Advanced spectral unmixing and classification methods for hyperspectral remote sensing data

Alberto Villa

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Electronique, Electrotechnique, Automatique & Traitement du Signal**

Arrêté ministériel : 7 août 2006

Et de

DOCTOR OF UNIVERSITY OF ICELAND

Spécialité : **Electrical and Computer Engineering**

Présentée par

Alberto VILLA

Thèse dirigée par **Christian Jutten** et
codirigée par **Jocelyn Chanussot** et **Jon Atli Benediktsson**

préparée au sein du **GIPSA Lab**
dans **Signal Image Parole Telecoms**
et de **Department of Electrical and Computer Engineering, University of Iceland**

Advanced spectral unmixing and classification methods for hyperspectral remote sensing data

Thèse soutenue publiquement le **29 Juillet 2011** devant le jury composé de :
Mme. Melba CRAWFORD
Professeur, Purdue University. Président
M. Jean-Yves TOURNERET
Professeur, INP-ENSEEIHT Toulouse. Rapporteur.
M. Paul SCHEUNDERS
Professeur, University of Antwerp. Rapporteur.
M. Antonio PLAZA
Professeur, University of Extremadura, Membre.
M. José BIOUCAS-DIAS
Professeur, Instituto Tecnico Superior, Membre.
M. Johannes SVEINSSON
Professeur, University of Iceland, Membre.
M. Christian JUTTEN
Professeur, Université Joseph Fourier, Membre.
M. Jocelyn CHANUSSOT
Professeur, Grenoble INP. Membre
M. Jon Atli BENEDIKTSSON
Professeur, University of Iceland.

Contents

List of Figures

List of Tables

Résumé de thèse

La thèse propose des nouvelles techniques pour la classification et le démelange spectral des images obtenus par télédétection iperspectrale. Les problèmes liées au données (notamment très grande dimensionalité, présence de mélanges des pixels) ont été considerés et des techniques innovantes pour résoudre ces problèmes. Nouvelles méthodes de classification avancées basées sur l'utilisation des méthodes traditionnel de réduction des dimension et l'integration de l'information spatiale ont été développés. De plus, les méthodes de démelange spectral ont été utilisés conjointement pour ameliorer la classification obtenu avec les méthodes traditionnel, donnant la possibilité d'obtenir aussi une amélioration de la résolution spatial des maps de classification grâce à l'utilisation de l'information à niveau sous-pixel.
Les travaux ont suivi une progression logique, avec les étapes suivantes:

1. Constat de base: pour améliorer la classification d'imagerie hyperspectrale, il faut considérer les problèmes liées au données : très grande dimensionalité, presence de mélanges des pixels.

2. Peut-on développer méthodes de classification avancées basées sur l'utilisation des méthodes traditionnel de réduction des dimension (ICA ou autre)?

3. Comment utiliser les different types d'information contextuel typique des imagés satellitaires ?

4. Peut-on utiliser l'information données par les méthodes de démelange spectral pour proposer nouvelles chaines de réduction des dimension?

5. Est-ce qu'on peut utiliser conjointement les méthodes de démelange spectral pour ameliorer la classification obtenu avec les méthodes traditionnel?

6. Peut-on obtenir une amélioration de la résolution spatial des maps de classification grâce à l'utilisation de l'information à niveau sous-pixel?

Les différents méthodes proposées ont été testées sur plusieurs jeux de données réelles, montrant resultats comparable ou meilleurs de la plus part des methodes presentés dans la litterature.

Introduction

The work presented in this thesis deals with the classification of hyperspectral remote sensing images. The main objective of the thesis is to propose advanced techniques able to exploit the different kinds of information contained in a hyperspectral image, in order to improve the classification of a scene, both from a quantitative point of view, by increasing the performances provided by recently proposed techniques, and qualitative point of view, obtaining thematic land cover maps at a finer resolution with respect to the analyzed scene. Earth observation optical remote sensing involves the acquisition of information about the surface of our planet without coming into physical contact with it. This is achieved trough airborne and satellite sensors exploiting the properties of the observed materials, which absorb, reflect and emit electromagnetic radiation according to their intrinsic structure. The main information which can be gathered from a remote scene can be divided into three types:

- Spectral information
- Spatial information
- Temporal information.

In case of image classification, that is the creation of thematic maps representing the land coverage of the observed scene, the most important information is represented by spectral and spatial data, while the temporal information is mainly used when change detection between different acquisitions of the same area must be performed.
Recent advances in imaging spectroscopy, motivated by a desire to extract increasingly detailed information about the material properties of pixels in a scene, made possible the development of advanced sensors which can collect a detailed spectral signature of a scene covering a wide range of wavelengths [1]. Hyperspectral sensors can simultaneously collect more than a hundred spectral bands of an area with a very detailed spectral resolution (typically in the order of few nanometers), and varying spatial resolution. A hyperspectral image can be viewed as an image cube where the third dimension is the spectral domain represented by hundreds of narrow, contiguous spectral bands corresponding to the spectral reflectance [2]. The wealth of information that resides in the spectral domain provides significant enhancements relative to traditional panchromatic and multispectral imagery. The ability of imaging sensors to acquire the reflectance spectrum of a pixel in a significant detail, leads to substantial differences in the reflectance values of the pixels belonging to disparate materials on the Earth's surface. For example, there is the potential to classify scene elements with subtle material differences as well as detect low contrast targets in complex background clutter [3]. However, when considering hyperspectral images for supervised classification, several drawbacks should also be taken into account. The first and most important limitation, which limits the possibility to apply traditional techniques used for the classification of other data types is the so-called curse of dimensionality [4]. The analysis of the huge amount of information contained in hyperspectral images poses some problems both from a theoretical and practical point of view. From the theoretical viewpoint, there is the appearance of the Hughes' phenomenon: Given a certain number of training samples and a traditional supervised classifier, if the number of spectral dimensions of the image are increased over a certain threshold, the classification accuracy will inevitably decrease [4]. From the practical viewpoint, the analysis of this huge amount of information requires a tremendous computational effort.
The second major drawback of hyperspectral images is the trade-off between spectral and spatial resolution, which lead, in the case of satellite sensors, to a spatial resolution of tens of meters [5]. The energy field arising from the Earth is, of course, finite in magnitude. The data collection process must divide this finite quantity spatially into pixels. The power level in each pixel can be divided up into a number of spectral bands. Given the finite nature of the field, there is then a trade-off between these two, because as one moves to finer spatial

resolution and spectral band intervals, less power is left to overcome the internal noise of the sensor system. Thus a less precise measure can be made of the signal level arriving from the surface. In the case of supervised classification, the low spatial resolution can lead to the appearance of mixed pixels, which are pixels of the images occupied by more than a single land cover class. These pixels cannot be classified without neglecting a (likely to be) important part of the information provided by the spectral signature. The problem is still more important since the new generation of satellite carrying hyperspectral sensors (like EnMap and PRISMA) have a spatial resolution of about 30 m, making this problem an important issue to analyze in order to provide efficient methods for the classification of acquired images.

The traditional approach to hyperspectral remote sensing classification was presented in [2], and can be summed up as the application of feature selection / extraction technique, in order to decrease the dimensionality if the data considering only informative features, that will be used as input for the supervised classifier. Several dimensionality reduction techniques were proposed over the last years, both supervised and unsupervised. The supervised techniques use the information provided by the ground truth training samples to estimate the most suitable features to separate the classes, while unsupervised do not rely on such information. Standard unsupervised techniques are Principal Component Analysis (PCA), Independent Component Analysis (ICA) [6]; supervised ones include Discriminant Analysis Feature Extraction (DAFE), Decision Boundary Feature Extraction (DBFE), and Non-parametric Weighted Feature Extraction (NWFE) [7,8]. It is preferable to use supervised transformations because the projection into the subspace minimizes a classification error criterion. However, the performance of such methods is closely related to the quality of the reference or training set. Consequently, unsupervised methods are of interest, but since the minimization criterion is not related to the classification error, the projection is not optimal for the purpose of classification. However, the aim of feature-reduction algorithms is not necessarily classification, but also representation. Since PCA considers only second order statistics, it may neglect important information which is not significant in terms of variance. Some variations of the PCA algorithm were more recently proposed to overcome such problems [9–12], but the analysis of methods exploiting higher order statistics could be an interesting alternative.
A possible modification of this framework is the inclusion of contextual spatial information: the algorithm considered for classification takes into account not only the spectral information of the pixel, but also the spatial context in which the pixels is comprised. Markov Random Fields (MRFs) are usually used within a statistical framework [13]. A survey of the current techniques for the analysis of remotely-sensed data can be found in [14]. The use of contextual information can be very useful, especially in case of high spatial resolution remote sensing data. For hyperspectral imaging, high spatial resolutions are usually provided by airborne sensors, as the ROSIS sensor. High spatial resolution data contain a lot of contextual information: for a given pixel we can extract the size, shape, and gray level distribution of the structure to which it belongs [15]. This information will not be the same if the pixel belongs to a roof or to a green area. This is also a way to discriminate various structures made of the same materials. If spectral information alone is used, the roofs of a private house and of a larger building will be different.
Feature reduction techniques and the investigation of contextual information has recently been exhaustively investigated. However, the large number of works recently presented show that there is still room for a significant improvement in these fields. In particular, the study of dimensionality reduction techniques able to provide data related to the physical nature of the data is a very important topic.
Moreover, recent advances in mathematical morphology applied to remote sensing data open

new possibility for incorporating contextual information within the supervised classification framework [15].

An important issue related to hyperspectral imaging is the low spatial resolution [5]. While the curse of dimensionality and the use of contextual information have been extensively studied in the last years, the problems related to mixtures of classes have only been little touched in the literature concerning supervised classification. However, widely used data sets that can be considered as benchmarks in the remote sensing community, like the AVIRIS Indian Pines data set, are known to be dominated by mixtures of classes rather than pure pixels representing a single land cover type. A large number of mixed pixel techniques, providing information about a scene at a sub-pixel level, were proposed in the literature. The aim of such techniques is to provide a series of maps of abundance about the different land cover classes. The possibility to incorporate these techniques into a supervised classification framework remains however an almost unexplored topic. Previous efforts in this direction were presented in [16, 17], and show the interest of using jointly full pixel and mixed pixel techniques for classification purposes, but an exhaustive investigation of the beneficial effect provided by this additional source of information is required.

In summary, hyperspectral data can be seen as a very important instrument to obtain accurate land cover maps, thanks to the great amount of spectral information which is helpful for distinguishing very similar classes. However, in order to properly exploit all the different types of information contained in such data, several challenges should be taken into account:

- The use of feature reduction techniques to avoid problems related to the unfavorable ratio between number of training samples and dimensionality of the data.
- The possibility to include proper contextual information to improve the classification results, in terms of accuracy.
- The exploitation of sub-pixel information, which can be helpful in case of low spatial resolution and mixtures of classes, as is often the case with this kind of data.

In this thesis, advanced methodologies for the analysis and classification of hyperspectral remote sensing data are proposed. Satellite images contain different sources of information: The large number of spectral bands provides a detailed spectral description of each pixel acquired by the sensor; the spatial structures observed in a scene make it possible to exploit contextual information highly suitable for improving the accuracy of thematic maps. Moreover, sub-pixel information obtained by performing spectral unmixing is another important source of data which can be helpful for accurate classification of pixels, especially in case of moderate resolution sensors. The different experiments carried out in this work try to make a complete analysis of the possibility to incorporate all these different types of information within the classification process.
Particular attention is paid to the possibility of exploiting sub-pixel information for improving the quality of classification maps, due to the little investigation done in the literature about on topic. An important point of this work, is the attempt to obtain results which could be as general as possible, thus not being scenario-dependent. In order to achieve such a result, a large number of hyperspectral data sets are analyzed, provided by four different sensors (three airborne sensors and one satellite sensor), covering urban areas, agricultural zones, geological scenarios. A total of eight different hyperspectral scenes were analyzed in the different experiments, in order to assess the effectiveness of the approaches proposed.
The work is divided into three main parts and seven chapters:

I) The first part of the thesis is devoted to the investigation of classical feature reduction - extraction techniques for the improvement of the classification accuracy. Particular

attention is paid to the use of Independent Component Analysis (ICA), due to its attractive properties. This part is divided into three chapters.

