Classification of Hyperspectral Images With Regularized Linear Discriminant Analysis

Tatyana V. Bandos, Lorenzo Bruzzone, *Senior Member, IEEE*, and Gustavo Camps-Valls, *Senior Member, IEEE*

***Abstract*—This paper analyzes the classification of hyperspectral remote sensing images with linear discriminant analysis (LDA) in the presence of a small ratio between the number of training samples and the number of spectral features. In these particular ill-posed problems, a reliable LDA requires one to introduce regularization for problem solving. Nonetheless, in such a challenging scenario, the resulting regularized LDA (RLDA) is highly sensitive to the tuning of the regularization parameter. In this context, we introduce in the remote sensing community an efficient version of the RLDA recently presented by Ye *et al.* to cope with critical ill-posed problems. In addition, several LDA-based classifiers (i.e., penalized LDA, orthogonal LDA, and uncorrelated LDA) are compared theoretically and experimentally with the standard LDA and the RLDA. Method differences are highlighted through toy examples and are exhaustively tested on several ill-posed problems related to the classification of hyperspectral remote sensing images. Experimental results confirm the effectiveness of the presented RLDA technique and point out the main properties of other analyzed LDA techniques in critical ill-posed hyperspectral image classification problems.**

***Index Terms*—Hyperspectral images, ill-posed problem, image classification, linear discriminant analysis (LDA), regularization.**

I. Introduction

DURING the last decade, many supervised methods have been developed for classification of hyperspectral data. The problem is complex due to the high number of spectral channels (which results in a high-dimensional feature space), the relatively small number of labeled data, the presence of different sources of noise, and the nonstationary behavior of land-cover spectral signatures [2]–[4]. Many different techniques have been proposed in the literature for hyperspectral image classification, ranging from regularized maximum-likelihood methods to classifiers based on simple models and a very small number of parameters. In recent years, the framework of *kernel methods* in general, and support vector machines in particular, has offered state-of-the-art performances in ill-posed classification problems associated with hyperspectral images. The effectiveness of these approaches is mainly due to the intrinsic regularization and robustness that they exhibit in high-dimensional problems [4]–[8]. However, despite their good classification performance, the effectiveness of kernel methods and support vector machines depends on the selection of some critical free parameters, which define the specific classification model. The phase of free-parameter selection is also called *model selection*. Thus, the use of these methods should be carried out with the supervision of experts who can adequately conduct the model selection learning process. Despite the fact that several criteria are available for selecting the support-vector-machine (SVM) free parameters (see [9] and the references therein), the common approach consists of the exhaustive search through cross-validation strategies (e.g., v-fold or leave-one-out cross-validation). These methodologies are computationally demanding, particularly when several (sensitive) free parameters are to be tuned. Nonetheless, in many operational domains, where the classification task may be driven from nonexpert users, it is useful to exploit simple classifiers, which do not require the tuning of many different parameters but still allow one to achieve competitive accuracies.

In this context, a possible solution is to adopt adequately defined simple (potentially linear) classifiers, which can merge easy implementation and clear physical interpretation with high accuracy. *Subspace methods* are a particular class of algorithms focused on finding projections of the original hyperdimensional space to a lower dimensional space where the class separation is maximized. Linear discriminant analysis (LDA) is an effective subspace technique as it optimizes the Fisher score [10]. In addition, it does not require the tuning of free parameters. These good capabilities have resulted in its extensive use and practical exploitation in remote sensing applications mainly focused on image classification and feature reduction. In [11], a geobotanical investigation based on linear discriminant was conducted, in which the good explanatory capabilities of LDA were exploited for profile analysis. In [12], the classical LDA was used for conifer species recognition with hyperspectral data. Lately, in [13], LDA has been used for classification of tropical rainforest tree species using hyperspectral data at different scales. In [14], the canonical LDA has been used for identifying land-cover units in ecology.

Despite its good performance in many applications, classical-LDA-based methods cannot work in ill-posed problems, when the number of features is higher than the number of training

Manuscript received March 26, 2008; revised July 22, 2008. Current version published February 19, 2009. This paper was supported in part by the Italian and Spanish Ministries for Education and Science under Grant "Integrated Action Programme Italy–Spain" and under Projects DATASAT/ESP2005-07724-C05-03 and CONSOLIDER/CSD2007-00018.

T. V. Bandos was with the Departament d'Enginyeria Electrònica, Universitat de València, 46100 València, Spain, and also with the Department of Information and Communication Technology, University of Trento, 38050 Trento, Italy. She is now with the Institute of Pure and Applied Mathematics, Polytechnic University of Valencia, 46022 Valencia, Spain.

L. Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, 38050 Trento, Italy (e-mail: lorenzo.bruzzone@ing.unitn.it).

G. Camps-Valls is with the Departament d'Enginyeria Electrònica, Escola Tècnica Superior d'Enginyeria, Universitat de València, 46100 València, Spain (e-mail: gustavo.camps@uv.es).

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.2008.2005729

samples. This is due to the fact that the standard LDA algorithm is based on the computation of the following three scatter matrices: the *within-class*, the *between-class*, and the *common scatter* matrices, defined via the sample covariance matrix. In this framework, obtaining the solution requires the common scatter matrix to be inverted, and thus, this matrix must be nonsingular. Several strategies can be followed to fulfill this condition. The simplest one consists in applying a preliminary feature selection/extraction step and then using LDA on the new (low-dimensional) feature space. In this way, the ratio between the number of training samples and the number of features can increase, thus allowing one to obtain good estimations of the covariance matrices. This is commonly obtained by applying either the filter or the wrapper method [15], [16]. For instance, a common procedure consists in retaining few (informative) features extracted through principal component analysis (PCA) and then applying LDA to them [17], [18]. However, this (*ad hoc*) preliminary stage poses the important problem of defining a criterion for selecting the spectral channels (or the extracted components) to be used for training the LDA. An attractive alternative is to modify the LDA method so that it can deal with ill-posed problems directly, without any preprocessing step of the input data. This can be accomplished by introducing *regularization* in order to mitigate the effects of the relatively small number of training samples.

Since the early presentation of the Fisher LDA [10], many techniques have been proposed to deal with ill-posed problems. The following two excellent algorithms have been proposed in the literature for the regularization of the LDA equations: 1) the penalized LDA (PLDA) [19] and 2) the regularized LDA (RLDA) [20]. These methods (which have been recently used in remote sensing applications [21]) are focused on the class discriminant score rather than on the feature extraction [22]. However, these algorithms show numerical instability when the regularization parameter tends to zero, involving again a singularity problem. Lately, other forms of regularization have been explored based on the concepts of sharing covariance structure [23] and through the efficient estimation of inverse covariance matrices by sequential Cholesky factorization [24]. Other techniques address ill-posed problems constructing the new (projected) features in such a way that they have to be uncorrelated or orthogonal in the projection subspace. This results in the definition of the uncorrelated LDA (ULDA) [25] and the orthogonal LDA (OLDA) [26], respectively. Nevertheless, model selection turns to be critical again, since the regularization parameter is difficult to be selected, and many values need to be tested.

In the aforementioned framework, this paper presents two main contributions. First, we conduct an exhaustive comparison of several LDA-based methods for hyperspectral remote sensing image classification in ill-posed problems. To this purpose, after reporting results on simulated toy examples, we analyze the effectiveness of five different LDA-based classifiers (LDA, PLDA, RLDA, OLDA, and ULDA), along with soft-margin linear and nonlinear SVMs on five standard hyperspectral image data sets. Second, we introduce an efficient version of the RLDA in which, instead of regularizing the scatter matrix, its nonzero eigenvalues are regularized (which has been proved to be equivalent [1]). This strategy alleviates the critical problem of traditional RLDA related to the computation of the inverse of the (regularized) total scatter matrix when the regularization parameter tends to zero. Hence, it permits an exhaustive exploration of the values of the regularization parameter and, thus, the selection of the optimal one for the considered problem. This allows one to overcome the critical drawback of the model selection in standard RLDA.

This paper extends our previous work [27] by analyzing in more detail the theoretical relations among methods and by showing more results in different real scenarios to assess the capabilities of the different techniques. The rest of this paper is organized as follows. Section II reviews the basic notation for the classical LDA method used in the paper. Section III presents the regularization framework for LDA-based methods. Specifically, the standard formulations for the PLDA and RLDA are introduced. In Section IV, the proposed version of the RLDA is analyzed in detail and compared theoretically with the OLDA, the ULDA, and the standard RLDA method. Section V presents the experimental results obtained in several ill-posed situations on both illustrative toy examples and several hyperspectral image classification problems. Finally, Section VI draws the conclusion of this work and gives some directions for further investigation.

II. Linear Discriminant Analysis (LDA)

This section presents an overview of the classical LDA and devotes special attention to highlight its limitations due to the nonsingularity assumption. In supervised remote sensing image classification, we are given a set of n labeled samples, $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^m$ represents the m-dimensional feature vector for the ith pixel with label $y_i \in \Omega$. Here, m denotes the spectral bands (or channels), and Ω defines the universe of all possible labeled classes in the image. The notation used throughout this paper is summarized in Table I.

