Hyperspectral Image Classification With Independent Component Discriminant Analysis

Alberto Villa, Jon Atli Benediktsson, Jocelyn Chanussot, Christian Jutten

Hyperspectral Image Classification With Independent Component Discriminant Analysis

Alberto Villa, *Student Member, IEEE*, Jón Atli Benediktsson, *Fellow, IEEE*, Jocelyn Chanussot, *Senior Member, IEEE*, and Christian Jutten, *Fellow, IEEE*

***Abstract*—In this paper, the use of Independent Component (IC) Discriminant Analysis (ICDA) for remote sensing classification is proposed. ICDA is a nonparametric method for discriminant analysis based on the application of a Bayesian classification rule on a signal composed by ICs. The method uses IC Analysis (ICA) to choose a transform matrix so that the transformed components are as independent as possible. When the data are projected in an independent space, the estimates of their multivariate density function can be computed in a much easier way as the product of univariate densities. A nonparametric kernel density estimator is used to compute the density functions of each IC. Finally, the Bayes rule is applied for the classification assignment. In this paper, we investigate the possibility of using ICDA for the classification of hyperspectral images. We study the influence of the algorithm used to enforce independence and of the number of IC retained for the classification, proposing an effective method to estimate the most suitable number. The proposed method is applied to several hyperspectral images, in order to test different data set conditions (urban/agricultural area, size of the training set, and type of sensor). Obtained results are compared with one of the most commonly used classifier of hyperspectral images (support vector machines) and show the comparative effectiveness of the proposed method in terms of accuracy.**

***Index Terms*—Bayesian classification, curse of dimensionality, hyperspectral data, Independent Component (IC) Analysis (ICA).**

I. Introduction

HYPERSPECTRAL images are composed of hundreds of bands with a very high spectral resolution, from the visible to the infrared region. The wide spectral range, coupled with an always increasing spatial resolution, allows us to better characterize materials and gives the ability to pinpoint ground objects laying on the observed surface and to distinguish between spectrally close ground classes, making hyperspectral imagery suitable for land cover classification. Due to their features, hyperspectral data have, in recent years, gained a continuously growing interest among the remote sensing community [1], [2].

The huge quantity of information and the high spectral resolution of hyperspectral images give the possibility to solve problems which usually cannot be solved by multispectral images. In classification of hyperspectral images, the higher dimensionality of the data increases the capability to detect and distinguish various classes with improved accuracy. However, several significant issues need to be considered in the classification process for this kind of images. The most common of such issues are the spatial variability of the spectral signature, the high cost of true sample labeling, the quality of data, and problems associated with the very high number of spectral channels. The large dimensionality of the data in the spectral domain leads to theoretical and practical problems. For example, for high-dimension data, normally distributed data have a tendency to concentrate on the tails of the distribution, conversely to what appears for low-dimension data (one or two dimensions). Consequently, intuitive considerations, based on the "bell shape" 1- or 2-D distribution, fail for high-dimension Gaussian distributions. For the purpose of classification, these problems are related to the course of dimensionality. Therefore, very important information for land cover classification can be hidden in a relatively small dimensional area of the data space and can be easily neglected.

In the context of supervised classification, one of the most challenging issues is related to the small ratio between the number of samples used for training and the number of features of the data. As the dimension of the data space becomes higher, the number of training samples necessary to define the statistical behavior of the data increases exponentially [7], which makes it impossible to obtain reasonable estimates of the class-conditional probability density functions used in standard statistical classifiers. The first consequence is that when increasing the number of features of the data used as input of the classifier over a given threshold (which depends on the number of training samples and the kind of classifier adopted), the classification accuracy decreases, according to the so-called Hughes' phenomenon [6]. Because of the aforementioned limitations, parametric classifiers such as the maximum likelihood classifier [5] or the Bayesian classifier [3], [4], which model probability density functions for individual classes with parameters estimated from the training samples, are often ineffective when used for classification of hyperspectral data.

Manuscript received April 19, 2010; revised September 13, 2010, December 2, 2010, and January 27, 2011; accepted April 29, 2011. This work was supported by the European Community's Marie Curie Research Training Networks Programme under Contract MRTN-CT-2006-035927, Hyperspectral Imaging Network (HYPER-I-NET).

A. Villa is with the GIPSA-Lab, Grenoble Institute of Technology (Grenoble INP), 38402 Grenoble, France, and also with the University of Iceland, 107 Reykjavik, Iceland (e-mail: alberto.villa@hyperinet.eu).

J. A. Benediktsson is with the Faculty of Electrical and Computer Engineering, University of Iceland, 107 Reykjavik, Iceland (e-mail: benedikt@hi.is).

J. Chanussot is with the GIPSA-Lab, Grenoble Institute of Technology (Grenoble INP), 38402 Grenoble, France (e-mail: jocelyn.chanussot@gipsa-lab.grenoble-inp.fr).

C. Jutten is with the GIPSA Lab, Université Joseph Fourier, 38402 Grenoble, France, and also with the Institute Universitaire de France, 75005 Paris, France (e-mail: christian.jutten@gipsa-lab.grenoble-inp.fr).

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.2011.2153861

It is well known that the probability of error achieved with a Bayes classifier is the best that can be obtained [5], thus making Bayesian classifiers attractive in pattern recognition. Nevertheless, the construction of an optimal Bayesian classifier is very challenging when dealing with high-dimensional data, which would require large training sets in order to obtain accurate estimates of the density functions of the classes. To overcome the problem of a small size of the labeled samples, when these classifiers are applied, it is often assumed that each class can be represented by a multivariate normal model depending on the mean and covariance matrix of the class-specific data. This assumption can be accepted in the case of low input dimensionality, but it is usually far from the reality for hyperspectral remote sensing data, leading to low generalization capabilities for nonlabeled data and, consequently, to poor classification accuracies. Many efforts have been reported in the literature to overcome the intrinsic problems of high-dimensionality data [8]–[13]. The main approaches that can be found in the literature are regularization of the sample covariance matrix, semisupervised classification for the exploitation of the classified (semilabeled) samples, and projection of the data into a lower dimensional space by preprocessing hyperspectral data with feature selection/extraction.

In order to reduce the small sample size problem and mitigate the course of dimensionality, several improved covariance matrix estimators have been proposed to reduce the variance of the estimate [8], [9]. The main problem involved by this kind of estimators is the possibility that the estimated covariance matrices overfit the few available training samples and lead to a poor approximation of statistics for the testing set and the whole image to be classified. Semisupervised classifiers give a rough classification of the data using the information of the training set and then iteratively update the class statistics according to the results of the classification [10]. The main drawbacks of this approach are the required high computational burden and the risk of overfitting when a limited number of training samples is available. Finally, another approach that has been proposed to overcome the problem of high dimensionality of the data is to use feature reduction/selection algorithms in order to reduce the dimensionality of the input space. Many techniques have been proposed, such as Decision Boundary Feature Extraction [11], Projection Pursuit [12], and Nonparametric Weighted Feature Extraction [13]. Nevertheless, the higher computational time required or the inevitable loss of information introduced by these techniques often represents an obstacle in obtaining high performances, in terms of processing time or classification accuracy.