1. A general introduction of spectral feature reduction and spatial feature extraction is given in Chapter 1. The reasons for making these approaches suitable for remote sensing images are explained. Independent Components Analysis is described in detail as a technique which could be possibly used as an alternative to PCA. The objective is to reduce the dimension of the data without any ground truth for classification, using conventional classifiers that are adversely affected by dimensionality. The use of Morphological Profiles is also introduced, as a possibility to extract spatial information techniques from the image with a sensor-independent approach.
2. Independent Component Discriminant Analysis (ICDA) is presented in Chapter 2 as a classifier suited to the problem of classification of remote-sensing data. This approach exploits the characteristics of ICA, which allows to obtain independent components estimated from the spectral feature of a hyperspectral data set, to obtain a state of the art Bayes classifier. The approach is tested by using four hyperspectral data sets and the results are compared with those obtained by the SVM, one of the most widely used approaches for remote sensing classification. The comparison shows the comparative effectiveness of the proposed approach.
3. The use of spatial information to better describe structures in a urban environment is investigated in Chapter 3. The possibility to exploit several types of Morphological Attribute Profiles in combination with the use of ICA is investigated, and two different types of multi-attribute classifiers are proposed. These classifiers provide a fusion of the data and of the classifier output to improve the overall classification accuracy. The proposed approach was tested on two very high resolution images, acquired over a urban environment.

II) The second part of this work investigates the use of sub-pixel information to provide a better description of the hyperspectral scene containing mixed pixels. The use of sub-pixel information is particularly useful when images containing mixed pixels, since these images are hard to be represented with traditional techniques. Two chapters compose this part:

4. When analyzing a moderate resolution image (for example, images obtained with the AVIRIS satellite sensor, which has a 20 m/pixel resolution), several mixture of classes can be found in the data. Techniques based on spectral unmixing, which could provide information at a sub-pixel level, are likely to better represent this kind of images, with respect to traditional techniques. Chapter 4 presents a number of dimensionality reduction chains based on the use of endmember extraction algorithms combined with spectral unmixing techniques. The results obtained are compared with those provided by traditional techniques such as PCA, ICA and MNF.
5. When performing dimensionality reduction of hyperspectral images before classification, the main differences in terms of accuracy are provided by the type of algorithm and by the number of features retained. This chapter aims at investigating these issues. A comparison of the main feature reduction techniques proposed in the literature; both unsupervised and supervised, is proposed, and the influence of the most suitable number of dimensions to retain is investigated.

III) The last part of this thesis explores the possibility to use sub-pixel information in order to obtain a representation of the thematic maps which can be more precise in case of

moderate spatial resolution. The traditional classification maps obtained by analyzing hyperspectral data are not suitable when mixtures of classes can be found in the image. If a pixel is jointly occupied by more than a single land cover class, it can not be represented by a single label without loss of information. The issue is addressed in the last two chapters, and a suitable alternative is proposed in order to obtain thematic maps at a finer spatial scale.

6. A new technique is proposed in Chapter 6 to handle the problem of mixed pixels. After retrieving pure spectral components by mean of endmember extraction or unsupervised clustering methods, spectral unmixing is used to obtain the fractional abundance of each land cover class. In order to obtain an improvement of the spatial resolution of the thematic maps, the results are first upsampled and filled with the results of the first step, and then a spatial regularization is applied to correctly locate sub-pixels within the original pixel. The experimental results show the interest of the proposed method from a qualitative and quantitative point of view.
7. Chapter 7 shows a way to incorporate the super-resolution method proposed in the previous chapter into a supervised classification framework. The base classifier considered in this work is a probabilistic SVM. The assumption done is that the pixel classified with a highest probability can be considered as pure pixels, while the others are handled in a second step. The framework proposed in Chapter 6 is introduced at this point, by using an adaptive technique to compute, for each pixel, the possible "endmember candidates" representing the classes which could be contained within the pixel. Results show the possibility to jointly use supervised classification and spectral unmixing in a single framework in order to improve the classification accuracy of thematic maps.

The Appendices at the end of the manuscript present the data set analyzed in the work and the indexes of performance considered for evaluating the validity of the proposed methodologies.

In summary, the main contribution of the thesis are:

- The investigation of advanced methods making use of both spectral and spatial information, showing results in terms of accuracy comparable or better than the state-of-the-art approaches proposed in the literature.
- The investigation of innovative feature reduction techniques based on spectral unmixing concepts, to incorporate information about the (possibly) mixed nature of the training samples and of the whole data set.
- The proposition of a new method allowing to obtain classification maps at a finer spatial resolution, in a totally unsupervised way, to mitigate the problem of low spatial resolution and presence of mixed pixels into the data set.
- The exploitation of sub-pixel information in supervised classification framework to improve thematic maps from a qualitative and a quantitative viewpoint.

Part I

ICA for the Classification of Hyperspectral Images

Chapter 1

Feature Reduction - Extraction

Abstract

This chapter provides a theoretical introduction to the feature reduction/extraction techniques which will be used in the first part of the thesis. Firstly, two spectral based techniques for feature reduction are presented, Principal Component Analysis (PCA) and Independent Component Analysis (ICA). Theoretical differences are highlighted, and advantages provided by ICA, which will allow one to integrate such a technique into advanced classification approaches, are described. In the second part of the chapter, the details of two well known spatial based techniques, usually considered for feature extraction, are depicted. Traditional mathematical morphology approaches are compared to the use of attribute filters, which allow a better description of the image avoiding issues related to the image distortion.

Contents

1.1 Introduction

Hyperspectral images are composed of hundreds of bands with a very high spectral resolution, from the visible to the infra-red region. The wide spectral range, coupled with an always increasing spatial resolution, allows 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 [18].

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, this characteristic poses several significant issues, which need to be considered in the classification process for this kind of images. For example, from a theoretical point of view, data have a tendency to assume a *super-Gaussian* distribution and concentrate on the tails of the distribution, and the assumption of normal distribution does not stand any longer. At the same time, the quantity of data is very challenging for processing, and long computational time are usually required to analyze a hyperspectral image. For the purpose of classification, further problems are related to the curse of dimensionality. For instance, when a supervised classifier is applied to classification problems in high-dimensional feature spaces, the Hughes phenomenon [4] can be observed; that is, when the number of input features exceeds a given limit for a fixed training sample size, the classification accuracy will decrease.

The possibility to exploit all the advantages provided by hyperspectral images is therefore limited by the complexity of the data. On one hand, if compared to traditional remote sensing images, a much higher quantity of information is available, which could allow to detect very subtle differences between the materials of the observed scene. However, traditional techniques seems to be not effective for the analysis of such data, due to the above mentioned issues.

Dimensionality reduction in a high-dimensional data space can decrease the computational cost and may also improve the accuracy during the classification process [7, 8]. Given a hyperspectral data image, dimensionality reduction is usually achieved, in the spectral domain, in essentially two ways: feature extraction and feature selection [2]. Feature extraction is to find the transformation from a higher dimension to a lower dimensional feature space with most of the desired information content preserved [2, 19]. This transformation can be either a linear or nonlinear combination of the original variables and may be supervised, if taking into account the ground truth data information, or unsupervised.
Alternative methods of feature extraction are based on the spatial properties of the analyzed scene [20]. Recent developments of spatial based feature extraction approaches made possible to analyze the objects observed within an image by taking into account characteristics like the shape, the diagonal of the bounding box, or the moment of inertia [21]. All these features may be very useful for the classification of remote sensing images, since objects belonging to the same land cover classes will be likely to have similar feature values. For example, if we consider a land cover class corresponding to "road" (or, if we consider materials, to "asphalt"), it will be very likely that pixels belonging to this class show a high value of the moment of inertia, due to the elongated shape of these objects. In this chapter, we will introduce the mainly problematics related to spectral feature reduction (Section 1.2) and spatial feature extraction (Section 1.3). Some Conclusions are drawn at the end of the chapter.

1.2 Spectral Feature Reduction

Each pixel of a hyperspectral image is a vector which contains all the spectral information provided by different spectral channels at different wavelengths. The dimensionality of the vector depends on the number of spectral bands collected by the sensor and, in the case of hyperspectral data, usually corresponds to few hundred of values. As pointed out in the introductory section, this huge quantity of information allows better discrimination of land cover material, but it also poses several critical issues, both from a theoretical and practical point of view.
The behavior of the pattern in a hyper-dimensional space is rather different of what is intuitive for us, being used to bi- and three-dimensional spaces we use everyday. As pointed out in [2,22] geometric concepts that are self-evident in low dimensional spaces do not necessarily apply in higher-dimensional space. Actually, is all the information contained in a hyperspectral scene needed to a detailed analysis of the data? Hyperspectral sensors acquire images on a wide spectral range, but they consider very narrow spectral channels, so that redundant information may be contained in the data. Moreover, the useful information in hyperspectral images is contained in a small part of the feature space, while a big part of the data does not contain useful information. As a result of this, the number of features which can be used as input for the classifier can be reduced, without any significant (or with very few) loss of information.
As a matter of fact, feature selection / extraction techniques are usually performed as preprocessing to hyperspectral data analysis [2]. Such processing can also be performed for multispectral images, to enhance class separability or to remove a certain amount of noise.
One of the most widely used techniques for data compression and dimensionality reduction is the Principal Component Analysis, which estimates the data eigenvalues in order to perform a projection into a new feature space where the maximum of the data variance is preserved. However, since PCA considers only statistic up to the second order, it may be not effective when used for hyperspectral data [14]. In order to overcome the well-known problems connected to the nature of PCA, some variations of the algorithm were more recently proposed [9–12], making use of contextual information [23], or in order to improve data visualization [24].
Independent Component Analysis has received increasing attention in the last years [25], [26], [27]. This technique projects the data into a new space were the estimated components are as independent as possible from a statistical point of view. Despite the interesting properties of this method, when compared to PCA, the incorporation of ICA into more advanced classification approaches for hyperspectral image classification has only been tested in [10, 28].
In the following discussion, we introduce the well-known problems of hyperspectral data in the spectral domain in order to highlight possible advantages provided by Independent Component Analysis, which is the core method used in the next chapters. The mathematical basis of PCA and ICA are then presented respectively in 1.2.2 and 1.2.3.

1.2.1 Problems with high dimensional data

In this section, classical problems encountered in high-dimensional space are presented. We will mainly refer to previous works of the literature which investigated the high dimensional data paradigm. A great quantity of theoretical information and experimental validations can be found in [2, 22]. In the following, we will outline the main conclusions of the mentioned works. An extensive discussion of mathematical background, which would be out of the scope of this introduction, is omitted. However, a clear understanding of the following statements is needed to introduce problematics related to high dimensionality data.

1. Hughes effect: with a limited number of training samples, there is a classification accuracy penalty as the number of features increases beyond a certain point [4].

2. It has been proven that for good estimation of the parameters, the required number of training samples is linearly related to the dimensionality for a linear classifier and to the square of the dimensionality for a quadratic classifier [29].

3. The second-order statistics play a more important role in classification: it has been shown that when the dimensionality increases, considering only the variance of multivariate data led to significantly better classification results than considering only the mean [30]. However, the variance itself is not enough to describe all the information contained by hyperspectral data. When considering only the variance, important information residing in small part of the data space (that is, parts corresponding to low amount of variance) can be neglected.

4. As pointed out before, data tend to have a *super*-Gaussian distribution, with concentration of the information in the tails. In a corresponding way, uniformly distributed data tend to concentrate on the corner of the simplex enclosing the information.

Points 1 and 2 are both critical in case of supervised classification of hyperspectral images. Having a number of training samples equal to the square of the spectral channels is a condition seldom verified, thus leading to overfitting phenomena. To alleviate such a problem, feature reduction methods are often used, since the hyperspace in which data live is mostly empty, and therefore the components of the image can be reduced without (or with few) loss of information.
Points 3 and 4 should be taken into account when considering the approach to perform feature extraction. As a matter of fact, they suggest that techniques based on second order statistics like variance (*e.g.*, Principal Component Analysis) may not be the most suitable to retain all the useful information. Alternative techniques could be more interesting for classification purposes. In this work, we will analyze Independent Component Analysis, due to its attractive properties explained in Section 1.2.3.

1.2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a classic technique in statistical data analysis. PCA de-correlates the variables $x_1, ..., x_n$ of a given random vector $\mathbf{x} \in \mathbb{R}^n$. In order to perform the dimensionality reduction, only a number of the first components, which span the largest variance of the data, are retained. The variables of the projected vector $\mathbf{y} = \mathbf{P^t x}$ are uncorrelated with the other variables. This means that its covariance matrix $\sum_y = \mathrm{E}[\mathbf{y}_c \mathbf{y}_c^t]$ is diagonal, where $\mathbf{y}_c$ is the centered vector $\mathbf{y}$. The computation of the covariance matrix can be written as:

$$\begin{aligned} \Sigma_y &= \mathrm{E}[(\mathbf{y} - \mathbf{m}_y)(\mathbf{y} - \mathbf{m}_y)^t] \\ &= \mathrm{E}[(\mathbf{P}^t\mathbf{x} - \mathbf{P}^t\mathbf{m}_x)(\mathbf{P}^t\mathbf{x} - \mathbf{P}^t\mathbf{m}_x)^t] \\ &= \mathbf{P}^t E[(\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)^t]\mathbf{P} \\ &= \mathbf{P}^t \Sigma_x \mathbf{P}. \end{aligned} \tag{1.1}$$

Σ_x is a real-valued symmetric matrix of finite dimension. By the spectral theorem, Σ_x can be diagonalized by an orthogonal matrix $\mathbf{M}(\mathbf{M}^t = \mathbf{M}^{-1}) : \mathbf{M}^{-1}\Sigma_x\mathbf{M} = \Sigma_y$ (from (1.1)). By identification, $\mathbf{P}$ is an orthonormal matrix which is found by solving the eigenvalue problem with unitary norm condition on the eigenvectors (v):

$$\begin{aligned} \lambda \mathbf{v} &= \Sigma_{\mathbf{x}} \mathbf{v} \\ ||\mathbf{v}||_2 &= 1. \end{aligned} \tag{1.2}$$

It turns out that $\mathbf{P}$ consists of the set of all eigenvectors $\mathbf{v}$ of Σ_x, with one eigenvector per column.