The standard LDA classifier allows us to find a *linear* transformation matrix $\mathbf{G}$ that reduces an original m-dimensional feature vector $\mathbf{x}$ to an l-dimensional vector, $\mathbf{a} = \mathbf{G}^\top \mathbf{x} \in \mathbb{R}^l$, where $l < m$. This low-dimensional feature space is selected to fulfill a given maximization criterion of separability among class distributions [28]. The widely used Fisher criterion [10] is based on maximizing the distance among the means of the classes and, at the same time, minimizing their intraclass variances on the basis of the following function, $J(\mathbf{w}) = (\mu_2 - \mu_1)^2/(\sigma_2^2 + \sigma_1^2)$. See Fig. 1 for an illustrative example in a two-class problem.

Since the decision function is $y = \mathbf{w}^\top \mathbf{x}$, and the means and variances can be trivially defined, one can easily demonstrate that maximizing the Fisher score is equivalent to maximizing the following Rayleigh coefficient with respect to the decision function weight vector $\mathbf{w}$:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \{J(\mathbf{w})\} = \arg\max_{\mathbf{w}} \left\{ \frac{\mathbf{w}^\top \mathbf{S}_b \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}} \right\} \quad (1)$$

where $\mathbf{S}_b = (1/n) \sum_{k=1}^{K} n_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top$ is the between-class variance, $\mathbf{S}_w = (1/n) \sum_{k=1}^{K} \sum_{i \in \mathbf{I}_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top$

TABLE I
NOTATION USED IN THIS PAPER

Notation	Description	Notation	Description
n	number of training samples	m	dimension of the input space (bands)
l	number of features in the mapped space	K	number of classes
n_i	number of training samples from class i	$\mathbf{X}_i \in \mathbb{R}^{m\times n_i}$	data matrix of elements in class i
$\mathbf{x} \in \mathbb{R}^m$	vector of m features (bands)	$\mathbf{X} \in \mathbb{R}^{m\times n}$	data matrix of all elements
$\mathbf{G} \in \mathbb{R}^{m\times l}$	transformation matrix	$\mathbf{a} \in \mathbb{R}^l$	mapped vector of l features
$\boldsymbol{\mu}_i$	mean vector of class i	$\boldsymbol{\mu}$	mean vector of the data
$\boldsymbol{\Sigma}_i$	covariance matrix for the ith class	$\boldsymbol{\Sigma}$	pooled covariance matrix
$\mathbf{S}_b \in \mathbb{R}^{m\times m}$	between-class scatter matrix	$\mathbf{S}_w \in \mathbb{R}^{m\times m}$	within-class scatter matrix
$\mathbf{S} \in \mathbb{R}^{m\times m}$	total scatter matrix	λ, γ	regularization parameters
$\mathbf{I} \in \mathbb{R}^{m\times m}$	identity matrix in the feature space	$\mathbf{I}_r \in \mathbb{R}^{r\times r}$	identity matrix of size r
$\widehat{\ }$	estimate symbol	$\top$	transpose symbol

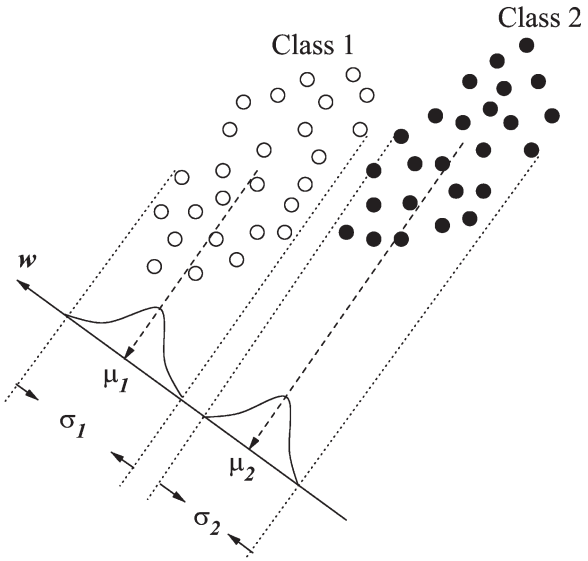

Fig. 1. Illustration of Fisher's discriminant for two classes. One searches for a direction $\mathbf{w}$ such that both the difference between the class means projected onto this direction (μ_1 and μ_2) is large and the variance around these means ($\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$) is small.

is the within-class variance, and $\boldsymbol{\mu}_k$ and $\mathbf{I}_k$ denote the sample mean and the index set for class k, respectively.

The maximization criterion in (1) can be rewritten as the following maximization problem:

$$G^* = \arg\max_G \left\{\text{trace}\left((\mathbf{G}^\top \mathbf{S}_w \mathbf{G})^{-1}\mathbf{G}^\top \mathbf{S}_b \mathbf{G}\right)\right\} \tag{2}$$

which can be demonstrated to be equivalent to [28]

$$G^* = arg\max_G \left\{\text{trace}\left((\mathbf{G}^\top \mathbf{S} \mathbf{G})^{-1}\mathbf{G}^\top \mathbf{S}_b \mathbf{G}\right)\right\} \tag{3}$$

where $\mathbf{S} = \mathbf{S}_b + \mathbf{S}_w$ is the total scatter matrix, which is the estimate of the common covariance matrix. Note that solving any of the aforementioned equivalent problems is only possible if $\mathbf{S}_w$ and $\mathbf{S}$ are nonsingular.

In this context, the scatter matrices can be redefined as

$$\mathbf{S} = \mathbf{H}\mathbf{H}^\top \quad \mathbf{S}_w = \mathbf{H}_w\mathbf{H}_w^\top \quad \mathbf{S}_b = \mathbf{H}_b\mathbf{H}_b^\top \tag{4}$$

where

$$\mathbf{H} = \frac{1}{\sqrt{n}}(\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^\top) \tag{5}$$

$$\mathbf{H}_w = \frac{1}{\sqrt{n}}\left[\mathbf{X}_1 - \boldsymbol{\mu}_1\mathbf{1}_1^\top, \ldots, \mathbf{X}_K - \boldsymbol{\mu}_K\mathbf{1}_K^\top\right] \tag{6}$$

$$\mathbf{H}_b = \frac{1}{\sqrt{n}}\left[\sqrt{n_1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \sqrt{n_K}(\boldsymbol{\mu}_K - \boldsymbol{\mu})\right] \tag{7}$$

with $\mathbf{1}$ being a column vector of n ones and $\mathbf{1}_i$ being a column vector of n_i ones.

It is worth noting that (2) and (3) are generalized eigendecomposition problems for the scatter matrix $\mathbf{S}_w^{-1}\mathbf{S}_b$ or $\mathbf{S}^{-1}\mathbf{S}_b$, respectively. As a consequence, there exists a maximum of $K-1$ eigenvectors with nonzero eigenvalues, as this is the upper bound on the rank of $\mathbf{S}_b$. Several problems arise from this theoretical fact. First, selecting a maximum of $K-1$ eigenvectors of matrix $\mathbf{S}_w^{-1}\mathbf{S}_b$ for projecting data and classification may not be always sufficient, because essential information can be lost. Second, $\mathbf{S}$ (or $\mathbf{S}_w$) is often singular in high-dimensional small-sized data sets, and thus, the solution cannot be obtained. As it will be analyzed in the following section, *regularization* permits one to alleviate this problem.

III. REGULARIZED LDA

This section is focused on the regularization of LDA equations. First, we introduce the general concept of regularization and how it is usually included in the LDA formulation, thus leading to the following two regularized algorithms: the PLDA and the RLDA. Second, we introduce in the remote sensing community a version of the RLDA for efficiently tuning the regularization parameter, which was originally proposed in [1] for general-purpose machine learning applications.

A. Regularization of LDA

The LDA approach makes use of a linear transformation to reduce the dimensionality of data in classification problems [22]. However, in hyperspectral remote sensing applications,

the ratio between the number of training samples and the features is small due to the fact that labeling is expensive and that the feature space has high dimensionality. One of the consequences is that the sample covariance matrices become singular, anisotropic, and usually have highly variable parameters when few labeled samples are used. On the one hand, the classical discriminant analysis can be used only if more than $m+1$ training samples are available. On the other hand, when the ratio between the number of samples n and the number of spectral features m is too small, the problem becomes ill posed, and the Hughes phenomenon occurs [29]. In the context of LDA, the regularization of the class sample covariance matrices is shown to be efficient to address such problems [20], [30]. This regularization is sometimes applied in combination with a preliminary feature selection/extraction step in order to increase the aforementioned ratio [31].