Advanced classifiers like artificial neural networks (NNs) [14]–[16] and kernel-based classifiers [17], [23], [26]–[28] have more recently been applied for hyperspectral data classification, because they are distribution free and do not make assumptions about the density functions of the data. Multilayer NNs [14] basically suffer from two main limitations. The first limitation is that the number and the size of hidden layers need to be set, and this is not a straightforward task. The second limitation is that a very large number of iterations are sometimes needed to find a solution, making feature reduction a very useful step before the classification process. RBF NNs [15] overcome these shortcomings, but their classification accuracy strongly depends on the selection of the centers and widths of the kernel functions associated with the hidden neurons of the network. Kernel methods have been widely investigated in the last decade for remote sensing and hyperspectral data analysis. Such methods show even better performances than NNs in terms of accuracies, also providing good results in case of very limited training sets. During recent years, a number of powerful kernel-based learning classifiers (e.g., support vector machines (SVMs) [23], kernel Fisher discriminant analysis [26], support vector clustering [27], and the regularized AdaBoost algorithm [28]) have been proposed in the machine learning community, providing successful results in various fields. SVMs, in particular, have been widely investigated recently in the remote sensing community [24], [25], [29], [30]. Camps-Valls and Bruzzone compared in [17] a number of these methods. SVMs provided the most attractive classifier when dealing with high-dimensional data, providing very good results in terms of accuracy and robustness to common levels of noise. The main limitations of SVMs are the training time and the need to find the optimal parameters for the kernel. The training time, although much smaller than other kernel methods, quadratically depends on the size of the training set and can be very large, particularly when a large number of labeled samples is available. The choice of the parameters of the kernel is usually done using a cross-validation approach. Bazi and Melgani recently proposed a classification approach based on Gaussian processes [31], which showed results similar to those of SVMs but with a bigger computational burden. Due to the challenging problems of hyperspectral data classification, several approaches have been recently proposed to exploit also the spatial information of the data [32], [33].

In this paper, a nonparametric method for discriminant analysis based on the application of a Bayesian classification rule on a signal composed by independent components (ICs) originally presented in [34] is proposed for the classification of hyperspectral images. The main characteristics of the method are the use of IC Analysis (ICA) to retrieve ICs from the original data and the estimate of the multivariate density in the new data space computed with the ICA. When the data are projected in an independent space, the estimates of their multivariate density function can be computed in a much easier way as the product of univariate densities. The use of ICA is an elegant way which allows us to overcome the problem of the high dimensionality of input data, obtaining reliable estimates of the class conditional densities which can be used to build a Bayesian classifier. A nonparametric kernel density estimator is used to compute the density function of each IC. Finally, the Bayes rule is applied for classification assignment. The main contributions of this paper are the following: First, we propose an in-depth experimental analysis to highlight the potentialities of the method when used to classify hyperdimensional data. Second, we propose a simple but effective approach to choose the number of ICs which has to be retained for the classification process, in order to make the classifier suitable for hyperspectral data analysis. Finally, we perform a detailed comparison with respect to the SVM, one of the most used hyperspectral classifiers, considered as the one providing the best results.

The remainder of this paper is organized as follows. In Section II, the general framework of IC Discriminant Analysis (ICDA) is introduced. The experimental part is shown in Section III, and finally, conclusions are drawn in Section IV.

II. ICDA

The proposed method is a generalization of the quadratic discriminant analysis, where the ability of ICA to retrieve components as independent as possible is exploited to estimate the class-conditional joint densities $f_k(\mathbf{x})$ as the product of the marginal densities of the transformed components. The joint densities, which are hard to estimate when dealing with high-dimensional data, can be computed in a much simpler way in an independent space. The risk incurred when performing a classification of a measured vector $\mathbf{x}$ into one of K possible classes is given by

$$R(\hat{k}|\mathbf{x}) = \frac{\sum_{k=1}^{K} L(k,\hat{k}) f_k(\mathbf{x})\pi_k}{\sum_{k=1}^{K} f_k(\mathbf{x})\pi_k} \tag{1}$$

where π_k is the *a priori* probability that a sample could belong to the class k, f_k is the class-conditional *a priori* density of class k, and L is the cost or loss incurred when assigning the sample $\mathbf{x}$, belonging to the class k, to the class $\hat{k}$. In the case of hard classification (i.e., classification where only one class is selected), this cost is expressed by the so-called *symmetrical* loss function

$$L(k,\hat{k}) = \begin{cases} 0, & \text{if } k = \hat{k} \\ 1, & \text{if } k \neq \hat{k}. \end{cases}$$

By choosing $\hat{k}$ such that the numerator of (1) is minimized, this leads to the so-called Bayes decision rule. In the case of hard classification, the Bayes rule reduces to the following rule: Assign $\mathbf{x}$ to the class $\hat{k}$ such that

$$\hat{k} = d(\mathbf{x}) = \arg\max \{f_k(\mathbf{x})\pi_k\} \qquad k = 1, \ldots, K. \tag{2}$$

The design of the Bayes classifier is then determined by the conditional densities $f_k(\mathbf{x})$ and by the prior probabilities π_k. While the prior probabilities can be easily obtained from the training set, following the relation

$$\pi_k = N_k/N \tag{3}$$

where N_k is the number of samples of the class k and N is the overall number of samples of the training set, the determination of the class-conditional density is much more challenging. Owing to its analytical tractability, the Gaussian (or *normal*) multivariate density is the most often used density for classification. The general expression of a multivariate normal density in d dimensions is written as

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{4}$$

where $\mathbf{x}$ is a d-component column vector, $\boldsymbol{\mu}$ is the d-component mean vector, Σ is the d by d covariance matrix, and $|\Sigma|$ and Σ^{-1} are its determinant and its inverse. Finally, $(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}$ denotes the transpose of $(\mathbf{x}-\boldsymbol{\mu})$. These classification rules are, in general, derived by assuming that the class-conditional densities are p-variate normal with mean vectors $\boldsymbol{\mu}_k$ and that variance–covariance matrices Σ_k are nonsingular [37]. These two parameters are estimated from the training samples according to

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \mathbf{x}_{ik} \tag{5}$$

$$\hat{\Sigma}_k = \frac{1}{N_k - 1}\sum_{i=1}^{N_k} (\mathbf{x}_{ik} - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_{ik} - \hat{\boldsymbol{\mu}}_k)^{\mathrm{T}} \tag{6}$$

where $\mathbf{x}_k = \{x_{ik}, i = 1, \ldots, N_k\}$ are the training samples of the class k. This approach works well when the class-conditional densities are approximately normal and good estimates can be obtained from the training samples. However, it is highly affected by substantial divergence from normal density and by a limited training set [35], as it is often the case for hyperspectral remote sensing data.

In order to overcome problems linked to the aforementioned limitations, the parametric approach to discriminant analysis has been extended to the case where nothing is known about the densities f_k except for some assumptions about their general behavior [36]. The idea is to apply a nonparametric density estimator to the training samples and then to substitute the obtained estimates into the Bayes decision rule (2). This family of density estimators does not assume any prior knowledge about the distribution of the data like parametric estimators do. Many other nonparametric estimators can be found in the literature, such as the histogram approach, k-nearest neighbor, and the expansion by basis function method [38]. Owing to its properties in terms of computation and accuracy, one of the most common procedures is to use a multivariate kernel density estimator of the form

$$\hat{f}_k(\mathbf{x}) = \sum_{i=1}^{N_k} \mathcal{K}\{\mathbf{x} - \mathbf{x}_{ik}; \mathbf{H}_k\} \tag{7}$$

where $\mathcal{K}$ denotes an unbiased kernel function and $\mathbf{H}_k$ is a diagonal covariance matrix. It has been shown that the choice of the kernel used for the estimation is, in general, not crucial [38]. In our experiments, we have considered one of the most widely used kernels, the Gaussian one

$$\mathcal{K} = \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2h^2}\right\}. \tag{8}$$