1.2.2.1 Reducing the dimensionality using PCA

The eigenvalues obtained represent the variance of the variable $\mathbf{y}$, i.e., var($\mathbf{y}_i$)= λ_i. They are stored in decreasing order $\lambda_1 > \ldots > \lambda_n$.Feature reduction is performed using the following postulate: the greater the variance, the greater the contribution to the representation. Thus, variables associated with high eigenvalues need to be considered and should remain after feature reduction. The problem lies in selecting sufficient principal components so that the reconstruction error is low. It can be shown [31] that the error in reconstruction, in the mean square sense, of $\mathbf{x}$ using only the k first principal components is

$$MSE = \sum_{i=k+1}^{n} \lambda_i. \tag{1.3}$$

Therefore, k is chosen in order to make the MSE fall below a given threshold t_{pca}, usually 5% or 10% of the total variance:

$$\frac{\sum_{i=k+1}^{n} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \leq t_{pca}. \tag{1.4}$$

Note this strategy is optimal for the purpose of representation [29]. As said in Section 1.2, PCA is an unsupervised algorithm whose objective is to represent the data in a lower dimensional space without discarding meaningful information. It does not use a criterion which is related to the classification error, and is, therefore, not appropriate as a dimensional reduction technique where low classification error is the main objective.

1.2.3 Independent Component Analysis

The PCA algorithm can be understood as an algorithm finding the direction of space with the highest sample variance, and moving on to the orthogonal subspace of this direction to find the next highest variance, and iteratively discovering an ordered orthogonal basis of highest variance. This is well adapted to normal processes, as their covariance is indeed diagonal in an orthogonal basis. In addition, the resulting vectors come with a "PCA score", i.e., the variance of the data projected along the direction they define. Thus when using PCA for dimension reduction, we can choose the subspace defined by the first k PCA vectors, on the basis that they explain a given percentage of the variance, and that the subspace they define is the subspace of dimension n that explains the largest possible fraction of the total variance.

However, in the case of non-Gaussian processes, as class distributions are in hyperspectral data, the variance may not be the quantity of interest. Although it is not optimal for classification, the PCA is often used for such a task, due to its simplicity and ease of use. In order to overcome the problem, Independent Component Analysis (ICA) is analyzed in this work. ICA was first proposed as a blind source separation technique in [6,31], and was immediately adopted in several fields of signal processing.

Due to its attractive properties, ICA is receiving a growing interest among the remote sensing

community, for feature reduction and spectral unmixing [27,32,33]. ICA tries to minimize all the dependencies of the retrieved components, using higher (than second) order statistics. In the following, the ICA model is presented into more detail.

1.2.3.1 The ICA model

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

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1.5}$$

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

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s}, \tag{1.6}$$

where $\mathbf{y}$ is a vector of independent components. 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 three general assumptions, in order to make sure that its basic model can be estimated [34]:

1. The components of $\mathbf{y}$, estimated from the observed signal $\mathbf{x}$, are statistically independent. This is the basic principle required for ICA to be valid.

2. At most one signal has a Gaussian distribution. If more than one component have 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 of full rank and invertible. This assumption is equivalent to saying that the number of independent components is equal to the number of observed mixture. It is done in order to simplify very much estimation, but it can sometimes be relaxed.

Under these three assumptions (or at least the first two), the independent components and the mixing matrix can be estimated, under some indeterminacies that will necessarily hold. In fact, for (1.5) if both $\mathbf{A}$ and $\mathbf{s}$ are unknown, at least two ambiguities cannot be avoided. First, the variances of independent components cannot be computed. In fact, any scalar multiplier in one of the sources could always be cancelled by dividing the corresponding column of the mixing matrix. Due to this, the energy of the components can be at first fixed by whitening in order to make all variances equal to unity, and consequently the mixing matrix is adapted. Second, because of similar reasons, the independent components cannot be ranked, because any change in their order will not change the possibility to estimate the model. Since the method is well known in the signal processing community, we will not give all the mathematical details of the ICA implementation. A detailed explanation of the basics of ICA would be out of the scope of this section. However, in the next subsection, we briefly review several possibilities which were proposed in the last years as a possibility to compute independence. Moreover, we give some details of ICA Jade, which is the algorithm considered in our experiments. We refer the reader interested in a complete explanation of the general framework of ICA to references [31,35,36].

1.2.3.2 Independence measures

As is stated by its name, the theoretical concept on which ICA rests is independence. Independence is a much stronger assumption than uncorrelatedness. Contrary to common decorrelation methods, such as Principal Component Analysis (PCA) and Factor Analysis (FA), 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:

- Nonlinear decorrelation. Find the matrix $\mathbf{W}$ so that the transformed components $g(\mathbf{y}_i)$ and $h(\mathbf{y}_i)$ are uncorrelated, where g and h are some suitable nonlinear functions. Possible non-linear functions can be derived through the maximum likelihood approach, or based on the mutual information.

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

Classical algorithms, such as FastICA [37], Infomax [38] have been developed using the above criteria. Another approach for the estimation of the Independent Components is JADE [39], which makes use of 4th 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 [40, 41], and since it has provided in a preliminary test better results than FastICA and Infomax.

Cumulant tensors can be considered as the generalization of the covariance matrix at orders higher than the second order. If we consider a random vector $\mathbf{x}$ with a probability density function (pdf) $p(\mathbf{x})$, its characteristic function is defined as the inverse Fourier transform of the pdf [42]

$$\Phi(\omega) \quad = \mathrm{E}\{\exp(j\omega\mathbf{x})\} = \quad \int_{-\infty}^{\infty} \exp(j\omega\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{1.7}$$

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

$$\Psi(\omega) = \sum_{k=0}^{\infty} a_k \frac{(j\omega)^k}{k!} \tag{1.8}$$

the ath cumulant is defined as the derivative

$$a_k = (-j)^k \quad \left. \frac{d^k \Psi(\omega)}{d\omega^k} \right|_{\omega=0}. \tag{1.9}$$

It can be shown that the second, and third order cumulants for a zero mean random vector are [43]

$$\mathrm{cum}(x_i, x_j) \quad = \quad \mathrm{E}\{x_i x_j\} \tag{1.10}$$

$$\mathrm{cum}(x_i, x_j, x_k) \quad = \quad \mathrm{E}\{x_i x_j x_k\} \tag{1.11}$$

We refer the reader interested in a more detailed explanation of the cumulants and their properties, to [44, 45]. 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 analogue 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:

$$\mathbf{F}_{ij}(\mathbf{M}) = \sum_{kl} m_{kl}\text{cum}(x_i, x_j, x_k, x_l) \tag{1.12}$$

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

Joint Approximate Diagonalization of Eigenmatrices (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

$$\mathbf{F}(\mathbf{M}) = \lambda\mathbf{M}, \tag{1.13}$$

where λ is a scalar eigenvalue.
Let us consider data which follow the ICA model, with whitened data

$$\mathbf{z} = \mathbf{VAs} = \mathbf{H^T s} \tag{1.14}$$

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

$$\mathbf{M} = \mathbf{h_m h_m^T} \tag{1.15}$$

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 independent components, and all the other eigenvalues are zero. By determining these eigenvalues, we can obtain the independent sources we are looking for. We refer the reader interested in a more detailed explanation of the theoretical framework of JADE to [39].

1.2.3.3 Applying ICA to hyperspectral images

When applying ICA to hyperspectral images, two representations can be considered: spectral and spatial mixture models. Roughly speaking, the spectral mixture model corresponds to retrieve the independent spectral components of the image, while the spatial mixture models provide a description of the material abundances which are ideally independent. If we consider a hyperspectral data cube with N_f images of $N_z = (N_x \times N_y)$ pixels obtained from N_f frequency bands and for simplicity, assume raw vectorized images $\text{I}(n, \lambda_k)$, with $1 \leq n = (i-1)N_y + j \leq N_z$ (where i and j are the initial row and column image indices) is the spatial index and k, $k = 1, \ldots, N_f$, is the spectral index for wavelength λ_k, the two representations can be described as follows:

- Spectral Mixture Model: Each pixel of spatial index n gives an observed spectrum of N_f frequency samples, which is represented by the linear approximation:

$$I_n(\lambda_k) \approx \sum_{p=1}^{N_c} a_{(n,p)}\phi_p(\lambda_k), \qquad \forall n = 1, \ldots, N_z \tag{1.16}$$

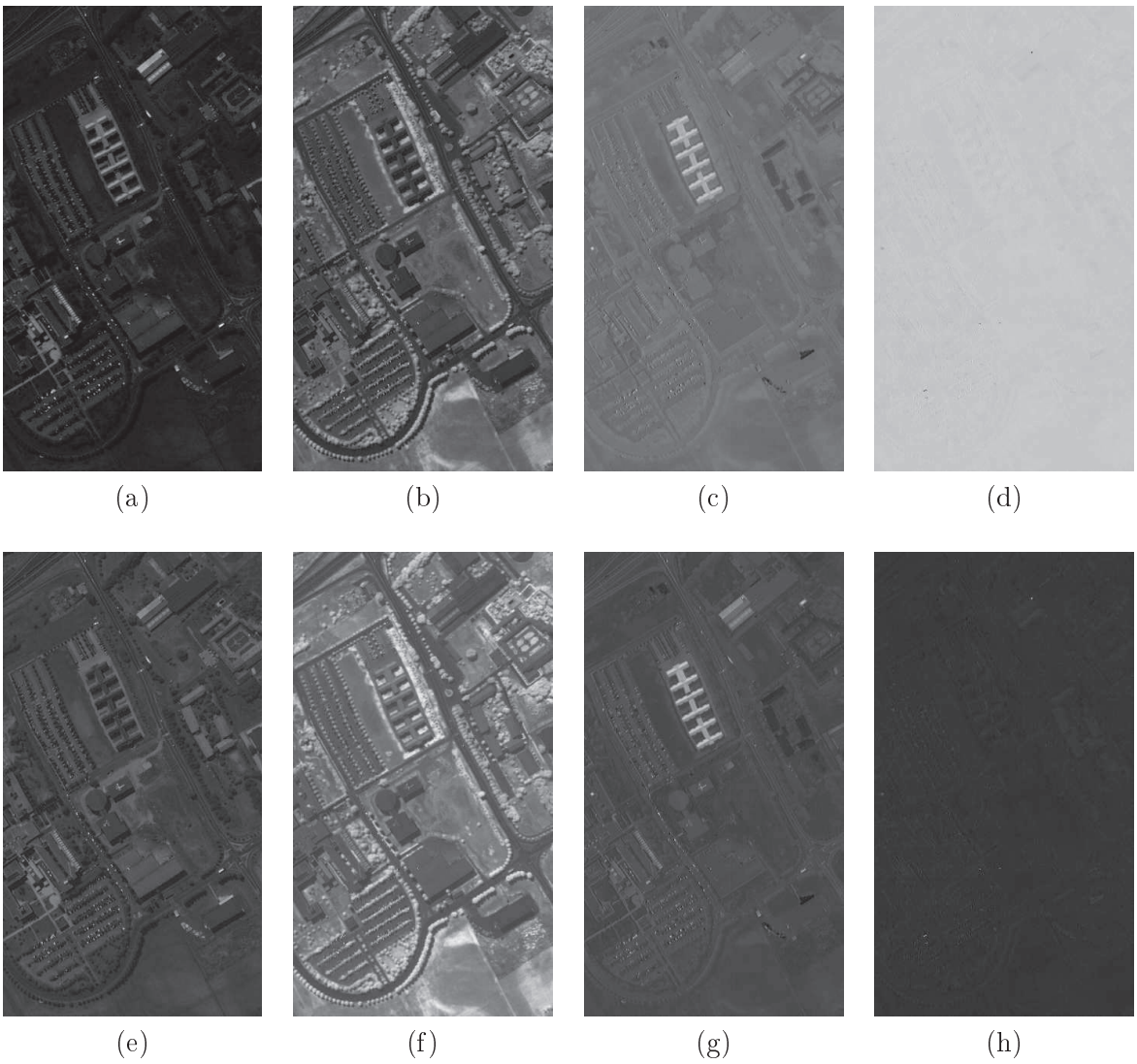


Figure 1.1: Principal Component Analysis vs. Independent Component Analysis: (a)-(d) First four principal components, ROSIS University data set; (e)-(h) First four independent components.

where $\phi_p(\lambda_k)$, for $p = 1, \ldots, N_c$, are the constituent reflectance spectra, and the number N_c is chosen according to the desired accuracy of the approximation. Denoting as $I(\lambda_k)$ the vectorized image (of dimension $N_x \times N_y$), the $(N_z \times N_c)$ mixing matrix A and $\phi(\lambda_\mathbf{k}) = [\phi_1(\lambda_k), \ldots, \phi_{N_c}]$, this spectral mixture model is then expressed as:

$$\mathbf{I}(\lambda_k) \approx \mathbf{A} \cdot \phi(\lambda_k) \tag{1.17}$$

Practically, this spectral model intends to approximate the spectrum of each pixel as a sum of N_c component spectra of the area corresponding to these pixel coordinates. If ICA is used for the estimation, then the N_c basis spectra ϕ_p, $p = 1, \ldots, N_c$, should be statistically independent. Moreover, the p-th column of the matrix A is the unfolded image associated to the basis spectrum ϕ_p. According to this model, we have $N_x \times N_y$ = 50,000 sensors and a small number of samples N_f = 174, for estimating the large matrix $\mathbf{A}$ which has $N_z \times N_c$ = 250,000 parameters (taking $N_c = 5$).