Making use of the pooled covariance estimate may decrease the variance. Indeed, while the scatter matrix of each class may differ greatly, the common covariance matrix provides a more accurate classification since a higher number of samples are used for its estimation. This is certainly a straightforward type of regularization, which reduces the number of parameters to be estimated. Regularization can be carried out according to the use of the pooled covariance estimate, which is more reliable and stable than the scatter matrix of each class (as estimated with a higher number of samples). In order to improve the estimate, a linear combination of the sample covariance and joint matrices can be used

$$\hat{\mathbf{S}}_i(\gamma) = (1-\gamma)\hat{\mathbf{S}}_i + \gamma\hat{\mathbf{S}}. \tag{8}$$

The low-dimensional subspace spanned by the eigenvectors with the largest eigenvalues of the common covariance matrix comprises the representative directions for classification in the feature space, but it results in high overlapping of the marginal probability density functions. On the contrary, overlapping is lower in directions corresponding to the smallest eigenvalues [20]. Therefore, they give the highest contribution to the discriminant function. To improve the estimate, the following linear combination has been also proposed [22]:

$$\hat{\mathbf{S}}(\lambda,\gamma) = (1-\lambda)\hat{\mathbf{S}}_i(\gamma) + \lambda\mathbf{I} \tag{9}$$

where λ and γ are regularization parameters, which control the variance by shrinking the eigendecomposition toward the average eigenvalue of the estimate $\hat{\mathbf{S}}$, i.e., $\lambda \rightarrow \text{trace}(\hat{\mathbf{S}}_i(\gamma)\lambda/n)$. Note that by fixing the regularization parameter to one, the classical (unregularized) LDA is obtained. Since a decrease in variance is achieved at the expense of a certain increase in bias, cross-validation is used to select good values for the regularization parameter. However, unless efficient versions of the RLDA are used, this constitutes a difficult task because of the sensitivity of the solution to this parameter and of the cost involved when dealing with high-dimensional data.

B. PLDA and RLDA

Two main forms of regularizing LDA equations have been presented so far in the literature, namely, the PLDA and the RLDA. In this section, we review their basic formulations, relations, and theoretical properties.

The PLDA regularizes the between-class scatter matrix by adding a symmetric positive semidefinite penalty matrix [19]. Specifically, the $\mathbf{\Sigma}_w$ matrix is replaced with $\boldsymbol{S}'_w = \mathbf{S}_w + \lambda\mathbf{\Theta}$, where $\mathbf{\Theta}$ is an $m \times m$ matrix such that $\mathbf{w}^\top\mathbf{\Theta}\mathbf{w}$ is large for "bad" solutions of $\mathbf{w}$, that is high values of $\|\mathbf{w}\|$. The *smoothness* matrix $\mathbf{\Theta}$ can be designed to highlight some specific features in the data, such as spatial homogeneity, relevance of certain spectral channels, etc. Note that the classical LDA is the particular case of the PLDA when $\lambda = 0$. This method has been successfully used in remote sensing applications [21]. However, the proper design of the penalization matrix is still an open (and very challenging) issue.

A simpler (yet effective) alternative to the design of the regularizer in the PLDA is the use of the RLDA. The RLDA deals with the singularity of $\mathbf{S}$ by adding a constant value to its diagonal elements, $\mathbf{S} + \lambda\mathbf{I}$, where $\lambda > 0$ and $\mathbf{I}$ is an identity matrix. It is easy to verify that the regularized matrix is positive definite and, thus, nonsingular [20]. Following the same notation as in Section II, and by substituting the regularized matrix $\mathbf{S}$ in (3), the maximization problem is now

$$\mathbf{G}^* = \arg\max_{\mathbf{G}} \left\{ \text{trace}\left(\left(\mathbf{G}^\top(\mathbf{S}+\lambda\mathbf{I})\mathbf{G}\right)^{-1}\mathbf{G}^\top\mathbf{S}_b\mathbf{G} \right) \right\}. \tag{10}$$

Note that the classical LDA is the limiting case of the RLDA when $\lambda = 0$.

The solution of the problem in (10) is obtained by computing the eigendecomposition of $(\mathbf{S}+\lambda\mathbf{I})^{-1}\mathbf{S}_b$. Therefore, one has to compute first the singular value decomposition (SVD) of $\mathbf{H}$, i.e., $\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal square matrices and $\mathbf{D}$ is diagonal. Then, $\mathbf{S} = \mathbf{H}\mathbf{H}^\top = \mathbf{U}\mathbf{D}\mathbf{D}^\top\mathbf{U}^\top$ and $\widetilde{\mathbf{S}} = \mathbf{S}+\lambda\mathbf{I} = \mathbf{U}\widetilde{\mathbf{D}}\mathbf{U}^\top$, where $\widetilde{\mathbf{D}} = \mathbf{D}\mathbf{D}^\top + \lambda\mathbf{I}$ can be large for high-dimensional data. After some operations, one can demonstrate that the optimal transformation $\mathbf{G}^*$ for the RLDA consists of the first $q < m$ columns of a matrix whose calculation involves inverting a matrix of the same size as $\widetilde{\mathbf{D}}$, i.e., of size m. As a consequence, the selection of the parameter λ and, more importantly, the computational cost involved in its selection are serious shortcomings to the generalized use of the RLDA in the context of (high-dimensional) image classification. Table II gives the pseudocode for the standard RLDA algorithm.

IV. Efficient RLDA

This section is devoted to the analysis of the version of the RLDA specifically introduced in this paper, and the relationships that this method has with other presented RLDA-, OLDA-, and ULDA-based algorithms.

A. Proposed Efficient RLDA

The high computational cost of traditional RLDA has been recently alleviated in [1] by noting that regularizing the total scatter matrix $\mathbf{S}$ is equivalent to regularizing its nonzero eigenvalues. This strategy has the advantage of a much lower computational cost for computation and, thus, for model

TABLE II
PSEUDOCODE FOR THE ULDA, THE OLDA, THE STANDARD RLDA, AND THE PROPOSED EFFICIENT RLDA

	ULDA	**OLDA**
1 :	Construct $\mathbf{H}_b$, $\mathbf{H}_w$ and $\mathbf{H}$	Construct $\mathbf{H}_b$, $\mathbf{H}_w$ and $\mathbf{H}$
2 :	Compute SVD of $\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$	Compute SVD of $\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$
3 :	$\mathbf{B} \leftarrow \mathbf{D}^{-1}\mathbf{U}\mathbf{H}_b$	$\mathbf{B} \leftarrow \mathbf{D}^{-1}\mathbf{U}\mathbf{H}_b$
4 :	Compute SVD of $\mathbf{B} = \mathbf{P}\mathbf{D}\mathbf{Q}^\top$	Compute SVD of $\mathbf{B} = \mathbf{P}\mathbf{D}\mathbf{Q}^\top$
5 :	$\mathbf{X} \leftarrow \mathbf{U}\mathbf{D}^{-1}\mathbf{P}$	$\mathbf{X} \leftarrow \mathbf{U}\mathbf{D}^{-1}\mathbf{P}$
6 :	$\mathbf{G} \leftarrow \mathbf{X}_q$, $q \leftarrow rank(\mathbf{B})$	QR-decompose $\mathbf{X}_q = \hat{\mathbf{Q}}\hat{\mathbf{R}}$
7 :		$\mathbf{G} \leftarrow \hat{\mathbf{Q}}$
	Standard RLDA	**Proposed RLDA [1]**
1 :	Construct $\mathbf{H}_b$, $\mathbf{H}_w$ and $\mathbf{H}$	Construct $\mathbf{H}_b$, $\mathbf{H}_w$ and $\mathbf{H}$
2 :	Compute SVD of $\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$	Compute SVD of $\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$
3 :	$\widetilde{\mathbf{D}} \leftarrow \mathbf{D}\mathbf{D}^\top + \lambda\mathbf{I}$	$\widetilde{\mathbf{H}}_b \leftarrow \mathbf{U}^\top\mathbf{H}_b$
4 :	Compute SVD of $\widetilde{\mathbf{D}}^{-1/2}\mathbf{U}^\top\hat{\mathbf{H}}_b = \mathbf{U}_b\mathbf{D}_b\mathbf{V}_b^\top$	$\widetilde{\mathbf{D}}_s \leftarrow \mathbf{D}_r^2 + \lambda\mathbf{I}_r$, $r \leftarrow rank(\mathbf{H})$
5 :	$\mathbf{G} \leftarrow \mathbf{U}\widetilde{\mathbf{D}}^{-1/2}\mathbf{U}_b$	Compute SVD of $\widetilde{\mathbf{D}}_s^{-1/2}\mathbf{U}^\top\mathbf{H}_b = \mathbf{U}_b\mathbf{D}_b\mathbf{V}_b^\top$
6 :		$\mathbf{G} \leftarrow \mathbf{U}\widetilde{\mathbf{D}}_s^{-1/2}\mathbf{U}_b$

selection. Note that the rank of $\mathbf{H}$, r, is typically lower than the dimensionality of the problem, m, and thus, it is not necessary to compute its whole SVD, i.e., $\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, but only the first r columns and rows of matrix $\mathbf{D}$ (hereafter denoted as $\mathbf{D}_r$). The main result in [1] can be written as follows:

$$(\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{S}_b = \mathbf{U}_r(\mathbf{D}_r\mathbf{D}_r + \lambda\mathbf{I}_r)^{-1}\mathbf{U}_r^\top\mathbf{S}_b \qquad (11)$$

where $\mathbf{I}$ is the m-dimensional identity matrix, r is the rank of $\mathbf{S}$ (which is equal to the number of its nonzero eigenvalues), and $\mathbf{I}_r$ is the identity matrix of size r. Here, the squared matrix $\mathbf{D}_r$ comprises the upper block $r \times r$ of the complete matrix $\mathbf{D}$ in the SVD decomposition of $\mathbf{H}$, which is $\mathbf{H} = \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r^\top$, where the subscript r marks the first rows and columns of $\mathbf{U}$ and $\mathbf{V}^\top$. Essentially, (11) states that the regularization of the total scatter matrix results in the regularization of its nonzero eigenvalues. Therefore, finding the eigenvectors of matrix $(\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{S}_b$ is equivalent to finding the eigenvectors of matrix $\mathbf{U}_r(\mathbf{D}_r\mathbf{D}_r + \lambda\mathbf{I}_r)^{-1}\mathbf{U}_r^\top\mathbf{S}_b$.