Multidimensional density estimation is highly affected by the high dimensionality of the data and, practically, is not tractable when the dimension of the data is comparable with the size of the training data, as it often happens in hyperspectral data analysis. In these cases, the kernel $\mathcal{K}$ is substituted by the product of univariate Gaussian kernel function, leading to estimates of the form [39]

$$\hat{f}_k(\mathbf{x}) = (2\pi)^{-p/2}\mathcal{H}N_k^{-1}\sum_{l=1}^{N_k}\prod_{j=1}^{p}\exp\left\{-\frac{(x_j - x_{lkj})^2}{2h_{kj}^2}\right\} \tag{9}$$

where h_{kj} is the jth element in the $\mathbf{H}_k$ diagonal matrix, x_{ljk} is the ljth observation of the samples belonging to the class k, and

$$\mathcal{H} = \frac{1}{\prod_{j=1}^{p} h_{kj}}. \tag{10}$$

The main drawback of this approach is that some important information for the classification process is not retrieved. In fact, it particularly occurs when dealing with high-dimensional data where very important information for the classification process can be hidden in relatively low density regions. Therefore, the estimation of the tails of the distribution becomes crucial in order not to degrade the final results. Consequently, in such a case, a Gaussian kernel product estimator can be inappropriate, due to the short-tailed normal density.

In [34], Amato *et al.* proposed an interesting approach to circumvent the problems of nonparametric multivariate kernel density estimators and used ICA to enforce independence to the components of the analyzed data. In their approach, the components become as independent as possible after a transformation based on ICA, which allows one to estimate a multivariate density as the product of univariate densities, which is then fitted to normality with the use of normal densities. The results obtained are finally substituted in the Bayes rule for the class assignment.

The basic steps of the proposed approach are stated as follows.

1) Center the data on the k class, for each class $k = 1, \ldots, K$, and use the ICA to derive the optimal transform $\hat{\mathbf{A}}_k$ according to the training samples of the class.
2) Project the data using the computed transform and use an adaptive univariate kernel density estimator to estimate the density of each component.
3) For a new observation $\mathbf{x}$, the joint density of $\mathbf{Y} = \hat{\mathbf{A}}_k \mathbf{x}$ is first computed for each class as the product of the estimated marginal densities, since the components are independent. The density of $\mathbf{x}$ can then be derived from that of $\mathbf{Y}$ with a simple change of variable. The results are then substituted into the Bayes rule to obtain the final assignment.

In the rest of this paper, we refer to the aforementioned approach as ICDA.

A. ICA

ICA consists of finding a linear decomposition of observed data into statistically ICs. Given an observation model

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{11}$$

where $\mathbf{x}$ is the vector of the observed signals, $\mathbf{A}$ is a scalar matrix of the mixing coefficients, and $\mathbf{s}$ is the vector of the source signals, ICA finds a separating matrix $\mathbf{W}$ such that

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s} \tag{12}$$

where $\mathbf{y}$ is a vector of ICs. This means that the value of any of the components does not give any information about the value of the other components. ICA basically makes the following three general assumptions, in order to make sure that its basic model can be estimated.

1) The components of $\mathbf{y}$, estimated from the observed signal $\mathbf{x}$, are statistically independent. This is the basic principle on which all the ICA models rest.
2) At most one signal has a Gaussian distribution. If more than one component has a Gaussian distribution, we do not have enough information to separate mixtures of Gaussian sources. In the case of two or more Gaussian distributions, the higher order cumulants are equal to zero. This information is essential in order to estimate the ICA model; thus, the algorithm cannot work under these conditions.
3) The unknown mixing matrix $\mathbf{A}$ is square and invertible. This assumption is equivalent to saying that the number of ICs is equal to the number of observed mixtures. It is done in order to simplify the estimation very much, but it can sometimes be relaxed.

Under these three assumptions (or at least the first two), the ICs and the mixing matrix can be estimated under some indeterminacies that will necessarily hold. In fact, for (11), if both $\mathbf{A}$ and $\mathbf{s}$ are unknown, at least two ambiguities cannot be avoided. First, the variances of ICs cannot be computed. In fact, any scalar multiplier in one of the sources could always be canceled by dividing the corresponding column of the mixing matrix. Due to this, the energy of the components is, at first, fixed by whitening in order to make them all have variances equal to unity, and consequently, the mixing matrix is adapted. Second, because of similar reasons, the ICs cannot be ranked because any change in their order will not change the possibility to estimate the model. Due to its attractive properties, ICA is receiving a growing interest among the remote sensing community for feature reduction and spectral unmixing [52]–[54]. A detailed explanation of the basics of ICA is out of the scope of this paper. In the next section, we briefly review several possibilities to compute independence. We refer the reader interested in a complete explanation of the general framework of ICA to [49]–[51].

B. Independence Measures

Independence is a much stronger assumption than uncorrelatedness. Unlike common decorrelation methods, such as Principal Component Analysis and Factor Analysis, which use information provided by a covariance matrix in order to retrieve uncorrelated components, ICA considers higher (than second) order statistics. However, starting from the probabilistic definition of independence, several practical independence criteria can be defined. In addition to the basic concept of contrast functions, two of the most classical criteria are based on nonlinear decorrelation and maximum non-Gaussianity.

1) Nonlinear decorrelation. Find the matrix $\mathbf{W}$ so that the components $\mathbf{y}_i$ and $\mathbf{y}_j$ are uncorrelated and the transformed components $g(\mathbf{y}_i)$ and $h(\mathbf{y}_i)$ are uncorrelated, where g and h are some suitable nonlinear functions.

Possible nonlinear functions can be derived through the maximum likelihood or the mutual information.

2) Maximum non-Gaussianity. Find the local *maxima* of non-Gaussianity of a linear combination under the constraint that the variance of $\mathbf{y}$ is constant and the components not correlated (i.e., after prewhitening). Each local maximum gives one IC.

Classical algorithms, such as FastICA [40] and Infomax [41], have been developed using the aforementioned criteria. Another approach for the estimation of the ICs is joint diagonalization of eigenmatrices (JADE) [42], which makes use of fourth-order cumulant tensors. In the following experiments, we have used JADE as ICA algorithm to enforce independence, due to the effectiveness shown when dealing with hyperspectral remote sensing data [46], [47] and since it has provided better results than FastICA and Infomax in a preliminary test.

Cumulant tensors can be considered as the generalization of the covariance matrix at an order higher than the second. If we consider a random vector $\mathbf{x}$ with a probability density function $p(\mathbf{x})$, its characteristic function is defined as the inverse Fourier transform of the pdf [43]

$$\Phi(\omega) = E\left\{\exp(j\omega\mathbf{x})\right\} = \int_{-\infty}^{\infty} \exp(j\omega\mathbf{x})p(x)\,dx \tag{13}$$

where j is equal to $\sqrt{-1}$ and ω is the transformed row vector corresponding to $\mathbf{x}$. Every probability density function corresponds to a unique characteristic function and vice versa. Due to its attractive properties, the natural logarithm of the characteristic function is often considered. Given the Taylor series expansion of the characteristic function

$$\Phi(\omega) = \sum_{k=0}^{\infty} a_k \frac{(j\omega)^k}{k!} \tag{14}$$

the ath cumulant is defined as the derivative

$$a_k = (-j)^k \left.\frac{d^k\Phi(\omega)}{d\omega^k}\right|_{\omega=0}. \tag{15}$$

It can be shown that the second- and third-order cumulants for a zero mean random vector are [44]

$$\text{cum}(x_i, x_j) = E\{x_i x_j\} \tag{16}$$

$$\text{cum}(x_i, x_j, x_k) = E\{x_i x_j x_k\}. \tag{17}$$

We refer the reader interested in a more detailed explanation of the cumulants and their properties to [45] and [48]. The fourth-order cumulants contain all the information about the fourth-order statistics of the data. In the case the data are independent, all the cumulants with two or more different indices are equal to zero.