- Spatial Mixture Model: This model assumes that for each wavelength λ_k, the measured image $I_{\lambda_k}(n)$ is a weighted sum of N_c basis images, denoted $II_p(n)$, $p = 1, \ldots, N_c$:

$$I_{\lambda_k}(n) \approx \sum_{p=1}^{N_c} b_{(\lambda_k,p)} II_p(n), \qquad \forall k = 1, \ldots, N_f \tag{1.18}$$

In vector notations, denoting the matrix $\mathbf{B}$ and $II(n) = [II_1(n), \ldots, II_{N_c}(n)]^T$, one can write:

$$I(n) \approx \mathbf{B} \cdot II(n) \tag{1.19}$$

Practically, this spatial model intends to approximate the whole image at each frequency as a sum of N_c basis images. If ICA is used for the estimation, then the N_c basis images II_p, $p = 1, \ldots, N_c$, should be statistically independent. Moreover, the k-th column of the matrix $\mathbf{B}$ is the spectrum associated to the basis image II_k. According to this model, we have N_f = 174 sensors, and a very large number of samples N_z = 50,000 for estimating the matrix $\mathbf{B}$ which has $N_f = N_c < 900$ parameters (taking $N_c = 5$).

In our work, we will consider the ICA spatial model, which allows a better representation of the data without incurring in problems related to overfitting and matrix bad conditioning. An example of independent components extracted from a hyperspectral image (compared with the corresponding principal components) is shown in Fig. 1.1.

1.3 Spatial Feature Extraction

Dimensionality reduction techniques are helpful in avoiding problems related to the high dimensionality of the data and increase classification performances. Another possibility to improve the accuracy of land cover thematic maps is represented by the use of feature extraction approaches.
In opposite to feature reduction algorithms, which project the data into a new feature space to put there all the most interesting information, feature extraction techniques select the information directly from the original data. These techniques can be applied either in the spectral domain (only a subset or a non linear transformation of the original spectral channels are considered as input for the classification, possibly the bands providing the most discriminative features) and in the spatial context, in order to extract features related to shape, texture and other spatial characteristics which characterize the objects of a remote sensing scene. In this work we will investigate in more detail spatial feature extraction techniques, mainly because of two reasons: i) Spatial feature extraction techniques can be jointly used

with spectral feature reduction to improve the classification accuracy, which would not be possible with spectral feature extraction algorithms; ii) The use of spatial information has proven to be useful to boost classification accuracy of thematic land cover maps more than spectral based techniques [15,46].
The exploitation of spatial information is still more important after recent advances in remote sensor technology, which made available data with very fine spatial resolution (in the case of airborne sensors like ROSIS and HYDICE, up to few meters). The finer the spatial resolution, the larger the spectral variability of the pixels [47]. Therefore, the contextual information provided by spatial based approaches is very important to improve classification accuracy of supervised methods. However, also in case of coarse spatial resolution, objects of interest can be characterized by their spatial features (for example, agricultural fields will be likely to have a rectangular shape, independently of the sensor resolution), thus making attractive the use of spatial based techniques.
The contextual information needs therefore to be extracted into structures and between structures. Several techniques were proposed in the last years to extract spatial features from remote sensed images. Mathematical Morphology provides a well-established theory for analyzing the spatial relationship between sets of pixels [48,49]. Readers will find a review of Mathematical Morphology applications in remote sensing in [50]. The characterization of spatial information at different scales can be obtained by the application of Morphological Profiles (MP) [10], which are particularly suitable for representing the multiscale variability of the structures in the image, but it is not sufficient to model other geometrical features. To avoid this limitation, the use of morphological attribute filters instead of the conventional operators based on the geodesic reconstruction was proposed [51]. The application of attribute filters in a multilevel way leads to the definition of attribute profiles (APs) [51], which permit to model other geometrical characteristics rather than the size of the objects. Attribute Filter operators can extract a great deal of structural information, such as:

- direction
- area
- moment of inertia
- standard deviation
- ...

In the following, we will briefly introduces the basics of mathematical morphology and of attribute filters. Then we will present the possibility to use this methodology on a multilevel approach, by considering the concept of profiles. The main differences between the approaches are highlighted, and previous work related with remote sensing are discussed.

1.3.1 Theoretical Notions of Mathematical Morphology

Mathematical Morphology is a theory for non-linear image processing. It aims to analyze spatial relationships between pixels. In this section, basic notions of Mathematical Morphology are presented. Readers interested in Mathematical Morphology can find additional material in [48–50].

The basic morphological operators are erosion and dilation. When applying the erosion (dilation) operator to an image, every pixel of the image will be replaced by the maximum (minimum) value of the neighborhood defined by a spatial structure called structuring element. Examples of SE are shown in Fig. 1.2. By combining erosion and dilation, opening

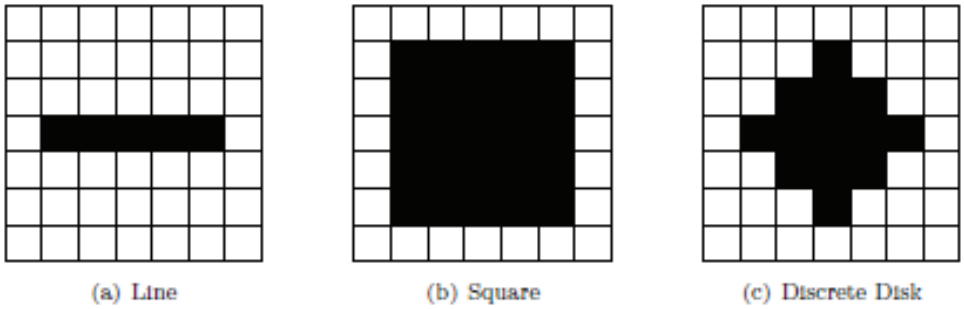


Figure 1.2: Examples of structuring elements.

and closing by reconstruction can be obtained. Opening by reconstruction is an image transformation that is increasing, idempotent and anti-extensive [49].
An opening by reconstruction consists of removing from a scalar image all the brighter connected components on which the SE does not fit in, with respect to the gray level of the neighboring regions. The use of opening transformation allows the appearance of distortion in the image to be avoided, since it transforms the image by merging flat regions. Therefore, the geometrical characteristic of the structures which are not removed are totally preserved, while the complexity of the image is decreased. The dual operator is the closing by reconstruction process. It removes from the original image all the dark connected components. The mathematical definitions of opening and closing by reconstruction of a gray-level image f are respectively given by:

$$\gamma_R^i(f) = R_f^\delta[\epsilon^i(f)], \tag{1.20}$$

$$\psi_R^i(f) = R_f^\epsilon[\delta^i(f)], \tag{1.21}$$

where ϵ^i and δ^i are the erosion and dilation with an SE of size i, respectively, R_f^δ and R_f^ϵ are the geodesic reconstruction by dilation and erosion, respectively [50].

A main issue is how to properly select the size of the structuring element. Remote sensing images contain object with totally different sizes, being therefore impossible to characterize all the structures with a single structuring element. In order to overcome the problem, the concept of morphological profiles was proposed as an interesting extension of mathematical morphology, for the analysis of remote sensing images.

A morphological profile is a concatenation of closing profiles and opening profiles, which form a sequence of closings and openings by reconstruction obtained by using SEs of different sizes. The mathematical definition of a Morphological Profile is:

$$MP(x) = \{\psi_n(x, \ldots, f(x), \ldots, \gamma_n(x))\} \tag{1.22}$$

and

$$MP_i(x) = \begin{cases} = & \psi_{n-i}(x) & if & 0 \leq i < n \\ = & f(x) & if & i = n \\ = & \gamma_i(x) & if & n < i \leq 2n \end{cases}$$

Thus, both the opening and closing profiles are computed with an SE of fixed shape and increasing size. The MP is effective for investigating the structure interactions of the

structures present in the image (i.e., objects in the surveyed scene) with the SE considered. According to its definition, an MP can only handle a scalar image (e.g., one band).

1.3.2 Extended Morphological Profiles

The generalization of EMPs to hyperspectral data was first proposed in [10]. The basic idea was to reduce the dimensionality of the data by using PCA, and to apply the transformation on each of the principal components retained. The features extracted by the transformation (i.e., eigenvectors ordered increasingly according to the values of the correspondent eigenvalues), called principal components (PCs), are meaningful for representation since they account most of the variance of the data in the original feature space. Thus, only the first PCs are considered for the analysis, while the others are discarded. As a general guideline, the number of the considered PCs should contain about 99% of the total variance of the data. Subsequently, on each of the PCs, a full MP is computed. Thus, the EMP of c PCs can be formalized by:

$$EAP = \{AP(PC_1), AP(PC_2), \ldots, AP(PC_c)\}, \tag{1.23}$$

where c defines the PC allowing to reach the threshold of 99% of variance. An example of morphological profiles built from a single image component is shown in Fig. 1.3.

Although the EMP was successfully used for including the spatial information in the classification of HR hyperspectral images, some drawbacks can be pointed out. The main shortcomings of EMPs can be identified as: (i) the spatial information can be modeled only through a small number of SE; (ii) the feature-extraction technique considered for reducing the dimensionality of the original feature space; and (iii) the increase in the number of features produced by the morphological transformations.
The first drawback mainly resides in the limited flexibility of the operators by reconstruction based on a SE in extracting informative features suitable to model the spatial characteristics of the objects. Obviously, a limitation in the capability of describing spatial features leads directly to a reduction in the discrimination power of the analysis. This drawback is addressed by the present work. The second limitation is related to the transformation used for reducing the dimensionality of the data prior to the morphological processing. The EMPs are computed on the few PCs extracted, which are suitable for the representation of the data, but might not be the most discriminative for classification [29]. Finally, the issue of the increase of the dimension of the feature space due to the computation of the EMP can be addressed again by applying some feature-extraction techniques, as already presented by [10], or by considering classifiers not sensitive to the curse of dimensionality.
In this work, the main emphasis is posed in the investigation of suitable techniques to preserve and model the spatial and spectral information of a hyperspectral data set. The main concerns regard therefore the first two drawbacks mentioned above.

1.3.3 Extended Attribute Profiles

A generalization of the concept of morphological operators by reconstruction leads to the definition of Morphological Attribute Filters [52]. These filters remove from the original image all the connected components that do not fulfill a given criterion, which can be defined by evaluating any attribute extracted from the scene. The main advantage provided by these filters is a great flexibility in the definition of the attributes. In fact, the attributes considered can be purely geometric (e.g. area, length of the perimeter, image moments, shape factors), textural (e.g. range, standard deviation, entropy), etc. Actually, an attribute can be any measure computable on the regions of the connected components present in the image.