The main consequence of the algorithm is that the computational cost mainly depends on the SVD computation of $\mathbf{H}$ and on the computation of $\mathbf{U}^\top\mathbf{H}_b$, which are independent of λ, and thus, its search scales almost at linear cost [1] (see Table II, lines 2–5 of the RLDA pseudocodes). The λ parameter is now tuned in the range $\lambda \in [0, \infty)$, rather than between $[\epsilon, \infty)$, where $\epsilon > 0$. A proper selection of this parameter is critical since large values may distort the original data, and small ones may not be sufficient to avoid the instability problem.

B. ULDA and OLDA

Let us finally review the formulation of the recently proposed ULDA and OLDA methods and their relationships with the RLDA proposed in the previous section. In the framework of LDA-based methods, the scatter matrices can be transformed to the low-dimensional space by using a transformation matrix $\mathbf{G}$ as a result of a prespecified optimization criterion, different from that given by the Fisher score in (1). For instance, with $\mathbf{S}$ being potentially singular, the methods of OLDA and ULDA have been recently proposed [25], [26], [32], [33], which generalize the classical LDA by solving an eigenvalue problem on $\mathbf{S}^+\mathbf{S}_b$, where $^+$ indicates pseudoinverse, instead of the inverse matrix $\mathbf{S}^{-1}$. In this way, the transformed feature vectors are uncorrelated (orthogonal) in the transformed space for ULDA (OLDA). More details on the OLDA and ULDA methods can be found in [1], [25], [26], [32], and [33].

Table II compares the pseudocode for the ULDA and OLDA, along with that for the standard and proposed RLDA algorithms. It must be highlighted first that the optimal transformation $\mathbf{G}$ for the standard RLDA consists of the first q columns of $\mathbf{X}$, where $q = \text{rank}(\mathbf{S}_b)$, which induces a dramatic computational burden in high-dimensional problems, particularly because of the search over the regularization parameter λ. On the contrary, the complexity imposed by the efficient RLDA is much lower (essentially dominated by lines 2 and 3). In addition, note the similarity between the OLDA and the ULDA (the first five lines are identical), whose theoretical relation is extensively studied in [26]. Finally, the ULDA can be regarded as a special case of the RLDA when $\lambda = 0$, so better results are expected by using the latter. This is an interesting property of the proposed RLDA, as stated in [1], by which the RLDA approaches the ULDA as the regularization parameter tends to zero.

V. EXPERIMENTAL RESULTS

In this section, experimental results are presented on synthetic data and real hyperspectral remote sensing classification problems. In the first battery of experiments, we deal with standard 2-D binary toy examples in order to illustrate the capabilities of the presented efficient RLDA compared to several LDA-based methods and SVMs. The second battery of experiments deals with a wide range of ill-posed hyperspectral

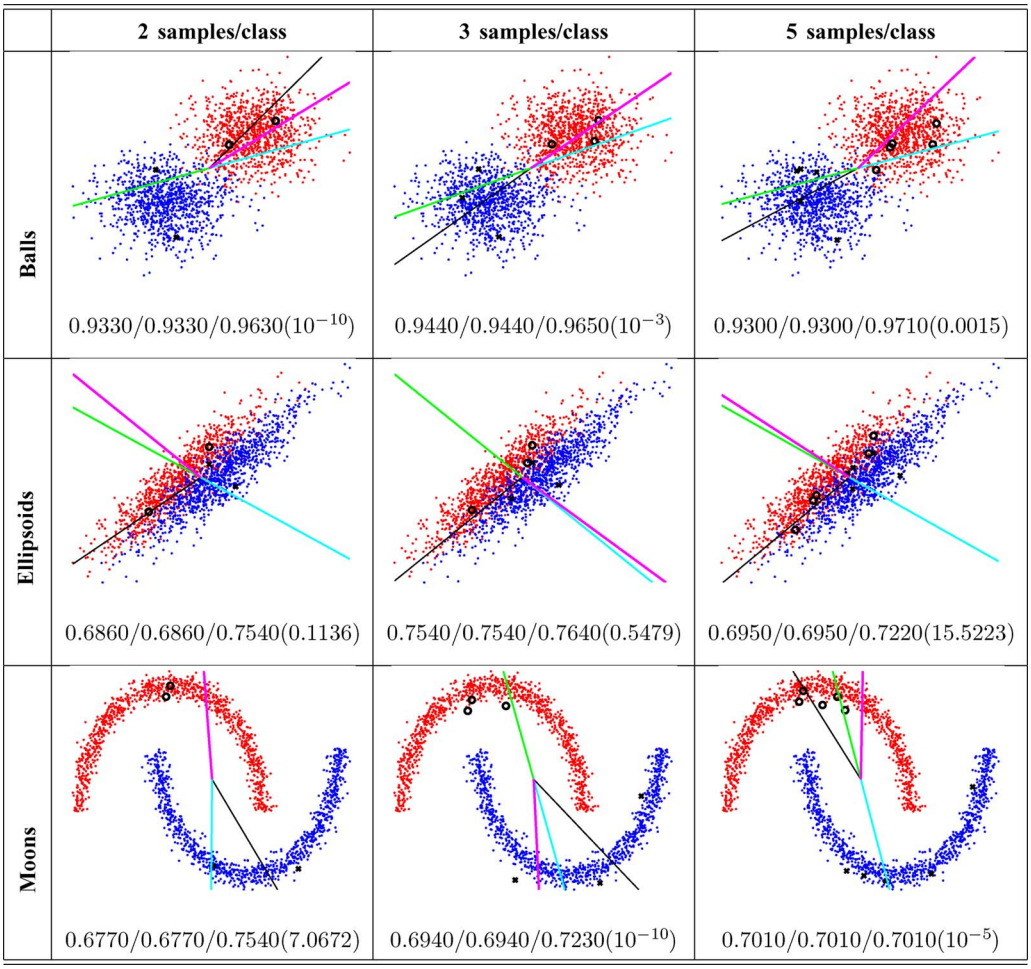

Fig. 2. Projections obtained with (black) PCA, (green) OLDA, (cyan) ULDA, and (pink) RLDA for different numbers of available samples for the classification of data in standard manifolds (overlapped Gaussian balls, ellipsoids, and two moons). The training samples are marked with black crosses (class -1) and circles (class $+1$), and the test samples are plotted in red (class -1) and blue (class $+1$) dots. On the bottom of each figure, we indicate the test κ statistic in the OLDA/ULDA/RLDA(λ) form.

remote sensing data classification problems, where a high number of images acquired by different sensors (and with different numbers of classes) are classified.

A. Experiment 1: Binary 2-D Toy Examples

In this section, we illustrate the projections obtained with the different LDA-based approaches used in this paper on a set of linear and nonlinear classification toy problems. For this purpose, we used the following three standard 2-D data sets: "balls," "ellipsoids," and "moons." We trained the classifiers with very few data ($n = \{2, 3, 5\}$ samples) and tested the extracted projections with 1000 test samples. Note that these are not strictly ill-posed situations since the input space dimension is $m = 2$ and the number of labeled samples is $n > m$. However, the examples are more intended to understand the capabilities of the methods in such challenging scenarios. Note that, in these scenarios, one can state a mild assumption for applicability; essentially, we require that the rank of the total scatter matrix be equal to the sum of the rank of the between-class scatter matrix and the rank of the within-class scatter matrix. Following this condition [33], a null-space LDA can be applied only marginally.

Fig. 2 shows the results for all data sets and LDA-based methods. We indicate the computed projections by PCA (solid gray), OLDA (dash-dotted gray), ULDA (dashed gray), and RLDA (solid black) and show at the bottom the κ statistic and the regularization parameter λ. Several conclusions can be extracted from these examples. First, the RLDA method outperforms the rest of the methods and allows us to extract more informative features with a very reduced number of training examples. For example, the extracted features in "balls" are near optimal with only two training samples/class. With more labeled samples, the RLDA fine-tunes the projection matrix $\mathbf{G}$. In the case of "ellipsoids," the RLDA also outperforms the rest of the methods. In particular, PCA yields a very poor feature extraction for this classification problem, while the RLDA is only slightly different from the OLDA/ULDA but still better. This is a direct consequence of the regularization, as can be noted by the increasing λ with the number of training samples. In the "two-moon" problem, results are poor for all methods since none of them takes into account the manifold structure of the data, and also, they produce linear projections only. In such difficult situations, the RLDA also yields better or equal accuracy than the OLDA/ULDA methods.