The cumulant tensor is a linear operator defined by the fourth-order cumulants $\text{cum}(x_i x_j x_k x_l)$ in an analog way to the case of a covariance matrix, which defines a linear operator. In this case, we have a linear transformation in the space of $n \times n$ matrices instead of the space of n-dimensional vectors. The space of such matrices is a linear space of dimension $n \times n$, so it is simple to define the linear transformation. The elements of this transformation can be defined as

$$\mathbf{F}_{ij}(\mathbf{M}) = \sum_{kl} m_{kl}\text{cum}(x_i, x_j, x_k, x_l) \tag{18}$$

where m_{kl} are the elements in the matrix $\mathbf{M}$ that is transformed.

JADE refers to one of the principles of solving the problem of equal eigenvalues of the cumulant tensor. As any symmetric linear operator, the cumulant tensor has an eigenvalue decomposition. An eigenmatrix of the tensor is, by definition, a matrix $\mathbf{M}$ such that

$$\mathbf{F}(\mathbf{M}) = \lambda\mathbf{M} \tag{19}$$

where λ is a scalar eigenvalue.

Let us consider data which follow the ICA model with whitened data

$$\mathbf{z} = \mathbf{VAs} = \mathbf{H^T s} \tag{20}$$

where $\mathbf{H^T}$ denotes the whitened mixing matrix. The eigenvalue decomposition allows us to point out some interesting features of the cumulant tensor of $\mathbf{z}$. Every matrix of the form

$$\mathbf{M} = \mathbf{h_m h_m^T} \tag{21}$$

is an eigenmatrix. The vector $\mathbf{h_m}$ represents here one of the rows of $\mathbf{H}$ and, thus, one of the columns of the whitened mixing matrix $\mathbf{H^T}$. Due to the independence of the sources, the corresponding eigenvalues are given by the kurtosis of the ICs and all the other eigenvalues are zero. By determining these eigenvalues, we can obtain the independent sources that we are looking for.

III. Experimental Results

A. Data Sets

In order to have a representation of the possible scenarios provided by the hyperspectral images as complete as possible (satellite/airborne sensors, urban/agricultural/geological area, and large/small/very small size of the training set), four hyperspectral data sets were considered in this paper.

The first one is an airborne data set from the ROSIS-03 with 115 spectral bands in the spectral range from 0.43 to 0.86 μm acquired over the University of Pavia, Italy. The spatial resolution is 1.3 m/pixel. The data set is 610 by 340 pixels. Twelve data channels were removed due to noise, and the remaining 103 spectral dimensions were processed. Nine classes of interest were considered. The training set is composed by about 10% of all the labelled samples that were available for the data.

The second data set is a small segment of an Airborne Visible InfraRed Imaging Spectrometer (AVIRIS) data set over the agricultural area of Indiana. For each spectral channel, the image contains 145 $\times$ 145 pixels. There are 220 spectral channels (spaced at about 10 nm) acquired in the 0.4–2.5 μm region. Four of the original 224 channels were removed, because they were containing only zeros. All the remaining 220 bands were

TABLE I
INFORMATION ABOUT THE TRAINING AND THE TESTING SET OF THE FOUR CONSIDERED DATA SETS

	ROSIS Data Set			AVIRIS Indian Pine			AVIRIS Hekla			HYDICE Washington		
No.	Name	Train	Test	Name	Train	Test	Name	Train	Test	Name	Train	Test
1	Asphalt	548	6641	Alfalfa	20	54	Andesite lava 1970	50	973	Roof	10	3794
2	Meadow	540	18649	Corn-no till	20	1434	And. lava 1980 I	50	634	Road	10	376
3	Gravel	392	2099	Corn-min till	20	834	And. lava 1980 II	50	408	Trail	10	135
4	Tree	524	3064	Corn	20	234	And. lava 1991 I	50	500	Grass	10	1888
5	Metal Sheet	265	1345	Grass-Pasture	20	497	And. lava 1991 II	50	1446	Tree	10	365
6	Bare Soil	532	5029	Grass-Trees	20	747	And. lava moss cover	50	650	Water	10	1184
7	Bitumen	375	1330	Grass-Mowed	20	26	Hyaloclastite formation	50	292	Shadow	10	57
8	Brick	514	3682	Hay-windrowed	20	489	Lava tephra covered	50	354	-	-	-
9	Shadow	231	947	Oats	20	20	Rhyolite	50	658	-	-	-
10	-	-	-	Soybean-no till	20	968	Scoria	50	663	-	-	-
11	-	-	-	Soybean-min till	20	2468	Firn-glacier ice	50	360	-	-	-
12	-	-	-	Soybean-clean t	20	614	Snow	50	268	-	-	-
13	-	-	-	Wheat	20	212	-	-	-	-	-	-
14	-	-	-	Woods	20	1294	-	-	-	-	-	-
15	-	-	-	Bldg-Trees-Drive	20	380	-	-	-	-	-	-
16	-	-	-	Stone-Steel Tower	20	95	-	-	-	-	-	-

processed, without discarding channels affected by atmospheric absorption. Sixteen reference data classes were considered. The ground truth is composed of 10 366 pixels. Different training sets were randomly constructed from the reference data with a total of, respectively, 320 pixels (20 samples per class). Due to the very small size of the training set, to increase the statistical significance of the test, the experiment was repeated ten times with different training sets and the average results reported.

The third study site is the region surrounding the central volcano Hekla in Iceland, one of the most active volcanoes in the country. Since 1970, Hekla has erupted quite regularly every ten years, in 1970, 1980–1981, 1991, and 2000. The volcano is located on the southwestern margin of the eastern volcanic zone in South Iceland. Hekla's products are mainly andesitic and basaltic lavas and tephra. AVIRIS data that were collected on a cloud-free day, June 17, 1991, were used for the classification. The image contains 560 $\times$ 600 pixels. As in the previous case, the sensor system has 224 data channels, utilizing four spectrometers, whereas the width of each spectral band is approximately equal to 10 nm [55]. During image acquisition, spectrometer 4, which operates in the wavelength range from 1.84 to 2.4 μm (64 bands), was not working properly. These 64 bands were deleted from the imagery along with the first channels for all the other spectrometers, so that only 157 data channels were considered. The training set was composed of 50 samples per class randomly chosen from the labeled data, and ten different training sets were selected.

Finally, airborne data from the Hyperspectral Digital Imagery Collection Experiment (HYDICE) sensor were used for the experiments. The HYDICE was used to collect data from flightline over the Washington DC Mall. Hyperspectral HYDICE data originally contained 210 bands in the 0.4–2.4-μm region. Noisy channels have been removed, and the set consists of 191 spectral channels. It was collected in August 1995, and each channel has 1280 lines with 307 pixels each. Seven information classes were defined. Also in this case, ten experiments were performed with the same training set size (ten samples per class). All the pieces of information about the ground truth (name of the classes, number of training, and testing samples) are listed in Table I.