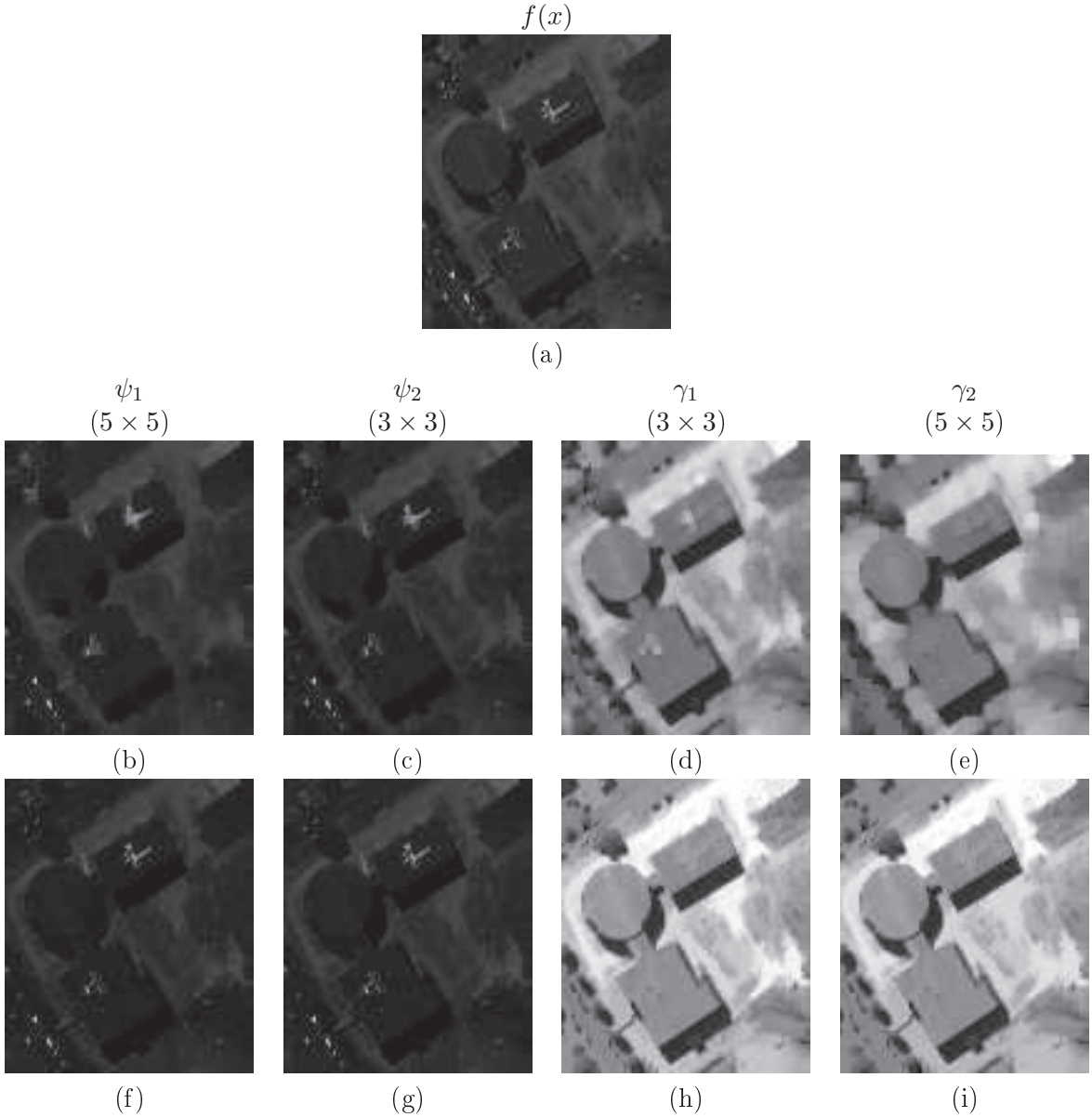


Figure 1.3: (a) The original image. (b)-(c) Morphological image closings obtained using respectively a 3×3 and a 5×5 square as structuring element. (d)-(e) Morphological image openings obtained using respectively a 3×3 and a 5×5 square as structuring element. (f)-(g) Attribute thinning obtained using as criterion the area of the object with a value of 100 and 200. (h)-(i) Attribute thickening obtained using as criterion the area of the object with a value of 100 and 200.

Analogously to the definition of EMPs, we can compute the APs on the c PCs extracted from the original hyperspectral data. This leads to the definition of the EAP:

$$AP(f) = \{\phi_n^T(f), \ldots, f, \ldots, \gamma_n^T(f)\}. \tag{1.24}$$

In conclusion, we remind that EAP includes in its definition the EMP (because the operators by reconstruction can be viewed as a particular set of morphological attribute filters) and, thus, it can be considered as its generalization. The main advantage in using the EAP instead of the EMP relies on the great flexibility given by the definition of the attributes used in the processing for modelling the spatial features that need to be extracted.

1.4 Conclusions

In order to improve the classification accuracy of hyperspectral images, suitable pre-processing techniques may be helpful to avoid problems related to the complex nature of the data. This chapter provides an overall introduction of traditional feature reduction and feature extraction techniques. In a first part, spectral based techniques, used to project hyperspectral data into a lower dimensional space where most of the spectral information can be retained, were presented. Two methods were described, Principal Component Analysis and Independent Component Analysis, and the reasons which make ICA particularly suitable in case of classification of hyperspectral remote sensing data were analyzed in detail. In the second part of the chapter, spatial feature extraction techniques were analyzed. Traditional morphological filters were introduced, and the concept of Attribute Filters and Profiles, an extension of traditional morphological filters, was described.
In the next chapters, the incorporation of these techniques into advanced classifier frameworks will be discussed, in order to obtain methods which can fully exploit the potentiality offered by hyperspectral data.

Chapter 2

Independent Component Discriminant Analysis

Abstract

In this chaper, Independent Component Discriminant Analysis (ICDA), a new classification technique for hyperspectral remote sensing classification data is presented. The method exploits the advantages provided by Independent Component Analysis to build a state-of-the-art classifier for high dimensionality data. The main concept is that, when projecting the data into a class-centered independent space, multivariate (joint) probability distributions can be easily approximated as the product of univariate (marginal) probability distributions. Therefore, a simple kernel density estimator can be considered to compute such probabilities, which are then included into a Bayesian classifier to compute the classification results. Experimental assessment shows the comparative effectiveness of the proposed approach.

Contents

2.1 Introduction

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 [2], 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 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 [4]. The drawbacks connected with very high dimensionality of hyperspectral data made methods traditionally used for land cover classification rapidly obsolete. Parametric classifiers widely used in pattern recognition such as the Maximum Likelihood classifier [53] or the Bayes classifier [29,54] model probability density functions for individual classes with parameters estimated from the training samples. Such approaches were proven to be ineffective when used for classification of hyperspectral data, due to unfavorable ratio between number of training samples and dimensionality of the data.
These methods estimate the multivariate densities of land cover patterns by considering only the mean and the variance values of the classes. When data distribution is not Gaussian, and when a limited training set is available, these estimates can diverge a lot from the real ones [55]. The use of non-parametric estimators like Parzen window and kernel density estimators, which do not make any assumption about the data probability distribution, is not effective in case of high dimensionality data, since important information can be hidden in low density regions [56].
In this chapter we will present a simple and effective method for the classification of hyperspectral images, based on the use of Independent Component Analysis. The projection of the data into an independent space allows one to estimate the multivariate probability densities as product of the univariate, since all the joint dependencies are nullified. This way, a simple Bayesian classifier with kernel density estimator can be used for estimation of classes, providing very good results in terms of classification accuracy. Comparative effectiveness is proven by comparison with one of the most common classifier of hyperspectral images, the Support Vector Machines (SVM). The remainder of the chapter is structured as follows. The next section describes the state-of-the-art approaches recently proposed to overcome critical points related to hyper-dimensionality. Section 2.3 describes the theoretical details of the proposed approach Independent Component Discriminant Analysis, and Section 2.4 the validation on real hyperspectral data sets. Conclusions are presented in the Section 2.5.

2.2 Supervised classification of high dimensional data in the literature

It is well known that the probability of error achieved with a Bayes classifier is the best that can be obtained [53], thus making Bayes classifiers attractive in pattern recognition. Nevertheless, the construction of an optimal Bayesian classifier is very challenging when dealing with high-dimensional data. These kind of data sets 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 labelled 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 non-labelled data and consequently to poor classification accuracies. Many efforts have been reported in the literature to overcome the intrinsic problems of high-dimensional data [8, 57–61]. 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 (semilabelled) samples; 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 curse of dimensionality, several improved covariance matrix estimators have been proposed to reduce the variance of the estimate [57,58]. 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 test 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 [59]. 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 (DBFE) [8], Projection Pursuit (PP) [60], Non-parametric Weighted Feature Extraction (NWFE) [61]. Nevertheless, the higher computational time required or the inevitable loss of information introduced by these techniques represent often an obstacle to obtain high performances, in terms of processing time or classification accuracy.

Advanced classifiers, like Artificial Neural Networks (ANN) [62–64] and kernel based classifiers [65–69], 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 Neural Networks [62] suffer basically of two main limitations. The first limitation is that the number and the size of hidden layers need to be set, which is not a straightforward task. The second limitation is that a very large number of iterations is sometimes needed to find a solution, making feature reduction a very useful step before the classification process. RBF Neural Networks [63] 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 a better performance than NN 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., SVM [66], KFD analysis [67], support vector clustering (SVC) [68], the regularized AdaBoost (Reg-AB) algorithm [69]) have been proposed in the machine learning community, providing successful results in various fields. SVM in particular have been widely investigated recently in the remote sensing community [70–73]. Camps-Valls and Bruzzone compared in [65] a number of these methods. SVM 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 SVM are the training time and the need to find the optimal parameters for the kernel. The training time, though much smaller than other kernel methods, quadratically depends on the size of the training set, and can be very large, especially when a large number of labelled 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 [74], which showed results similar to SVM but with a bigger computational burden. Due to the challenging problems of hyperspectral data classification,

several approaches have recently been proposed to exploit also the spatial information of the data [15, 46].
The need for an accurate classification method for hyperspectral data is still an important research topic. The use of Bayesian classifier is a very attractive possibility, thanks to the possibility to incorporate any kind of complementary information which could help improving the performances, as long as it can be stated as a probability function. In the following, we will review in detail the problems encountered by Bayesian classifier in high dimensional data space, and we will proposed an effective alternative to overcome them. The resulting algorithm, Independent Component Discriminant Analysis, is very efficient both in terms of classification accuracy and computational burden.

2.3 Independent Component Discriminant Analysis

In this work, inspired by [75], a non-parametric method for discriminant analysis based on the application of a Bayesian classification rule on a signal composed by independent components is proposed for the classification of hyperspectral images. The main characteristics of the method are the use of Independent Component Analysis (ICA) to retrieve independent components from the original data and the estimate of the multivariate density in the new data space computed with the ICA. The basic scheme can be represented as follows:

$$\mathbf{X}_k \quad \overset{(\mathbf{A}_k)}{\longrightarrow} \quad \mathbf{X}_k^* \tag{2.1}$$

$$f^m(\mathbf{X}_k) \quad \overset{(\mathbf{A}_k^{-1})}{\longleftarrow} \quad f^m(\mathbf{X}_k^*) = \prod f^u(\mathbf{X}_k^*), \tag{2.2}$$

where $\mathbf{A}_k$ is a proper transformation matrix which projects the data $\mathbf{X}_k$ centered on the class k into an independent space, and $\mathbf{A}_k^{-1}$ is the back transformation allowing one to estimate the multivariate probability distribution of the class k in the original space from the probability distribution of the independent space.
When the data are projected in an independent space, the estimates of their multivariate density function $f^m(\mathbf{X}^*)$ can be computed in a much easier way as the product of univariate densities $f^u(\mathbf{X}^*)$. 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 non parametric kernel density estimator is used to compute the density function of each independent component. Finally, the Bayes rule is applied for classification assignment. The main contributions of the work are the following: first, we propose an in-depth experimental analysis to highlight the potential of the method when used to classify hyperdimensional data. Second, we propose a simple but effective approach to choose the number of independent components which have 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 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.

In the following, we will explain problems encountered by traditional Bayesian classifiers when used for the classification of high dimensional data, and the theoretical basis of the proposed approach allowing one to overcome such issues.

2.3.1 Bayes Classification

The risk incurred when performing a classification of a measured vector $\mathbf{x}$ into one of K possible classes is given by:

$$R(\hat{k}|\mathbf{x}) = \frac{\sum_{k=1}^{K} L(k,\hat{k}) f_k(\mathbf{x})\pi_k}{\sum_{k=1}^{K} f_k(\mathbf{x})\pi_k} \tag{2.3}$$

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 can be 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 (2.3) 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

$$\hat{k} = d(\mathbf{x}) = \text{argmax}\{f_k(\mathbf{x})\pi_k\} \quad k = 1, .., K. \tag{2.4}$$

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

$$\pi_k = N_k/N \tag{2.5}$$

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. Thanks to its analytical tractability, the Gaussian (or *normal*) multivariate density is the most often used density for classification. However, multivariate estimation presents challenging problems when dealing with very high dimensional data.

