical engineering from The Ohio State University, Columbus, in 1990, 1992, and 1996, respectively.

In 1995, he was an Intern Researcher at AT&T Labs, Holmdel, NJ, and in 1997, he held a National Science Foundation-sponsored postdoctoral assignment at the Université de Nice-Sophia Antipolis, Nice, France. In 2004, he was a Visiting Professor with the Département Traitement du Signal et des

Images, École Nationale Supérieure des Télécommunications, Paris, France. He is currently a Billie J. Ball Professor and the Graduate Program Director of the Department of Electrical and Computer Engineering, Mississippi State University, Starkville. He is also a Researcher with the Geosystems Research Institute, Mississippi State.

Dr. Fowler is an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and the *EURASIP Journal on Image and Video Processing*. He formerly served as an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA and the IEEE SIGNAL PROCESSING LETTERS. He is the Vice-Chair of the Image, Video, and Multidimensional Signal Processing Technical Committee of the IEEE Signal Processing Society as well as the Publicity Chair of the Program Committee for the Data Compression Conference.



Lori Mann Bruce (S'90–M'96–SM'01) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from The University of Alabama, Huntsville, and the Georgia Institute of Technology, Atlanta.

She is the Associate Dean for Research and Graduate Studies in the Bagley College of Engineering, Mississippi State University, Starkville. She has been a Faculty Member for 14 years, during which she has taught approximately 40 engineering courses at the undergraduate and graduate levels. Her research in image processing and remote sensing has been funded by the Department of Homeland Security, the Department of Energy, the Department of Transportation, the National Aeronautics and Space Administration, the National Geospatial-Intelligence Agency, the National Science Foundation, the United States Geological Survey, and industry, resulting in over 100 peer-reviewed publications and the matriculation of more than 75 graduate students (25 as major professor and more than 50 as thesis/dissertation committee member).

Dr. Bruce is an active member of the IEEE Geoscience and Remote Sensing Society, and she is a member of the Phi Kappa Phi, Eta Kappa Nu, and Tau Beta Pi honor societies, and prior to becoming a faculty member, she held the prestigious title of National Science Foundation Research Fellow.