B. *Influence and Choice of the Number of ICs*

The proposed method ICDA was used to classify the four data sets, and the results of the experiments were compared with those obtained by a one-versus-one SVM, with a Gaussian kernel and tenfold cross-validation selection of the kernel's parameter [56] applied to the full feature space. When applying ICDA, the number of components considered to compute the density estimation has an influence both on the final classification accuracy and on the computational burden. The maximum number of ICs that can be used for the classification depends on the rank of the covariance matrix obtained from the training samples of each class, and it is equal to the number of training samples of a class. In order not to bias the final assignment of the analyzed samples, the number of ICs computed in step 1 of the proposed algorithm should be the same for each class. Therefore, when the covariance matrix obtained from the training samples of a class is singular, that will decrease the maximum number of components which can be retained and it will influence all the classes. Because of the singularity of the covariance matrix of some classes, the maximum number of components which can be retrieved in the case of AVIRIS Indian Pine data set is 19, while it is 9 in the case of the HYDICE DC Mall. A larger number of components could be computed for the other two data sets. Fig. 1 shows the variation of the coefficient of agreement (Kappa), the average class accuracy, which represents the average of the classification accuracies for the individual classes and the processing time with respect to the ICs retained for the four considered data

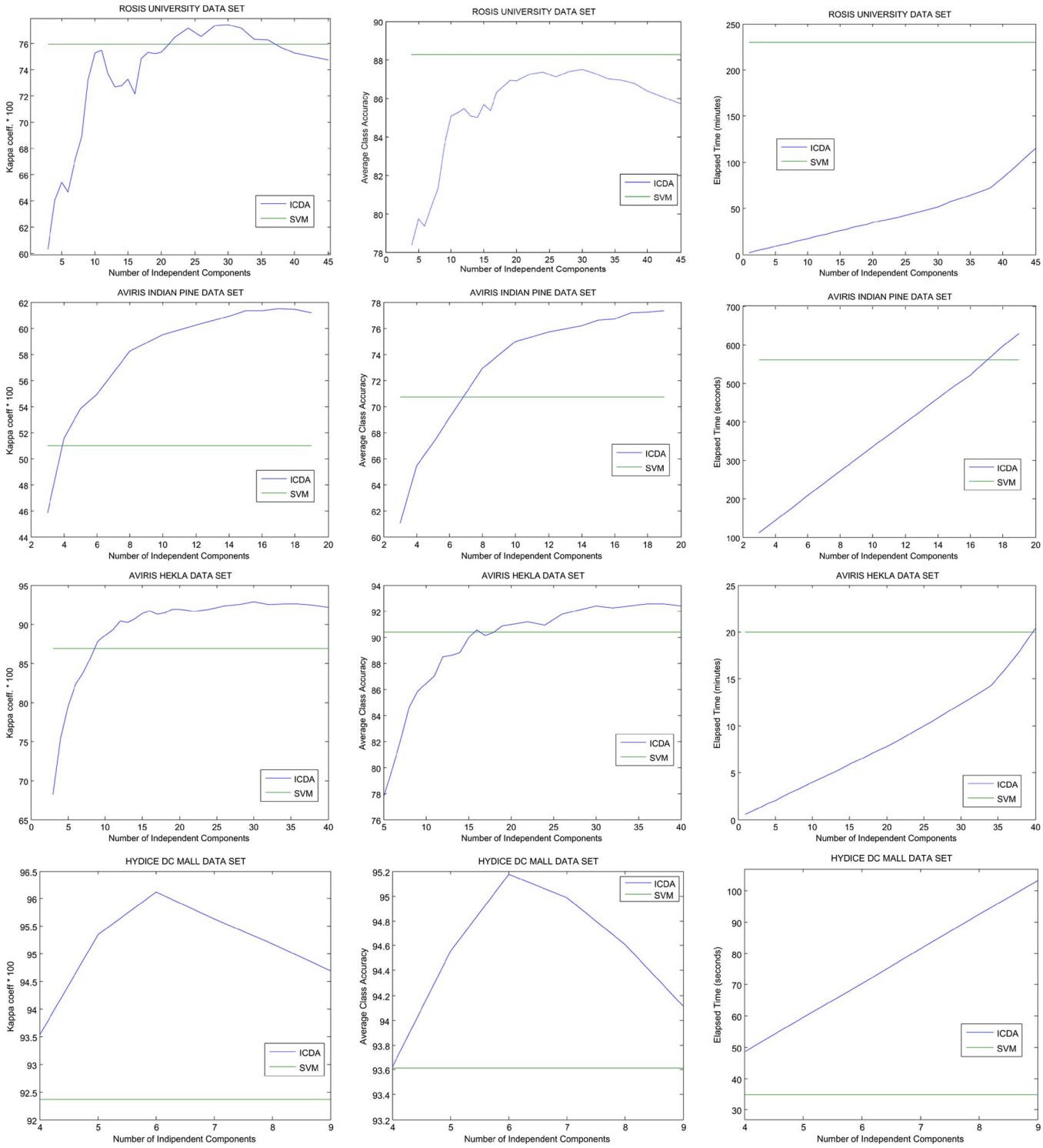


Fig. 1. Comparison of (first column) Kappa coefficient of agreement, (second column) average class accuracy, and (third column) classification processing time obtained with SVM (full feature space) and ICDA, with respect to different number of ICs retained for the four considered data sets. The rows respectively correspond to ROSIS, AVIRIS Indian Pine, AVIRIS Hekla, and HYDICE data sets.

sets. The Kappa coefficient of agreement is a parameter that estimates the correct percentage classification without considering the percentage accuracy that could be expected, performing a random classification [57].

Although the number of components has a large influence on the final results, it can be seen that there is a wide region where the proposed method outperforms SVM. In three among the four cases (AVIRIS Indian Pine, AVIRIS Hekla, and HYDICE DC Mall), once a minimum number of ICs is computed (in order to have enough information for the probability density estimation), the accuracy is much better than that of the SVM. In the case of the ROSIS data set, the trend is not so linear, but still, the choice of the number of ICs is not critical, i.e., we have a large range of values for which ICDA performs better

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY OBTAINED WITH THE PROPOSED METHOD ICDA (BEST CASE) AND SVM IN THE FOUR ANALYZED DATA SETS AND PROCESSING TIME OF THE TWO METHODS. WHERE SEVERAL TRAINING SETS HAVE BEEN SELECTED ON THE SAME DATA SET, THE STANDARD DEVIATION IS ALSO INDICATED