2.3.2 Multivariate density estimation

The general expression of a multivariate normal density in d dimensions is written as:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\Big[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big] \tag{2.6}$$

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, $|\Sigma|$ and Σ^{-1} are its determinant and its inverse. Finally, $(\mathbf{x}-\boldsymbol{\mu})^T$ denotes the transpose of $(\mathbf{x}-\boldsymbol{\mu})$. Some classification rules are derived by assuming that the class-conditional densities are p-variate normal with mean vectors $\boldsymbol{\mu}_k$ and variance-covariance matrices Σ_k assumed to be non singular [76]. These two parameters are estimated from the training samples according to:

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\mathbf{x}_{ik} \tag{2.7}$$

$$\hat{\Sigma}_k = \frac{1}{N_k - 1}\sum_{i=1}^{N_k}(\mathbf{x}_{ik} - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_{ik} - \hat{\boldsymbol{\mu}}_k)^T \tag{2.8}$$

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 [55], as it is often the case for hyperspectral remote sensing data.

In order to overcome problems linked to the above mentioned 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 behaviour [77]. The idea is to apply a non-parametric density estimator to the training samples and then to substitute the obtained estimates into the Bayes decision rule (2.4). This family of density estimators does not assume any prior knowledge about the distribution of the data, like parametric estimators do. Many other non-parametric estimators can be found in the literature, such as the histogram approach, k-nearest neighbour and the expansion by basis functions method [78]. Thanks to its properties in terms of computation and accuracy, one of the most common procedure is to use multivariate kernel density estimator of the form:

$$\hat{f}_k(\mathbf{x}) = \sum_{i=1}^{N_k}\mathcal{K}\{\mathbf{x} - \mathbf{x}_{ik}; \mathbf{H}_k\} \tag{2.9}$$

where $\mathcal{K}$ denotes an unbiased kernel function and $\mathbf{H}_k$ is a diagonal covariance matrix. The choice of the kernel used for the estimation can be very important. However, in case of classification, different kernels can show similar trends [78]. In our experiments, we have considered one of the most widely used kernel, the Gaussian one:

$$\mathcal{K}(\S) = \frac{1}{h\sqrt{2\pi}}\exp\Big\{-\frac{x^2}{2h^2}\Big\}. \tag{2.10}$$

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 [56]:

$$\hat{f}_k(\mathbf{x}) = (2\pi)^{-p/2}\mathcal{H}N_k^{-1}\sum_{l=1}^{N_k}\prod_{j=1}^{p}\exp\Big\{-\frac{(x_j - x_{lkj})^2}{2h_{kj}^2}\Big\} \tag{2.11}$$

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

$$\mathcal{H} = \frac{1}{\prod_{j=1}^{p} h_{kj}} \tag{2.12}$$

The main drawback of this approach is that some important information for the classification process is not retrieved. In fact, it especially 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. A new technique is needed to estimate multivariate densities of hyperspectral data.

2.3.3 The approach of Amato

In [75], Amato *et al.* proposed an interesting approach to circumvent the problems of non-parametric 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 the estimation of a multivariate density as the product of univariate densities, which is then fitted to normality with the use of normal densities. Fig. 2.1 shows an example of univariate densities computed in a real data set. The results obtained are finally substituted in the Bayes rule for the class assignment. The basic steps of the approach are stated below:

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 be then 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 the chapter, we refer to the above approach as Independent Component Discriminant Analysis (ICDA). Next section will show experimental evidence of the effectiveness of the method.

2.4 Experimental results

2.4.1 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, large/small/very small size of the training set), four hyperspectral data sets were considered to evaluate the performances of the ICDA approach.

The complete description of the data sets is given in Appendix A. In the following, however, we will briefly recall them for sake of completeness. The four analyzed data sets are:

- ROSIS University data set: 103 bands in the spectral range from 0.43 to 0.86 μm, 610 by 340 pixels with spectral resolution of 1.3 m. There are nine land cover classes of interest.

- AVIRIS Indian Pine data set: 220 bands in the spectral range from 0.40 to 2.50 μm, 145 by 145 pixels with spectral resolution of 20 m. There are sixteen land cover classes of interest.

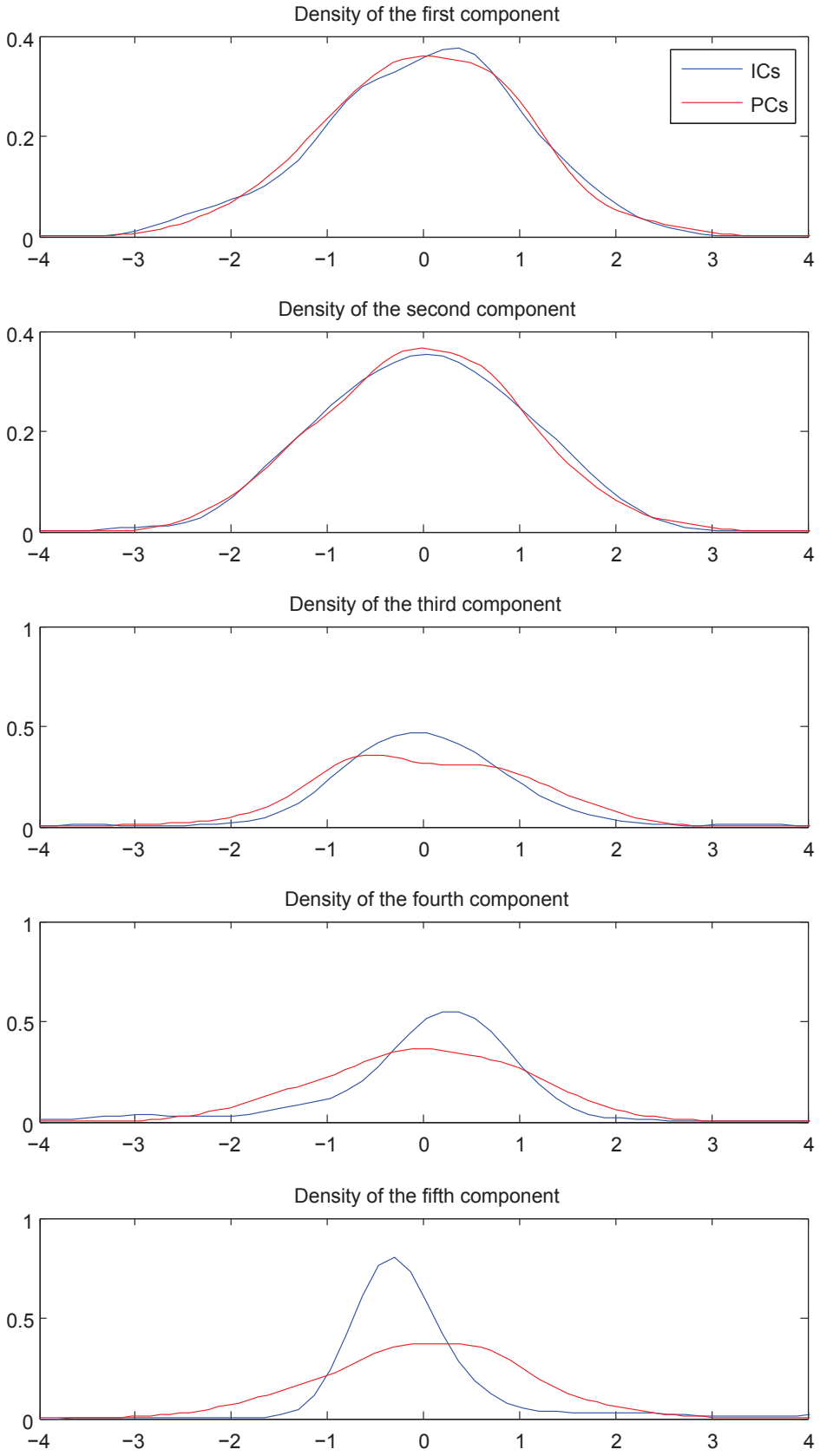


Figure 2.1: Example of the univariate densities estimation (ROSIS University data set, class 1), computed on the first principal components (red line), and independent components (blue line). Differences are due to the fact that, when projecting the data into an independent space, all the joint densities are minimized, which is not true in the case of an uncorrelated space. The advantage of using ICA is that these joint densities can not be easily estimated in high dimensional spaces.

Table 2.1: Information about the training and the test sets 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	-	-	-	-	-	-

- AVIRIS Hekla data set: 157 bands (a spectrometer was not working when the image was acquired), 560 by 600 pixels with spectral resolution of 20 m. There are twelve land cover classes of interest.

- HYDICE Washington DC Mall data set: 191 bands in the spectral range from 0.40 to 2.40 μm, 1280 by 307 pixels with spectral resolution of 3.6 m. There are seven land cover classes of interest.

For the ROSIS University data set, the standard training set provided with the data was used, corresponding to about 10% of the labeled samples (a single training set was therefore considered). In the other three cases, different training sets were randomly constructed from the reference data with a total of respectively 320, 600 and 70 pixels (corresponding to 20, 50 and 10 samples per class). In order to increase the statistical significance of the test, the experiments were repeated ten times with different training sets and the average accuracy results are reported, along with the standard deviation. As it can be noticed, the small size of available reference data poses major issues for the estimation of the data statistics, representing a very challenging test for the Independent Component Discriminant Analysis. A recapitulation of the training and test sets considered in the experiments can be found in Table 2.1.

2.4.2 Influence and choice of the number of independent components

The presented 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 10 fold cross-validation selection of the kernel's parameter [79], applied to hyperspectral image in the original spectral space (for comparison, Table 2.4 shows the results obtained with the SVM into the feature space obtained by applying PCA or ICA; it can be seen that results are similar to the original case). 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 independent components 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. Since $f_k(x)$ has to be computed for each class k, and the posterior probabilities of the different classes should be compared, all the multivariate densities have to be defined in the same subspace. Therefore, the number of independent components computed in step 1 of the presented 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 HYDICE DC Mall. A larger number of components could be computed for the other two data sets. Figure 2.2 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 processing time with respect to the independent components retained, for the four considered data sets. The Kappa coefficient of agreement is a parameter that estimates the correct percentage classification without considering the accuracy percentage that could be expected performing a random classification [80].

Although the number of components has a large influence on the final results, it can be seen that there is a wide region where ICDA 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 for the SVM. In the case of the ROSIS data set, the trend is not so linear, but again the choice of the number of ICs is not critical, i.e., we have a large range of values for which ICDA performs better than SVM in terms of classification accuracy. The difference of the behaviour 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 Independent Components used during the classification process is a very important task, in order to obtain a good accuracy of classification. In [75], the authors propose to retain the maximum possible number of IC's. This criterion is not always 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 Figure 2.2. 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, by considering a wide range of ICs to retain. The number of ICs which gives the best results is then used to perform the whole data set classification. Since we just have to choose the number of ICs which better characterize the classes of the image, we do not expect generalization problems, as could appear when selecting 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; 2) Splitting the number of samples in the cross-validation procedure influences the maximum number of components which can be retained. It should be noticed that the number of features to retain has only a limited influence on the final classification accuracy, as it is shown in Fig. 2.2.
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. In order 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 (as explained earlier in this section, the maximum number of components which can be selected depends on the rank of the covariance matrix obtained from the training samples of each class). This way, a finer step can be used, avoiding too much computational effort.

2.4.3 Performance analysis

Table 2.3 presents a comparison between the results obtained by the SVM (applied to the full feature space, where all the original spectral bands were considered) and the ICDA (with the proposed method to choose the number of IC's). The comparison is in terms of overall accuracy (OA), that 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 in terms of accuracy 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 for the ICDA. That happened for the AVIRIS Indian Pine and the HYDICE data set, where very small training sets were selected. The best results were obtained with a number of independent components 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 IC retained, also for small differences, thus leading to a higher value of standard

Table 2.2: 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

deviation. This phenomenon was less important for the AVIRIS Hekla data set due to the larger number of training samples selected. As an example, Figures 2.3 and 2.4 show some classification maps obtained with the SVM and with ICDA.
It has been observed in [81] that the comparison of two classification results in terms of overall accuracy 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 test process of the two different classifications, are exactly the same. In order to better evaluate the statistical significance of the difference between ICDA and SVM, McNemar's test was performed, which is based upon the standardized normal test statistic:

$$Z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} - f_{21}}} \tag{2.13}$$

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 test samples and thus is adapted to the situation at hand since the training and test sets were the same for each experiment. The results of McNemar's test [81] are shown in Table 2.2, and confirm the conclusions of previous experiments.

Finally, the computational burden of SVM and ICDA was 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 has to be calculated for each of the test samples. Figure 2.2 shows in the third column how the processing time of ICDA varies according to the number of Independent Component 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.