B. Experiment 2: Hyperspectral Image Classification

1) Data Collection: Five different hyperspectral image data sets with a total of nine images are used in our experiments. For each image data set, many ill-posed situations are simulated. Note that, in these scenarios, the condition in [33] suggests that a null-space LDA cannot be used, while the proposed regularized LDA can be safely applied.

a) AVIRIS Indian Pines: In our experiments, we used the standard Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) image taken over NW Indiana's Indian Pine test site in June 1992. Discriminating among the major crops can be very difficult (in particular, given the moderate spatial resolution of 20 m), which has made the scene a challenging benchmark to validate classification accuracy of hyperspectral imaging algorithms. The calibrated data are available online (along with detailed ground-truth information) from http://dynamo.ecn.purdue.edu/~biehl/.

Two different data sets were considered in the experiments. According to [34], we first used a part of the scene, called the *subset scene*, consisting of pixels $[27-94] \times [31-116]$ for a size of 68×86, which contains four labeled classes (the background pixels were not considered for classification purposes). In this subimage, there are four classes with uneven number of labeled samples, namely, "Corn-notill" (1008), "Grass/Trees" (732), "Soybeans-notill" (727), and "Soybeans-min" (1926). Second, we used the whole scene, consisting of the full 145×145 pixels, which contains 16 classes, ranging in size from 20 to 2468 pixels, and thus constituting a very difficult situation. In both images, we removed 20 noisy bands covering the region of water absorption and finally worked with 200 spectral bands.

b) DAISEX-1999 data set: This data set consists of labeled pixels of six different hyperspectral images (700×670 pixels) acquired with the 128-band HyMap airborne spectrometer during the DAISEX-99 campaign [5]. This instrument provides 128 bands across the reflective solar wavelength region of 0.4–2.5 μm with contiguous spectral coverage (except in the atmospheric water vapor absorption bands), bandwidths around 16 nm, very high signal-to-noise ratio, and a spatial resolution of 5 m. After data acquisition, a preliminary test was carried out to measure the quality of data. No significant signs of coherent noise were found. Bands 1, 2, 65, 66, 67, 97, and 128 were considered noisy bands due to their high variability but were included in our experiments. This issue constitutes an additional problem for classification and feature extraction. Training and validation sets were composed of 150 samples/class, and the best classifiers were selected using the cross-validation method.

c) Botswana data set: The NASA EO-1 satellite acquired a sequence of data over the Okavango Delta, Botswana, in 2001–2004. The Hyperion sensor on EO-1 acquires data at 30-m pixel resolution over a 7.7-km strip in 242 bands covering the 400–2500-nm portion of the spectrum in 10-nm windows. Preprocessing of the data was performed by the UT Center for Space Research to mitigate the effects of bad detectors, interdetector miscalibration, and intermittent anomalies. Uncalibrated and noisy bands that cover water absorption features were removed, and the remaining 145 bands were included as candidate features, [10–55, 82–97, 102–119, 134–164, 187–220]. The data, acquired on May 31, 2001, consist of observations from 14 identified classes intended to reflect the impact of flooding on vegetation. For more information, see [35] and visit http://www.csr.utexas.edu/.

d) KSC data set: The NASA AVIRIS instrument acquired data over the Kennedy Space Center (KSC), Florida, on March 23, 1996. AVIRIS acquires data in 224 bands of 10-nm width with center wavelengths from 400 to 2500 nm. The KSC data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. After removing water absorption and low-SNR bands, 176 bands were used for the analysis. Discrimination of land covers is difficult due to the similarity of spectral signatures for certain vegetation types. For classification purposes, 13 classes representing the various land-cover types that occur in this environment were defined. For more information, see [35] and visit http://www.csr.utexas.edu/.

2) Model Development and Free-Parameter Selection: In our experiments, we used LDA, PLDA, OLDA, ULDA, and the proposed version of the RLDA, along with the linear and the radial-basis-function (RBF) kernel SVMs. The only parameter to be tuned for the LDA-based algorithms is the regularization parameter λ for the PLDA and RLDA. We tuned it in the range $\lambda = \{10^{-10}, \ldots, 10^{6}\}$. Note that for $\lambda = 10^{-10}$, the RLDA behaves like the ULDA (cf. Section IV-B). In all our experiments with the LDA-based classifiers, after obtaining the transformed data, $a = G^{\top}x$, we fit a multivariate normal density to each class, with a pooled estimate of covariance.

For the case of the linear kernel machines, only the penalization factor C had to be tuned. For the case of the RBF kernel SVM classifiers, the σ parameter was additionally tuned. All RBF kernel widths were tuned in the range $\sigma = \{10^{-3}, \ldots, 10^{3}\}$, and the regularization parameter for the SVM was varied in $C = \{10^{-1}, \ldots, 10^{3}\}$. An exhaustive search among all free parameters is computationally unfeasible. Therefore, a nonexhaustive iterative search strategy (τ iterations) was used here. At each iteration, a sequential search of the minimum v-fold cross-validation-estimated kappa statistic on each parameter domain is performed by splitting the range of the parameter in L points.[1] Values of $\tau = 3$ and $L = 20$ exhibited good performance in our simulations. A *one-against-one* multiclassification scheme was adopted for the SVM classifier, while the intrinsic one-versus-all classification scheme (maximum vote) was used for the LDA-based algorithms. Before training, data were normalized to give zero mean and unit variance.

3) Numerical Comparison: Fig. 3 shows the overall accuracy (OA [%]) and kappa statistic (κ) as a function of the number of labeled samples for the different classifiers and considered images. We show mean and variance over 30 random realizations on the selection of training samples. These results show that the nonlinear SVM (with the RBF kernel) commonly outperforms linear methods, except in the case of images with Gaussian-like class distributions, such as in the DAISEX data

[1] In v-fold cross-validation, the training set is split in v disjoint groups: $v-1$ sets are used for training, while the remaining one is used for validation. The procedure is repeated v times. The best combination of free parameters is chosen by minimizing an averaged error measurement computed with the predictions on the v different validation sets.

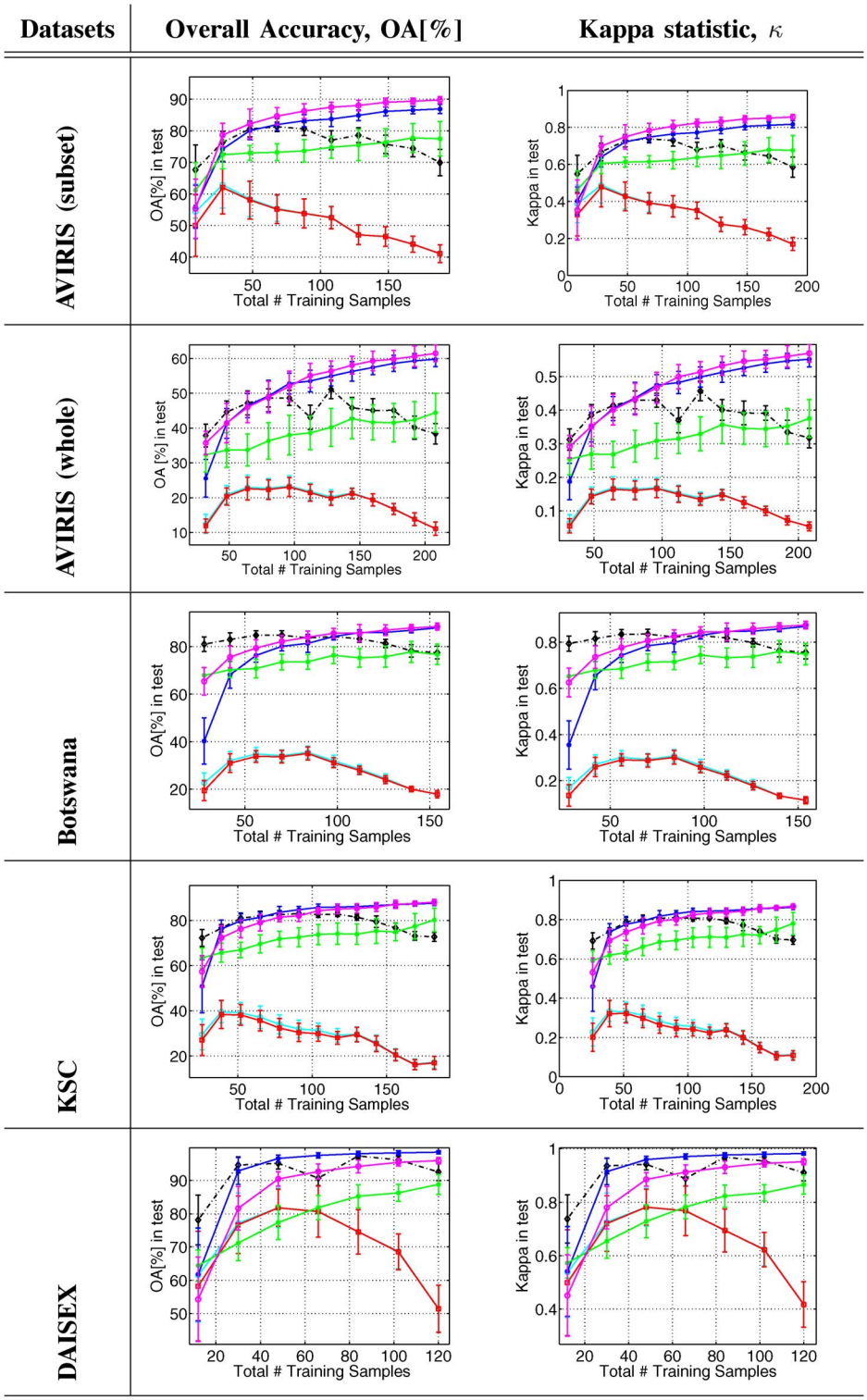

Fig. 3. (Left) OA [%] and (right) kappa statistic as a function of the number of labeled samples for different (in colors) classifiers and (in rows) images. Methods: LDA (not available in ill-posed situations), PLDA (black), OLDA (cyan), ULDA (red), linear SVM (green), RBF SVM (pink), and RLDA (blue). Mean and variance are shown over 30 realizations of random selection of training samples.

set, or in the case of extremely ill-posed situations (≤ 50 training samples). Certainly, simpler (linear) functions are best suited for learning in ill-posed situations.