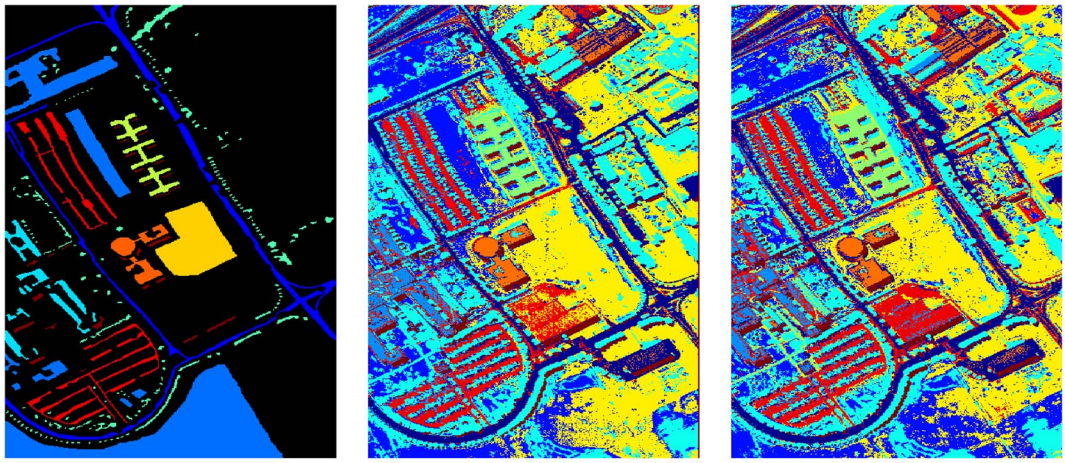
	ROSIS Data Set		AVIRIS Indian Pine		AVIRIS Hekla		HYDICE Washington	
Approach	SVM	ICDA	SVM	ICDA	SVM	ICDA	SVM	ICDA
OA	81.01%	**82.14%**	56.42 ± 1.45%	**64.98 ± 2.10%**	90.39 ± 1.17%	**94.22 ± 0.31%**	94.78 ± 1.00%	**97.23± 1.46%**
Kappa coef.	75.86%	**77.38%**	50.99 ± 1.35%	**60.80 ± 2.20%**	88.96 ± 1.31%	**93.32 ± 0.36%**	92.37 ± 1.44%	**95.95± 2.12%**
AA	**88.25%**	87.48%	70.10 ± 0.87%	**76.15 ± 1.02%**	90.37 ± 0.71%	**92.28 ± 0.50%**	93.62 ± 1.15%	**94.71± 1.93%**
Class 1	84.93%	76.40%	84.44%	83.33%	88.36%	96.92%	96.38%	97.79%
Class 2	70.79%	77.74%	36.79%	55.97%	87.25%	95.14%	99.16%	92.61%
Class 3	67.16%	77.42%	40.67%	51.61%	88.24%	94.19%	96.79%	94.48%
Class 4	97.77%	98.07%	72.31%	78.80%	84.94%	96.54%	98.41%	99.95%
Class 5	99.46%	100%	80.40%	84.87%	93.33%	86.11%	98.23%	98.15%
Class 6	92.83%	88.86%	78.93%	92.53%	94.24%	98.56%	81.85%	88.73%
Class 7	90.42%	91.35%	95.38%	96.15%	87.54%	96.06%	84.48%	90.69%
Class 8	92.78%	82.02%	76.11%	90.43%	91.69%	79.54%	-	-
Class 9	98.11%	95.35%	100%	100%	85.88%	87.29%	-	-
Class 10	-	-	53.80%	57.25%	74.20%	80.20%	-	-
Class 11	-	-	39.73%	49.78%	100%	100%	-	-
Class 12	-	-	49.12%	60.07%	97.59%	98.19%	-	-
Class 13	-	-	91.42%	99.15%	-	-	-	-
Class 14	-	-	81.31%	92.30%	-	-	-	-
Class 15	-	-	47.05%	51.89%	-	-	-	-
Class 16	-	-	94.11%	90.53%	-	-	-	-
Pr. time	240 m	53 m	580 s	420 s	20 m	12 m	36 s	66 s

than SVM. The difference of the behavior of the ROSIS data set with respect to the others has to be attributed to the way the training samples were collected. While in the other three cases the samples were randomly selected from the reference data, in this case, the training set was composed by spatially close zones, thus granting a worse capability of generalization.

Nevertheless, the choice of the appropriate number of ICs used during the classification process is a very important task in order to obtain a good accuracy of classification. In [34], the authors propose to retain the maximum possible number of ICs. This criterion is not appropriate for hyperspectral data: The computation of so many ICs can be very demanding from a computational viewpoint, and if the information provided by the components is redundant, the increase of ICs can lead to a decrease of classification accuracy, as pointed out in the first two columns of Fig. 1. In order to choose the number of ICs which has to be retained by the algorithm, we propose a simple but effective method. We apply the ICDA to the training set, using the same samples for testing. Since we just have to choose the number of ICs which better characterize the classes of the image, we do not expect problems of the generalization of the results, as could appear when selecting the kernel's parameter of SVM. The cross-validation approach has been discarded because of two reasons: 1) The very limited number of training samples of some data sets can lead to misleading results, and 2) splitting the number of samples in the cross-validation procedure influences the maximum number of components which can be retained. Since preliminary experiments have shown that the smallest and biggest values of ICs are not useful, because they do not contain enough information or they have redundant features, to avoid a large computational time, the range investigated was [10–30] in the case of ROSIS and AVIRIS Hekla data sets, [10–19] for the AVIRIS Indian Pine, and [3–9] for the HYDICE data set. This way, a finer step can be used, avoiding too much computational effort.

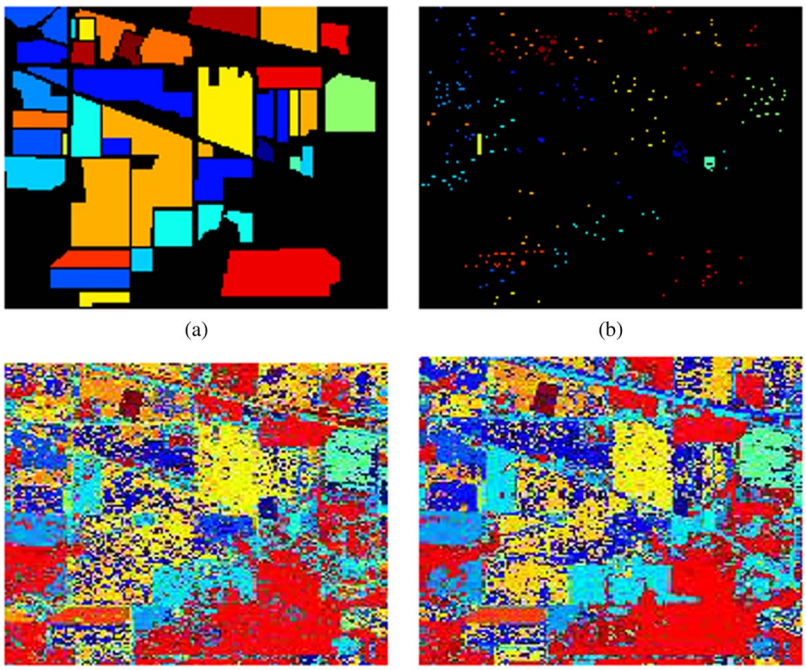
C. Performance Analysis

Table II presents a comparison between the results obtained by the SVM (applied to the full feature space) and the ICDA (with the proposed method to choose the number of ICs). The comparison is in terms of overall accuracy (OA), which is the number of correctly classified test samples with respect to the total number of test samples, average accuracy (AA), and the Kappa coefficient of agreement (κ). In all the considered data sets, the Kappa coefficient of agreement provided by the ICDA is better than the corresponding result for SVM. The ROSIS data set gives the only case where the average class accuracy of SVM is higher than ICDA. In the three experiments where multiple training sets were considered, the standard deviation was also computed. In two cases, the standard deviation obtained with the SVM was smaller than that for the ICDA.



(a) (b) (c)

Fig. 2. (a) Ground truth of the ROSIS data set. (b) Classification map obtained with the SVM. (c) Classification map obtained with the ICDA.



(a) (b) (c) (d)

Fig. 3. (a) Ground truth of the AVIRIS Indian Pine data set. (b) Training set n.3. (c) Classification map obtained with the SVM. (d) Classification map obtained with the ICDA.