Table 2.3: Comparison of classification accuracy obtained with ICDA (best case) and SVM in the four analysed 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

2.5 Conclusions

.

The incorporation of Independent Component Analysis into a Bayesian classifier framework was presented in this chapter as a possible alternative for the classification of hyperspectral images. By minimizing all the statistical dependencies up to the fourth order, the approach Independent Component Discriminant Analysis gives the possibility to estimate the multivariate densities as the product of univariate. Experiments have been carried out on four different real data sets. The results of the experiments showed the effectiveness of ICDA, which provided better results than one of the state-of-the-art hyperspectral classifier, the SVM. Moreover, ICDA presents several other advantages: i) 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; ii) it is suitable to be used jointly with spectral-spatial techniques recently developed for SVM [15,46].
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 ICDA has similar results and outperforms SVM in terms of classification accuracy. Moreover, a simple and effective technique to choose the number of components to retain was proposed, providing results significantly better than the SVM. The computational burden of ICDA 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.
The results obtained in this chapter suggest that Independent Component Analysis is an effective tool for the representation of the spectral information contained into hyperspectral remote sensing data, and the statistical properties of the method can be effectively exploited to build a state-of-the art classifier. Further developments of this work lead to the investigation of the possibility to incorporate contextual information present in the image, in order to further improve the classification accuracy of thematic maps: Is it possible to exploit the advantages offered by ICA integrating the complementary information provided by spatial structures? The issue will be investigated in Chapter 3.

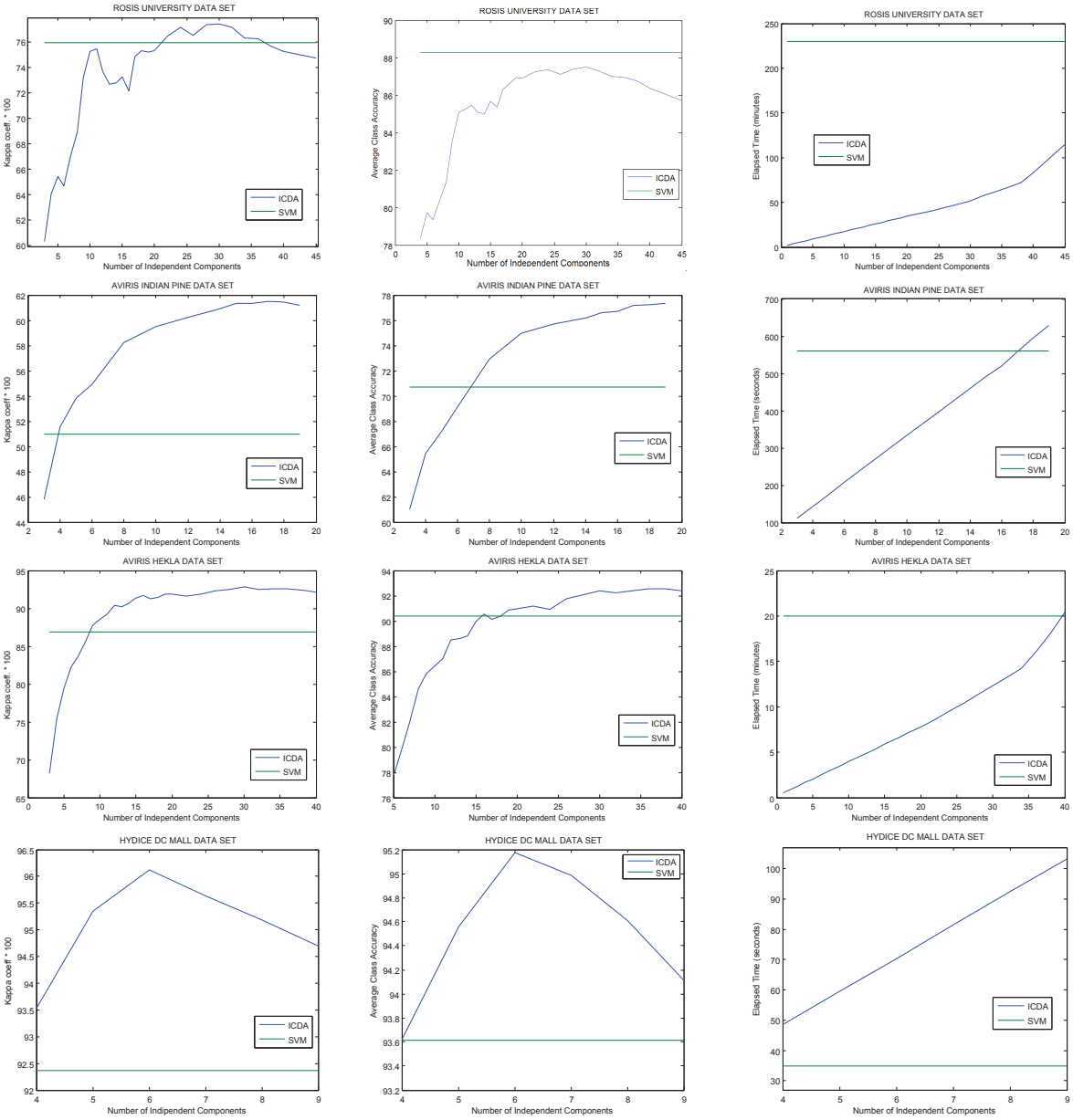
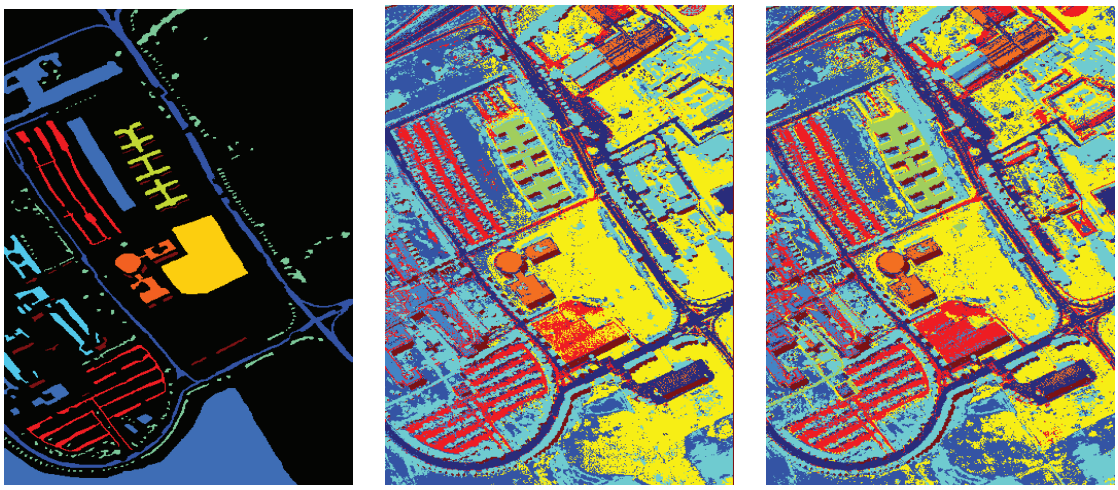
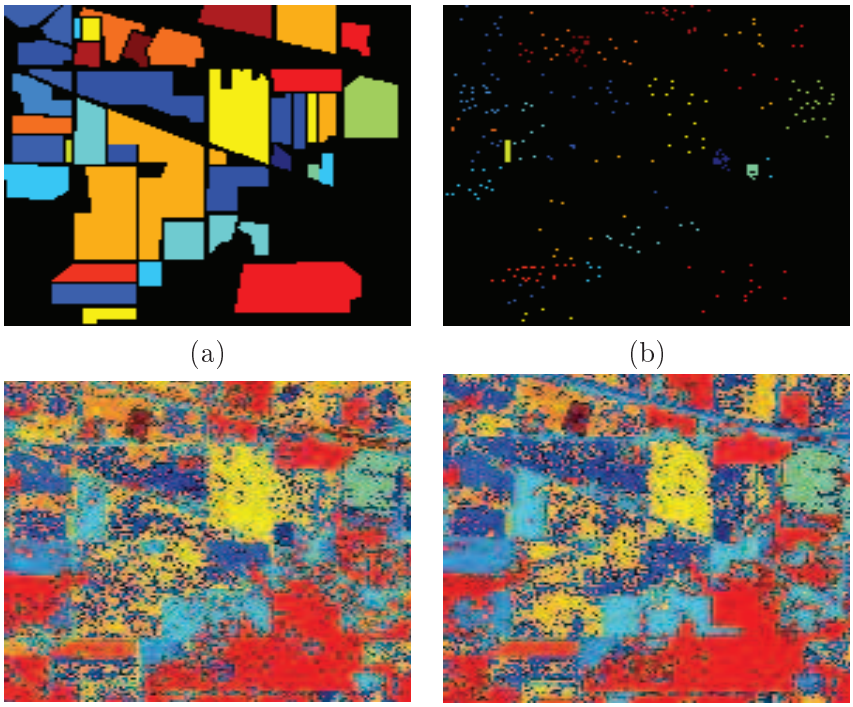


Figure 2.2: Comparison of Kappa coefficient of agreement (first column), average class accuracy (second) and classification processing time (third) obtained with SVM (by considering all the original spectral bands) and ICDA, with respect to different number of independent components retained, for the four considered data sets. The rows correspond respectively to: ROSIS, AVIRIS Indian Pine, AVIRIS Hekla and HYDICE data set.



(a) (b) (c)

Figure 2.3: (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)

Figure 2.4: (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.

Table 2.4: Influence of the dimensionality reduction on the overall classification accuracy. The number of components of PCA retained corresponds to the 99, the 99.9 and the 99.99% of the variance of the data. The same number of components was selected for the ICA.

	Hekla Data Set						
	PCA (2)	PCA (4)	PCA (20)	LDA (11)	ICA (2)	ICA (4)	ICA (20)
OA	81.03 ± 2.05	80.76 ± 2.89	82.12 ± 3.18	56.90 ± 28.33	81.06 ± 2.23	81.13 ± 2.98	82.03 ± 2.95
AA	72.97 ± 3.11	75.63 ± 3.93	80.60 ± 4.19	57.91 ± 27.45	72.97 ± 3.36	75.71 ± 4.26	80.56 ± 4.20
κ	78.02 ± 2.45	77.86 ± 3.32	79.59 ± 3.63	52.99 ± 29.33	78.86 ± 2.59	78.28 ± 3.43	79.41 ± 3.38
	AVIRIS Indian Pine						
	PCA (4)	PCA (31)	PCA (101)	LDA (15)	ICA (4)	ICA (31)	ICA (101)
OA	62.36 ± 1.85	59.42 ± 2.67	50.04 ± 2.32	9.82 ± 5.61	45.38 ± 3.45	62.13 ± 2.00	65.48 ± 1.28
AA	74.21 ± 1.10	73.86 ± 1.88	67.01 ± 1.66	22.74 ± 1.59	58.35 ± 1.83	74.09 ± 1.77	78.12 ± 1.02
κ	57.94 ± 1.97	54.78 ± 2.86	44.52 ± 2.50	3.17 ± 0.37	39.22 ± 3.40	57.73 ± 2.25	61.41 ± 1.36
	HYDICE Washington						
	PCA (24)	PCA (47)	PCA (-)	LDA (6)	ICA (24)	ICA (47)	ICA (-)
OA	85.80 ± 2.42	82.02 ± 4.61	-	-	84.00 ± 3.80	82.95 ± 4.70	-
AA	78.45 ± 4.69	78.23 ± 6.39	-	-	78.55 ± 3.99	78.32 ± 6.92	-
κ	79.41 ± 3.65	74.68 ± 5.99	-	-	77.31 ± 4.93	75.52 ± 6.62	-
	ROSIS data set						
	PCA (4)	PCA (15)	PCA (44)	LDA (8)	ICA (4)	ICA (15)	ICA (44)
OA	70.19	76.43	73.21	77.67	69.95	76.29	73.33
AA	80.37	85.16	84.12	84.48	79.84	85.46	84.01
κ	63.2	70.74	67.12	71.96	61.34	70.60	67.22

Chapter 3

Integration of Spectral and Spatial Information

Abstract

The combination of spectral based feature reduction technique and a spatial feature extraction method is presented in this chapter. The good results provided by performing dimensionality reduction with Independent Component Analysis are integrated with Extended Morphological Attribute Profiles, which allow the extraction salient features from the data. Several spatial features were tested, showing complementary results for the different data sets tested. Two techniques which allow to exploit the all information provided by the different filters are also proposed, providing an overall improvement of the classification accuracy.