For the case of the AVIRIS data set, it is noticeable that the nonlinear RBF SVM outperforms the rest of the methods, but that the proposed RLDA yields comparable results. The PLDA produces unstable results, probably due to the use of a trivial regularizer $\mathbf{\Theta}$, which was preset to a classwise weighted identity matrix. The linear SVM produces poor results, while the OLDA and ULDA perform very similarly, providing extremely poor results. For the Botswana data set, the PLDA outperforms the rest of the methods in hard ill-posed problems. However,

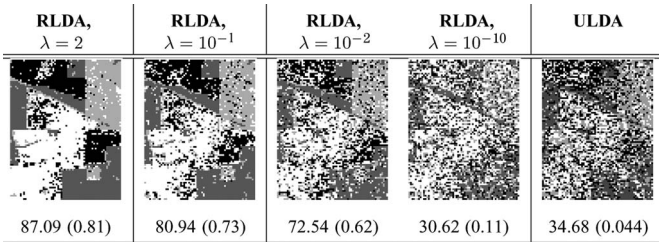

Fig. 4. Classification maps for the illustrative example of the AVIRIS subset image. The solutions obtained with different regularization parameters λ for RLDA and the results obtained with ULDA are shown. The classifiers were trained with as many samples as spectral bands, $N = n$, for illustration purposes of moderate ill-posed situations. The RLDA classifier was trained following a fivefold cross-validation strategy for free-parameter selection. We also include the obtained OA [%] and (in parentheses) kappa statistic κ.

in more realistic cases, the RBF SVM and RLDA perform better than other methods. Again, the linear SVM, OLDA, and ULDA produce poor results. In the case of the KSC data set, the RLDA outperforms the rest of the methods (including the RBF SVM) in all situations. Again, the PLDA produces unstable results, while the linear SVM needs an increasing number of training samples to produce acceptable results. These results are confirmed when analyzing the average and variance results for the six images in the DAISEX project data set.

In conclusion, the presented RLDA generally outperforms the rest of the linear methods, and, in some cases, also the SVM with the RBF kernel, resulting in an excellent tradeoff between accuracy and complexity. Both the OLDA and the ULDA produce poor results in all cases, which suggests that mapping to uncorrelated/orthogonal features is not a good choice in the case of ill-posed hyperspectral remote sensing problems, as the estimate of scatter matrices is rather loose on the small training set [19], [26]. Certainly, high variance of the within-class scatter matrices deteriorates their efficiency when the data density is sparse or badly represented. Following [26], the ULDA removes the correlation among the features in the transformed space, which is theoretically sound but may be sensitive to the noise in the data. A similar effect was observed with the OLDA. However, as far as the ULDA becomes the limiting case for the RLDA method when $\lambda \mapsto 0$ [1], this problem can be overcome by making use of a proper regularization, i.e., by adding the regularization term to nonzero eigenvalues of the scatter matrix. Fig. 4 shows how crucial the role of the regularization parameter is in the crossover region, i.e., when the total number of samples is about the number of spectral bands. On the one hand, it shows the effectiveness of the regularization applied to the proposed RLDA [1]. On the other hand, it explains the failure of the ULDA/OLDA in the classification of high-dimensional possibly noisy data (see Fig. 5).

C. Visual Comparison

In this section, we devote attention to the visual inspection of the classified maps for all test images in the collapse of ill-posing, i.e., we used as many samples as spectral bands, $N = n$. We drop from the comparison the OLDA and ULDA classifiers because of their poor performance in the ill-posed region. Also, the PLDA is withdrawn because of its unsatisfactory results in the collapse situation. Figs. 6–9 show the RGB composition

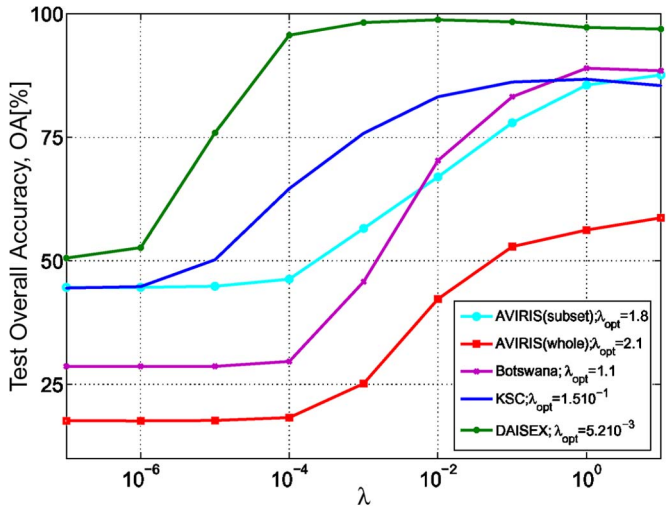

Fig. 5. OA [%] as a function of the value of regularization parameter λ with the ULDA classifier for various images. The total number of labeled samples is nearly equal to the number of spectral bands. Here, $\lambda_{\rm opt}$ corresponds to the best classification model.

maps and the obtained classification maps. All classifiers were trained following a fivefold cross-validation strategy for free-parameter selection.

It is noticeable that the performance of the linear SVM produces good but suboptimal results, suggesting that maximum-margin regularization may not be an optimal choice in ill-posed classification scenarios. The OLDA and ULDA both perform equally poor, as illustrated in the previous section. In general, it is observed that the nonlinear RBF SVM performs slightly better than the proposed RLDA method, but differences are not either numerically or statistically significant. Note that, in this (more realistic) scenario, the RLDA results are very competitive with the state-of-the-art nonlinear SVM, with the additional advantage of a much lower computational effort and user intervention.

It should be noted that the RLDA method also offers good spatial consistency in the classification maps; see, for instance, the river basin in the KSC image (Fig. 8), or the good spatial homogeneity, which is particularly significant when detecting the cirrus clouds in Fig. 7. In the case of the particular choice of the DAISEX image (Fig. 9), very good performance is also obtained with the RLDA. In this image, corn classification seems to be the most troublesome, and errors are mainly committed with the bare soil class. The reason for this is the presence of a whole field of two-leaf corn in the early stage of maturity, where soil was predominant and was not accounted for the reference labeled image (see the big round crop on the left which is misclassified by the linear SVM and correctly identified by the more complex RBF SVM and the proposed RLDA). Inclusion of contextual information in the form of composite covariances could improve even more this capability with eventually no additional cost for the proposed RLDA.

VI. Discussion and Conclusion

This paper has presented an exhaustive evaluation and review of LDA-based methods in the context of ill-posed hyperspectral image classification. In addition, an efficient version of

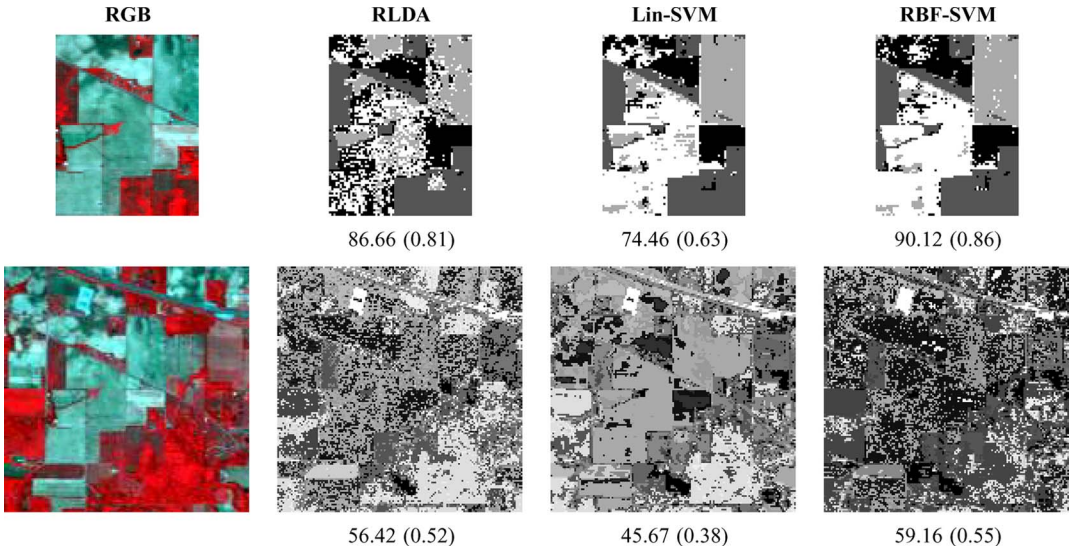

Fig. 6. RGB composition and classification maps for the AVIRIS Indian Pines images (subset scene in top row, whole scene in bottom row) and classification methods. The classifiers were trained with as many samples as spectral bands, $N = n$, for illustration purposes of moderate ill-posed situations. All classifiers were trained following a fivefold cross-validation strategy for free-parameter selection. We also include the obtained OA [%] and (in parentheses) kappa statistic κ.