That happened for the AVIRIS Indian Pine and the HYDICE data sets, where very small training sets were selected. The best results were obtained with a number of ICs retained, which was varying for the different training sets. Due to the small number of training samples, large variations were seen according to the number of ICs retained, also for small differences, thus leading to a higher value of standard deviation. This phenomenon was less important for the AVIRIS Hekla data set due to the larger number of training samples selected. As an example, Figs. 2 and 3 show some classification maps obtained with the SVM and with the proposed method. It has been shown in [58] that the comparison of two classification results in terms of OA may be inappropriate, being explicitly based on an assumption that the two sets considered are independent. This is not true in our experiments, where the samples used for the training and testing processes of the two different classifications are exactly the same. In order to better evaluate the statistical significance of the difference between ICDA and SVM, we performed McNemar's test, which is based upon the standardized normal test statistic

$$Z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} - f_{21}}} \tag{22}$$

where f_{12} indicates the number of samples classified correctly by classifier 1 and incorrectly by classifier 2. The difference in accuracy between classifiers 1 and 2 is said to be statistically significant if $|Z| > 1.96$. The sign of Z indicates whether classifier 1 is more accurate than classifier 2 ($Z > 0$) or vice versa ($Z < 0$). This test assumes related testing samples and, thus,

TABLE III
STATISTICAL SIGNIFICANCE OF THE DIFFERENCES BETWEEN THE TWO CLASSIFIERS. THE FIRST THREE COLUMNS REPRESENT HOW MANY TIMES SVM WAS PERFORMING SIGNIFICANTLY BETTER, THERE WERE NO STATISTICAL DIFFERENCES, AND ICDA WAS PERFORMING BETTER. DIFFERENCES ARE CONSIDERED SIGNIFICANT IF $|Z| > 1.96$

SVM better	No stat. differences	ICDA better	Mean Z
ROSIS data set			
0	0	1	-7.66
AVIRIS Indina Pine data set			
0	0	10	-13.20
AVIRIS Hekla data set			
0	0	10	-10.83
HYDICE data set			
0	1	9	-8.59

is adapted to our situation since the training and testing sets were the same for each experiment. The results of McNemar's test [58] are shown in Table III and confirm the conclusions of previous experiments.

Finally, the computational burdens of SVM and ICDA were investigated. The processing time of SVM quadratically depends on the size of the training set, and it is longer where a large number of training samples is used. In opposite, ICDA has a very short training time, due to the fast computation of density estimations of the training samples, and a longer testing time, because these densities have to be calculated for each of the testing samples. Fig. 1 shows in the third column how the processing time of ICDA varies according to the number of IC retained. As could be expected, SVM is computationally less demanding than the ICDA when considering data sets with small or very small number of training samples. The opposite situation occurs when medium/large training sets are available, as in the case of ROSIS or AVIRIS Hekla data sets.

IV. CONCLUSION

In this paper, a new approach for hyperspectral data classification, the ICDA, has been proposed. The proposed approach is based on the application of ICA to the data in order to retrieve ICs, the use of a kernel density estimate to obtain reliable estimation of class-dependent densities, and the substitution on the Bayes rule for the final assignment. Experiments have been carried out on four different real data sets. The results of the experiments showed the effectiveness of the proposed method, which provided better results than those of the state-of-the-art hyperspectral classifier, the SVM. Moreover, the proposed method presents several other advantages: 1) Its Bayesian nature allows the integration of any kind of prior information in the classification process, as long as they can be stated as a probability function, and 2) it is suitable to be used jointly with spectral–spatial techniques recently developed for SVM [32], [33].

Although the classification accuracy obtained by the ICDA is influenced by the number of components retained after applying ICA, this choice is not critical, since there is a large region around the optimal number for such accuracy for which the proposed method has similar results and outperforms SVM in terms of classification accuracy. Moreover, a simple and effective technique for choosing the number of components to retain was proposed, providing results significantly better than those of the SVM. The computational burden of the proposed method is smaller with respect of the SVM when a medium/large amount of training samples is available. The SVM is computationally less demanding for small training sets, but in such cases, time is not a critical issue. Further developments of this work include a comprehensive research of the influence of the ICA algorithm used to enforce independence and an investigation of the possibility of including contextual spatial information within the Bayesian framework.

ACKNOWLEDGMENT

The authors would like to thank Prof. A. Antoniadis, IMAG Université J. Fourier, for providing the ICDA code and Prof. P. Gamba, University of Pavia, for providing the ROSIS data set.

REFERENCES

[1] D. A. Landgrebe, S. B. Serpico, M. M. Crawford, and V. Singhroy, "Introduction to the special issue on analysis of hyperspectral image data," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1343–1345, Jul. 2001.
[2] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. New York: Wiley, 2003.
[3] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.
[4] J. Besag, "Toward Bayesian image analysis," *J. Appl. Statist.*, vol. 16, no. 3, pp. 395–407, 1989.
[5] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
[6] 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.
[7] P. H. Swain and S. M. Davis, *Remote Sensing: The Quantitative Approach*. New York: McGraw-Hill, 1978.
[8] 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.
[9] S. Tadjudin and D. A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 4, pp. 2113–2118, Jul. 1999.
[10] Q. Jackson and D. A. Landgrebe, "An adaptive classifier design for high dimensional data analysis with a limited training data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 12, pp. 2664–2679, Dec. 2001.
[11] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 4, pp. 388–400, Apr. 1993.
[12] L. O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and feature reduction via projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2653–2667, Nov. 1999.
[13] B. C. Kuo and D. A. Landgrebe, "A robust classification procedure based on mixture classifiers and nonparametric weighted feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2486–2494, Nov. 2002.
[14] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Prentice-Hall, 1999.
[15] J. A. Benediktsson, J. R. Sveinsson, and K. Arnason, "Classification and feature extraction of AVIRIS data," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 5, pp. 1194–1205, Sep. 1995.
[16] A. M. Filippi and J. R. Jensen, "Effect of continuum removal on hyperspectral coastal vegetation classification using a fuzzy learning vector quantizer," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1857–1869, Jun. 2007.
[17] 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.