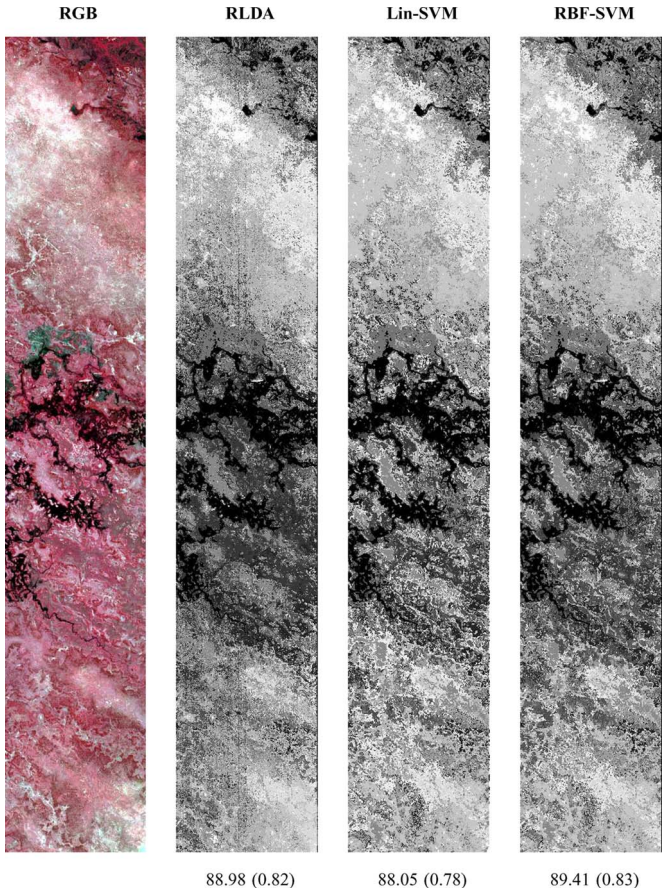

Fig. 7. RGB composition and classification maps for the Botswana image and classification methods. The classifiers were trained with as many samples as spectral bands, $N = n$, for illustration purposes of moderate ill-posed situations. All classifiers were trained following a fivefold cross-validation strategy for free-parameter selection. We also include the obtained OA [%] and (in parentheses) kappa statistic κ.

the RLDA has been introduced for solving remote sensing classification problems. The presented method consists in regularizing the nonzero eigenvalues of the total scatter matrix, rather than regularizing the scatter matrix, which results in faster solutions than the standard RLDA algorithm, and thus, the critical regularization parameter can be more accurately estimated.

The theoretical analysis and the empirical experimental results obtained on many different hyperspectral data sets (and confirmed on toy examples) point out the superiority of the presented RLDA over other linear classifiers, including the maximum-margin linear SVM. In greater detail, different kinds of regularization result in different effectiveness of the LDA algorithm. The RLDA proved also competitive with the nonlinear SVM based on RBF kernels, exhibiting an excellent tradeoff between classification accuracy and computational complexity. We should note that even though the nonlinear SVM exhibited higher accuracies, the computational cost associated to model selection was a clear disadvantage. It is worth noting that these conclusions are related to the analysis of strongly ill-posed classification problems, i.e., problems in which the number of training samples is similar to the number of features. Significantly different conclusions on the effectiveness of the nonlinear SVM (which favor this last classifier) can be obtained when the considered problem is moderately ill posed (see [7]).

The results obtained in this paper confirm that complex ill-posed problems associated with hyperspectral data often take advantage from the use of simple classifiers characterized by few parameters. This is particularly true in strongly ill-posed problems, when the ratio between the number of training samples and the number of features is very small.

Future work is tied to the use of the contextual information as efficient regularizer for LDA methods. The inclusion of the contextual information can be carried out through the use of contextual composite covariances, which were previously presented in the context of kernel methods [36]. In the near future, we are also interested on the kernelization of the discriminant functions, which has received some interest recently [37].

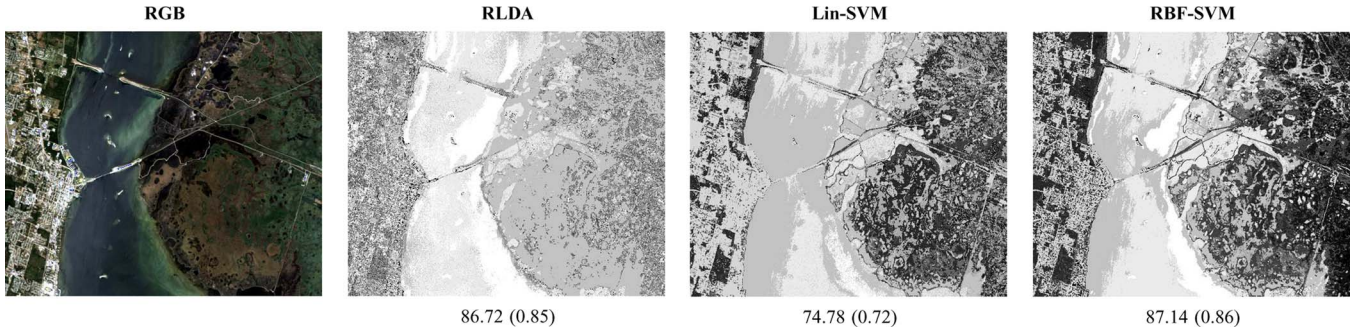

Fig. 8. RGB composition and classification maps for the KSC image and classification methods. The classifiers were trained with as many samples as spectral bands, $N = n$, for illustration purposes of moderate ill-posed situations. All classifiers were trained following a fivefold cross-validation strategy for free-parameter selection. We also include the obtained OA [%] and (in parentheses) kappa statistic κ.

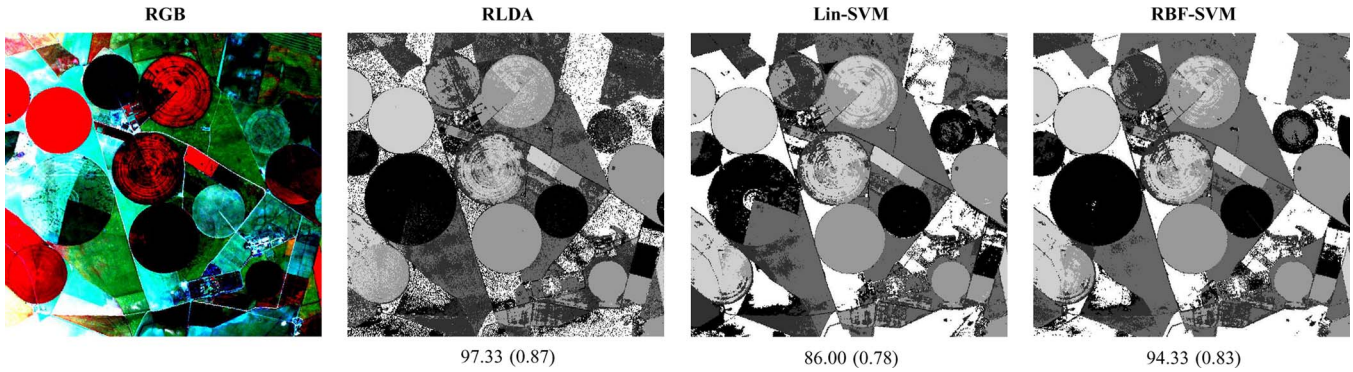

Fig. 9. RGB composition and classification maps for a representative scene of the DAISEX image data set and classification methods. The classifiers were trained with as many samples as spectral bands, $N = n$, for illustration purposes of moderate ill-posed situations. All classifiers were trained following a fivefold cross-validation strategy for free-parameter selection. We also include the obtained OA [%] and (in parentheses) kappa statistic κ.

ACKNOWLEDGMENT

The authors would like to thank Prof. M. Crawford for providing the Botswana and KSC data sets, Prof. D. Landgrebe for providing the AVIRIS Indian Pines data set, and Prof. J. Ye for providing the source code of the OLDA and ULDA algorithms. This paper was done while at the Departament d'Enginyeria Electrònica, Universitat de València, València, Spain, and during a short stage at the Department of Information and Communication Technology, University of Trento, Trento, Italy.

REFERENCES

[1] J. Ye, T. Xiong, R. Janardan, J. Bi, V. Cherkassky, and C. Kambhamettu, "Efficient model selection for regularized linear discriminant analysis," in *Proc. CIKM*, Arlington, VA, 2006, pp. 532–539.

[2] T. M. Lillesand, R. W. Kiefer, and J. W. Chipman, *Remote Sensing and Image Interpretation*, 5th ed. New York: Wiley, 2004.

[3] M. Chi and L. Bruzzone, "Semisupervised classification of hyperspectral images by SVMs optimized in the primal," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 6, pp. 1870–1880, Jun. 2007.

[4] L. Bruzzone, M. Chi, and M. Marconcini, *Hyperspectral Data Exploitation: Theory and Applications*, C.-I. Chang, Ed. Hoboken, NJ: Wiley, 2007.

[5] G. Camps-Valls, L. Gómez-Chova, J. Calpe, E. Soria, J. D. Martín, L. Alonso, and J. Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1530–1542, Jul. 2004.

[6] 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.

[7] 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.

[8] A. Plaza, J. Benediktsson, J. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 40, 2007, to be published.

[9] V. Cherkassky and M. Yunqian, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Netw.*, vol. 17, no. 1, pp. 113–126, Jan. 2004.

[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.

[11] M. R. Schwaller, "A geobotanical investigation based on linear discriminant and profile analyses of airborne thematic mapper simulator data," *Remote Sens. Environ.*, vol. 23, no. 1, pp. 23–34, Oct. 1987.

[12] G. Peng, P. Ruiliang, and Y. Bin, "Conifer species recognition: An exploratory analysis of *in situ* hyperspectral data," *Remote Sens. Environ.*, vol. 62, no. 2, pp. 189–200, Nov. 1997.

[13] M. L. Clark, D. A. Roberts, and D. B. Clark, "Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales," *Remote Sens. Environ.*, vol. 96, no. 3/4, pp. 375–398, Jun. 2005.

[14] A. Lobo, "Image segmentation and discriminant analysis for the identification of land cover units in ecology," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 5, pp. 1136–1145, Sep. 1997.

[15] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Int. J. Digit. Libr.*, vol. 1, pp. 108–121, 1997.

[16] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intell. Syst.*, vol. 13, no. 2, pp. 44–49, Mar./Apr. 1998.

[17] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.

[18] P. N. Belhumeour, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[19] T. Hastie, A. Buja, and R. Tibishirani, "Penalized discriminant analysis," *Ann. Stat.*, vol. 23, no. 1, pp. 73–102, 1995.

[20] J. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, Mar. 1989.

[21] B. Yu, M. Ostland, P. Gong, and R. Pu, "Penalized discriminant analysis of *in situ* hyperspectral data for conifer species recognition,"

IEEE Trans. Geosci. Remote Sens., vol. 37, no. 5, pp. 2569–2577, Sep. 1999.
[22] T. Hastie, R. Tibishirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.
[23] A. Berge and A. H. Schistad Solberg, "Structured Gaussian components for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3386–3396, Nov. 2006.
[24] A. Berge, A. C. Jensen, and A. H. S. Solberg, "Sparse inverse covariance estimates for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 5, pp. 1399–1407, May 2007.
[25] Z. Jin, J. Y. Yang, Z. S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognit.*, vol. 34, no. 7, pp. 1405–1416, 2001.
[26] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *J. Mach. Learn. Res.*, vol. 6, pp. 483–502, Dec. 2005.
[27] T. Bandos, L. Bruzzone, and G. Camps-Valls, "Efficient regularized LDA for hyperspectral image classification," in *Proc. SPIE Int. Symp. Remote Sens. XII*. Florence, Italy: SPIE-Int. Soc. Opt. Eng., Sep. 2007, vol. 6748.
[28] K. Fukunaga, *Introduction to Statistical Pattern Classification*. San Diego, CA: Academic, 1990.
[29] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
[30] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 763–767, Jul. 1996.
[31] J. P. Hoffbeck and D. A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 5, pp. 1096–1105, May 2004.
[32] J. Ye, R. Janarda, Q. Li, and H. Park, "Feature extraction via generalized uncorrelated linear discriminant analysis," in *Proc. ICML*, 2004, p. 113.
[33] J. Ye, R. Janarda, Q. Li, and H. Park, "Feature reduction via generalized uncorrelated linear discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1312–1322, Oct. 2006.
[34] J. A. Gualtieri, S. R. Chettri, R. F. Cromp, and L. F. Johnson, "Support vector machine classifiers as applied to AVIRIS data," in *Proc. 8th JPL Airborne Geosci. Workshop*, Feb. 1999, pp. 217–227.
[35] J. Ham, C. Yangchi, M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.
[36] G. Camps-Valls, L. Gómez-Chova, J. Muñoz Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.
[37] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Neural Netw. Signal Process. IX*, 1999, vol. 27, pp. 41–48.

Tatyana V. Bandos (Marsheva) received the M.S. degree in physics from the Kharkiv National University, Kharkiv, Ukraine, in 1981 and the Ph.D. degree (with emphasis in the theoretical condensed-matter physics) from the B. Verkin Institute for Low Temperature Physics and Engineering, National Academy of Sciences of Ukraine, Kharkiv, in 1995.

She was nominated to the Associateship at the International Center for Theoretical Physics in 1997, Trieste, Italy. Since her stint at the University of Valencia (UVEG), Valencia, Spain, in 2002, she has held the visiting position in the Spectroscopy Group and has been carrying out research with the Digital Signal Processing Group. She is currently a Researcher with the Institute of Pure and Applied Mathematics at the Polytechnic University of Valencia (IUMPA-UPV). Nowadays, her research interests include the field of heat and mass transfer and the semisupervised and supervised methods for hyperspectral data analysis in geoscience. Her publications are related to nonlinear dynamics of thermal waves in superconducting magnets, low-dimensional exactly solvable models of strongly correlated electrons, mesoscopic magnetic systems, support vector machines, and graph-based approaches to classification and regression problems.

Lorenzo Bruzzone (S'95–M'98–SM'03) received the Laurea (M.S.) degree in electronic engineering (*summa cum laude*) and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

From 1998 to 2000, he was a Postdoctoral Researcher with the University of Genoa. Since 2000, he has been with the University of Trento, Trento, Italy, where he his currently a Full Professor of telecommunications with the Department of Information Engineering and Computer Science. He teaches remote sensing, pattern recognition, and electrical communications. He is also the Head of the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento. His current research interests include remote sensing image processing and recognition (analysis of multitemporal data, feature extraction and selection, classification, regression and estimation, data fusion, and machine learning). He conducts and supervises research on these topics within the frameworks of several national and international projects. He is an Evaluator of project proposals for many different governments (including the European Commission) and scientific organizations. He is the author or coauthor of 60 scientific publications in referred international journals, more than 120 papers in conference proceedings, and seven book chapters. He is a Referee for many international journals and has served on the scientific committees of several international conferences. He is a Member of the Managing Committee of the Italian Inter-University Consortium for Telecommunications and a Member of the Scientific Committee of the India–Italy Center for Advanced Research.

Dr. Bruzzone ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (Seattle, July 1998). He was a recipient of the Recognition of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Best Reviewers in 1999 and was a Guest Editor of a Special Issue of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING on the subject of the analysis of multitemporal remote sensing images (November 2003). He was the General Chair and Cochair of the First and Second IEEE International Workshops on the Analysis of Multi-Temporal Remote Sensing Images (MultiTemp) and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. From 2004 to 2006, he was an Associated Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and is currently an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He is also a member of the International Association for Pattern Recognition and the Italian Association for Remote Sensing (AIT).

Gustavo Camps-Valls (M'04–SM'07) was born in València, Spain, in 1972. He received the B.Sc. degree in physics, the B.Sc. degree in electronics engineering, and the Ph.D. degree in physics from the Universitat de València, València, in 1996, 1998, and 2002, respectively.

He is currently an Associate Professor with the Departament d'Enginyeria Electrònica, Escola Tècnica Superior d'Enginyeria, Universitat de València, where teaches electronics, advanced time series processing, machine learning for remote sensing, and digital signal processing. His research interests include the development of machine learning algorithms for signal and image processing, with special attention to adaptive systems, neural networks, and kernel methods. He conducts and supervises research on these topics. He is the author or coauthor of 50 papers in refereed international journals, more than 70 international conference papers, and several book chapters and is the Editor of the books entitled *Kernel methods in bioengineering, signal and image processing* (IGI, 2007) and *Kernel methods for remote sensing data analysis* (Wiley, 2009). He has served as a Reviewer in many international journals and on the program committees of SPIE Europe, IGARSS, IWANN, and ICIP.

Dr. Camps-Valls was a member of the EUropean Network on Intelligent TEchnologies for Smart Adaptive Systems (EUNITE) and the Spanish Thematic Networks on "Pattern Recognition" and "Biomedical Engineering." He is active in the R&D sector through a high number of funded projects from both public and industrial partners, both at national and international levels. He is an Evaluator of project proposals and scientific organizations. Since 2009 he is member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society. Visit http://www.uv.es/gcamps for more information.