[18] G. M. Foody and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 6, pp. 1335–1343, Jun. 2004.
[19] G. Camps-Valls, L. Gómez-Chova, J. Calpe-Maravilla, E. Soria-Olivas, J. D. Martín-Guerrero, L. Alonso-Chord, 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.
[20] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
[21] B. Wohlberg, D. M. Tartakovsky, and A. Guadagnini, "Subsurface characterization with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 1, pp. 47–57, Jan. 2006.
[22] L. Zhang, X. Huang, B. Huang, and P. Li, "A pixel shape index coupled with spectral information for classification of high spatial resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2950–2961, Oct. 2006.
[23] C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[24] N. Ghoggali, F. Melgani, and Y. Bazi, "A multiobjective genetic SVM approach for classification problems with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 6, pp. 1707–1718, Jun. 2009.
[25] A. M. Filippi and R. Archibald, "Support vector machine-based endmember extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 771–791, Mar. 2009.
[26] S. Mika, G. Rätsch, B. Schölkopf, A. Smola, J. Weston, and K. R. Müller, "Invariant feature extraction and classification in kernel spaces," in *Advances in Neural Information Processing Systems*, vol. 12. Cambridge, MA: MIT Press, 1999.
[27] A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik, "Support vector clustering," *Mach. Learn. Res.*, vol. 2, pp. 125–137, 2001.
[28] G. Rätsch, B. Schölkopf, A. Smola, S. Mika, T. Onoda, and K.-R. Müller, "Robust ensemble learning," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999, pp. 207–219.
[29] Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3374–3385, Nov. 2006.
[30] B. Waske and J. A. Benediktsson, "Fusion of support vector machines for classification of multisensor data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3858–3866, Dec. 2007.
[31] Y. Bazi and F. Melgani, "Gaussian process approach to remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 186–197, Jan. 2010.
[32] M. Fauvel, J. Chanussot, J. A. Benediktsson, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 3804–3814, Oct. 2008.
[33] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.
[34] U. Amato, A. Antoniadis, and G. Gregoire, "Independent component discriminant analysis," *Int. Math. J.*, vol. 3, no. 7, pp. 735–753, 2003.
[35] J. Friedman, "Exploratory projection pursuit," *J. Amer. Statist. Assoc.*, vol. 82, no. 397, pp. 249–266, Mar. 1987.
[36] E. Fix and J. L. Hodges, "Discriminatory analysis—Nonparametric discrimination: Consistency properties," *Int. Stat. Rev.*, vol. 57, no. 3, pp. 238–247, Dec. 1989.
[37] B. Efrom and C. Morris, "Families of minimax estimators of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 4, no. 1, pp. 11–21, Jan. 1976.
[38] A. Webb, *Statistical Pattern Recognition*, 2nd ed. New York: Wiley, 2002.
[39] D. W. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: Wiley, 1992.
[40] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.
[41] A. J. Bell and T. J. Sejnowski, "A non linear information maximization algorithm that performs blind separation," in *Advances in Neural Information Processing Systems, 7*. Cambridge, MA: MIT Press, 1995, pp. 467–474.
[42] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *Proc. Inst. Elect. Eng.—Radar Signal Process.*, vol. 140, no. 6, pp. 362–370, Dec. 1993.
[43] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 3rd ed. New York: McGraw-Hill, 1991.
[44] M. Girolami, *Self-Organising Neural Networks—Independent Component Analysis and Blind Source Separation*. New York: Springer-Verlag, 1999.
[45] C. Nikias and A. Petropulu, *Higher-Order Spectral Analysis—A Nonlinear Signal Processing Framework*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[46] S. Moussaoui, H. Hauksdottir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Doute, and J. A. Benediktsson, "On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation," *Neurocomputing*, vol. 71, no. 10–12, pp. 2194–2208, Jun. 2008.
[47] A. Villa, J. Chanussot, C. Jutten, J. A. Benediktsson, and S. Moussaoui, "On the use of ICA for hyperspectral image analysis," in *Proc. IEEE IGARSS*, 2009, pp. IV-97–IV-100.
[48] M. Rosenblatt, *Stationary Sequences and Random Fields*. Boston, MA: Birkhäuser, 1985.
[49] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.
[50] T. W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski, "An unifying information-theoretic framework for Independent Component Analysis," *Comput. Math. Appl.*, vol. 39, no. 11, pp. 1–21, Jun. 2000.
[51] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation. Independent Component Analysis and Applications*. New York: Academic, 2010.
[52] J. M. P. Nascimento and J. M. Bioucas Dias, "Does independent component analysis play a role in unmixing hyperspectral data?" *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 1, pp. 175–187, Jan. 2005.
[53] J. Wang and C. I. Chang, "Independent Component Analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, Jun. 2006.
[54] P. Birjandi and M. Dactu, "Multiscale and dimensionality behavior of ICA components for satellite image indexing," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 103–107, Jan. 2010.
[55] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis, M. R. Olah, and O. Williams, "Imaging spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," *Remote Sensing of Environment*, vol. 65, no. 3, pp. 227–248, Sep. 1998.
[56] C. C. Chang and C. J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2007. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[57] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, pp. 37–46, 1960.
[58] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–633, May 2004.

Alberto Villa (S'09) received the B.S. and M.S. degrees in electronic engineering from the University of Pavia, Pavia, Italy, in 2005 and 2008, respectively. He has been working toward the Ph.D. degree (a joint degree) from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, and the University of Iceland, Reykjavik, Iceland, since 2008.

He is a Reviewer for the *Journal of Signal Processing Systems*. His research interests are in the areas of spectral unmixing, machine learning, hyperspectral imaging, signal and image processing.

He is a Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING.

Jón Atli Benediktsson (S'84–M'90–SM'99–F'04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984 and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, in 1987 and 1990, respectively.

He is currently the Pro Rector for Academic Affairs and a Professor of Electrical and Computer Engineering with the University of Iceland. His research interests are in remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has published extensively in those fields.

Dr. Benediktsson is the 2011 President of the IEEE Geoscience and Remote Sensing Society (GRSS) but he has been on the GRSS AdCom since 2000. He was Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING from 2003 to 2008. He was the recipient of the Stevan J. Kristof Award from Purdue University in 1991 as outstanding graduate student in remote sensing. In 1997, he was the recipient of the Icelandic Research Council's Outstanding Young Researcher Award. In 2000, he was granted the IEEE Third Millennium Medal. In 2004, he was a corecipient of the University of Iceland's Technology Innovation Award. In 2006, he was the recipient of the yearly research award from the Engineering Research Institute of the University of Iceland, and in 2007, he was the recipient of the Outstanding Service Award from the IEEE GRSS. He is a member of Societas Scinetiarum Islandica and Tau Beta Pi.

Jocelyn Chanussot (M'04–SM'04) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995 and the Ph.D. degree from the University of Savoie, Annecy, France, in 1998.

In 1999, he was with the Geography Imagery Perception Laboratory, Délégation Geńérale pour l'Armement (French National Defense Department). Since 1999, he has been with Grenoble INP, where he was an Assistant Professor from 1999 to 2005, was an Associate Professor from 2005 to 2007, and is currently a Professor of signal and image processing. He is currently conducting his research with the Grenoble Image Speech Signals and Automatics Laboratory, (GIPSA-Lab, affiliated with CNRS and Grenoble INP). His research interests include image analysis, multicomponent image processing, nonlinear filtering, and data fusion in remote sensing. He was an Associate Editor for *Pattern Recognition* (2006-2008).

Dr. Chanussot was an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (2005-2007). He has been an Associate Editor for theIEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING since 2007. He is currently the Editor in Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2011-2013). He is the founding President of the IEEE Geoscience and Remote Sensing Society (GRSS) French Chapter (2007), which was the recipient of the 2010 IEEE GRSS Chapter Excellence Award for excellence as a GRSS chapter demonstrated by exemplary activities during 2009. He is a member of the IEEE Geoscience and Remote Sensing AdCom (2009–2011). He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing—Evolution in Remote Sensing (2008) and the technical cochair of the following editions. He is the Chair (2009–2011) and was the Cochair (2005–2008) of the GRS-S Data Fusion Technical Committee. He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006–2008) and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009).

Christian Jutten (M'03–SM'06–F'08) received the Ph.D. and Docteur des Sciences degrees from the Grenoble Institute of Technology, Grenoble, France, in 1981 and 1987.

After being an Associate Professor with the Electrical Engineering Department, Grenoble Institute of Technology (1982-1989), and a Visiting Professor at the Swiss Federal Polytechnic Institute, Lausanne, Switzerland, in 1989, he became a Full Professor with the University Joseph Fourier of Grenoble, Grenoble, more precisely in the sciences and technologies department. For 30 years, his research interests are learning in neural networks, blind source separation, and independent component analysis, including theoretical aspects (separability and source separation in nonlinear mixtures) and applications (biomedical, seismic, speech, astrophysics, and chemical sensor array). He is the author or coauthor of 4 books, 18 invited papers, and more than 60 papers in international journals and 160 communications in international conferences. He has been a Scientific Advisor for signal and image processing at the French Ministry of Research (1996–1998) and for the French National Research Center (CNRS, 2003–2006).

Dr. Jutten is a member of the technical committee "Blind Signal Processing" of the IEEE Circuits and Systems Society and of the technical committee "Machine Learning for Signal Processing" of the IEEE Signal Processing Society. He has been the Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS (1994-1995). He was Co-organizer of the 1st International Conference on Blind Signal Separation and Independent Component Analysis (Aussois, France, January 1999). For his contributions in source separation and independent component analysis, he received the EURASIP Best Paper Award in 1992, the Medal Blondel in 1997 from SEE (French Electrical Engineering Society) and, in 2008, has been elevated as a senior member of the Institut Universitaire de France.