Contents

3.1 Introduction

Pixel-wise classification is the most used tool when trying to understand land cover with hyperspectral imaging. As stated in the introduction of Chapter 2, a number of techniques based on only the spectral content of the pixels were recently proposed, with or without a pre-processing step to reduce dimensionality of the data and mitigate problems related to overfitting. Pixel-wise classification enables an accurate characterization of the observed scene considering only the spectral information, especially in case of "good classification conditions" (number of training samples representative of the land cover classes, low amount of noise at the sensor). However, the contextual information provided by spatial features could be very important to improve the classification accuracy when these conditions are not verified.
The importance of contextual information can be seen by performing a simple toy experiment (Fig. 3.1). Figure 3.1 (a) shows a band of a hyperspectral image, while Figure 3.1 (b), illustrates the same area, after the pixels have been randomly relocated. All the spectral information is preserved in the latter image, but the contextual information is lost. If one should declare which image contains more information, the intuitive answer would be for sure 3.1 (a). However, if trying to classify the two images with a pixel-wise classifier under the same conditions of training, the results in terms of overall accuracy would be exactly the same.
This simple example gives an idea of the quantity of information which is missed when using non contextual classifiers. The exploitation of spatial characteristics can be very important for improving accuracy, especially in the following cases:

- Noisy images: when a large quantity of noise corrupts a hyperspectral image, the classification based only on spectral information could lead to a high error rate. The use of contextual information could help avoid errors [82].

- Images with very high spatial resolution. It was proven that images with very high spatial resolution are more sensitive to outliers (that are, pixels with anomalous values of reflectance, which could be referred to direct reflectance rather than to sensor problems). Also in this case, the exploitation of the neighboring pixels information can be helpful for avoiding mistakes during the classification process [51].

- More generally, all the remote sensing images characterized by clear spatial structures, can have a benefit when using spatial based techniques [83].

The exploitation of the spatial information is very important for the classification of high resolution hyperspectral images, especially when considering urban areas and it is advisable to consider geometrical features in the analysis in order to derive spatially accurate maps [84]. As a matter of fact, typical land cover classes present in urban images are characterized by complementary spatial features. For example, classes like road will be likely to have an elongated shape, while building and grass fields will probably be characterized by rectangular shapes.
In this chapter, we propose a new classification chain to integrate the characteristics provided by Independent Component Analysis with the contextual information extracted by Morphological Attribute Filter. As shown in the previous chapter, ICA maximizes the information of the projected data contained in the original hyperspectral image, thus being attractive also when very few components are considered. The use of Attribute Profiles allows to extract several complementary spatial features from the image.

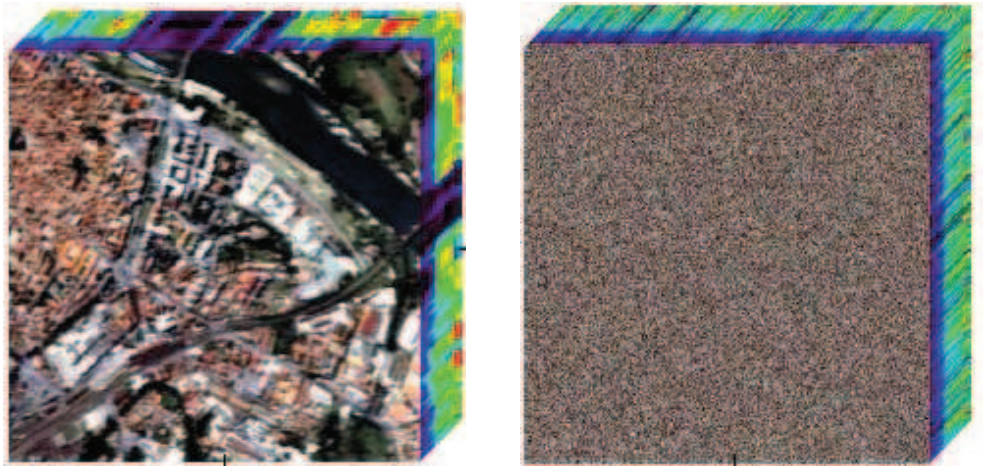


Figure 3.1: Contextual information: (a) A single band of a hyperspectral image (b) The same image, after a random shuffling of the pixels.

3.2 Related works

The spatial characteristics of the objects in the image can be modeled with several approaches. A number of techniques were proposed in the last years, based on clustering techniques [85], spectral spatial kernel [86], minimum spanning forest [46], conditional random fields [87], and many others.
A widely used technique for extracting spatial features is based on mathematical morphology [15]. Among all the operators belonging to this framework, morphological connected operators [52] have proven to be suitable for extracting spatial information while preserving the geometrical characteristics of the structures in the image (i.e., without distorting the borders). In [88], Morphological Profiles (MPs), a sequence of multi-scale connected operators, were applied to high resolution hyperspectral images by reducing the high dimensionality of the data by a Principal Component Analysis (PCA), and computing the profiles on the highest ranked principal components according to the size of the eigenvalues (leading to the definition of Extended Morphological Profiles, EMPs). Due to the limitations of PCA, when extracting information sources from high dimensional data, it was proposed to perform an Independent Component Analysis (ICA) before the computation of the MPs [28].

Analogously to the definition of EMP, the EAPs are obtained by computing Attribute Profiles (APs) [51] to the first principal components extracted form a hyperspectral image. The characterization of the spatial information obtained by the application of a MP is particularly suitable for representing the multi-scale variability of the structures in the image but it is not sufficient to model other geometrical features. To avoid this limitation, the use of morphological attribute filters instead of the conventional operators based on the geodesic reconstruction was proposed [51]. The application of attribute filters in a multi-level way leads to the definition of Attribute Profiles (APs), which permit to model other geometrical characteristics rather than the size of the objects. Moreover, APs show interesting characteristics when extended to hyperspectral images [21]. In greater details, analogously to [88], the APs were applied to the first principal components extracted from a hyperspectral image, generating an Extended Attribute Profile (EAP).

The aim of this chapter is to extend the work presented in [89] by presenting a technique based on Extended Attribute Profiles and Independent Component Analysis. Moreover, two approaches are investigated for combining the information extracted by EAPs computed with different attributes.

The chapter is organized as follows. In Section 3.3 the proposed classification technique based on ICA and EAPs is presented. The experimental results are illustrated in Section 3.4. Finally, conclusions are drawn in Section 3.5.

3.3 Proposed Classification Technique based on EAPs and ICA

The proposed approach is in two steps. In a first step, relevant spectral information is extracted from the image and concentrated into few components. Then, each retrieved components is analyzed by mean of attribute profiles to extract salient spatial information. Extended Attribute Profiles (EAPs) are based on the concept of the Attribute Profile (AP), which is a generalization of the widely used Morphological Profiles (MPs) [21]. Analogously to the definition of the extended morphological profiles, EAPs are generated by concatenating many APs computed on the components extracted by a PCA [51]. Conversely, in the proposed technique, the APs are computed on the c independent components (ICs) extracted by a ICA transformation applied to a multivariate image (e.g., the hyperspectral image):

$$EAP = \{AP(IC_1), AP(IC_2), \ldots, AP(IC_c)\}. \tag{3.1}$$

The AP is an extension of the MP, obtained by processing a scalar grayscale image f, according to a criterion T, with n morphological attribute thickening, (ϕ^T) and n attribute thinning (γ^T) operators, instead of the conventional morphological filters by reconstruction:

$$AP(f) = \{\phi_n^T(f), \ldots, f, \ldots, \gamma_n^T(f)\}. \tag{3.2}$$

Attribute filters are connected operators which operate on the connected components (i.e., regions of iso-intensity spatially connected pixels) that compose an image, according to a given criterion [52]. In general, the criterion compares the value of an arbitrary attribute, *attr* (e.g., area, volume, standard deviation, etc.) measured on the component C against a given reference value λ (which is the filter parameter), e.g., $T(C) = attr(C) > \lambda$. If the criterion is verified then the regions are kept unaffected, otherwise they are set to the graylevel of a darker or brighter surrounding region, according to if the transformation performed is extensive (i.e., thickening) or anti-extensive (i.e., thinning), respectively. When the criterion considered in the analysis is increasing (i.e., if it is verified for a connected component then it will be also verified by all the regions brighter or darker, according to the transformation, including the component) the attribute thinning and thickening operators are actually opening and closing transformations. Non-increasing criteria do not have a unique definition when considering grayscale images. In fact, different effects can be obtained by the operators with a non-increasing criterion according to the filtering rule selected [90].
Attribute filters can be efficiently computed by taking advantage of the representation of the input image as a rooted hierarchical tree of the connected components of the image. The tree is obtained by the Max-tree algorithm [90]. The approach based on this data representation is especially useful when computing an AP, since the image is converted to the tree only once (this is the most demanding stage of the filtering) and processed several times with the different criteria. An example of Extended Attribute Profiles computed on the first independent component of the ROSIS University data set can be seen in Fig. 3.2

3.3.1 Approaches to Deal with Multiple EAPs

The choice of the most suitable attribute and range of thresholding values (λs) for extracting the information on the geospatial objects is certainly a complex task, especially when *a piori* information on the scene is not available. A possible approach attempting to overcome

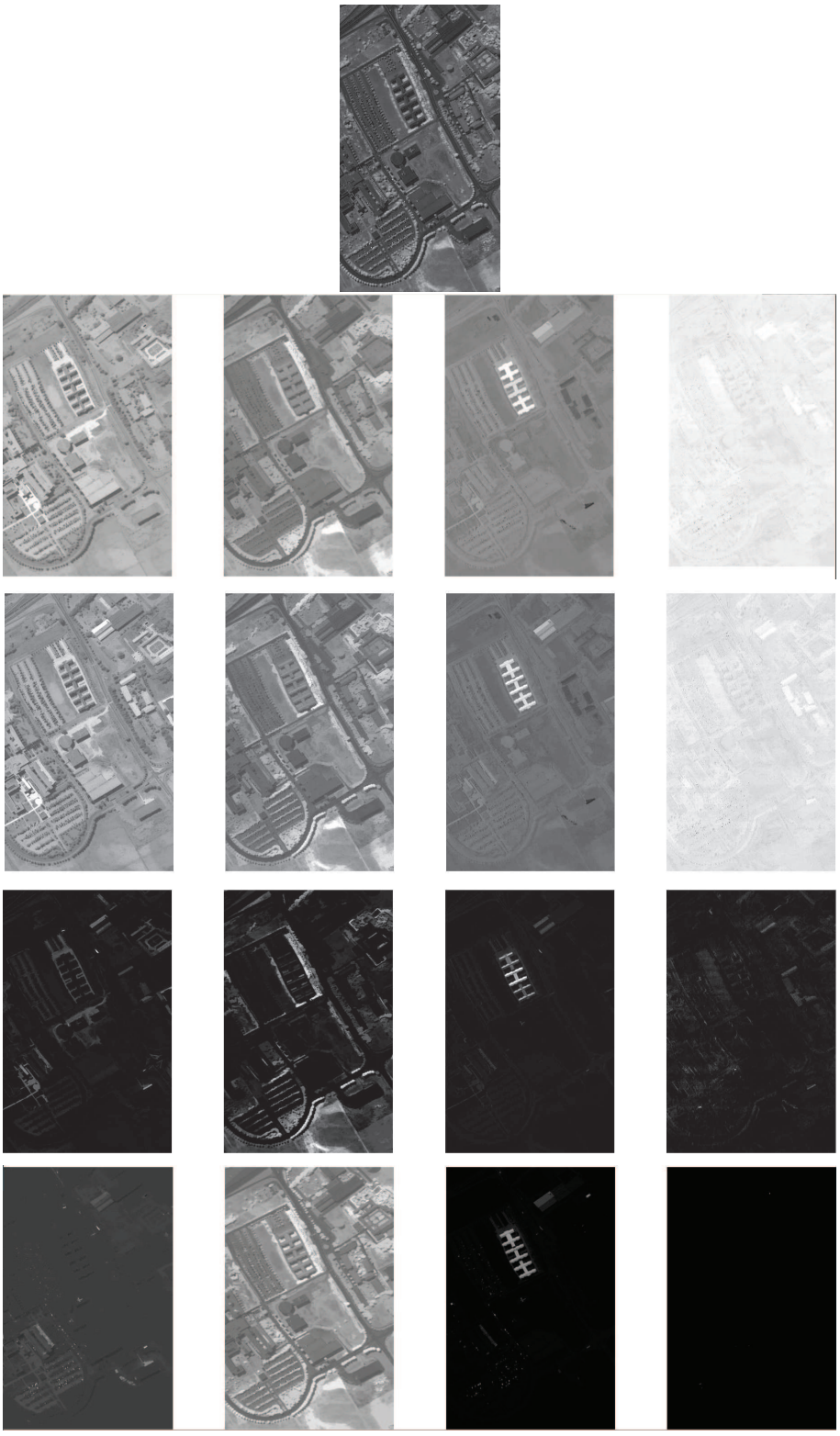


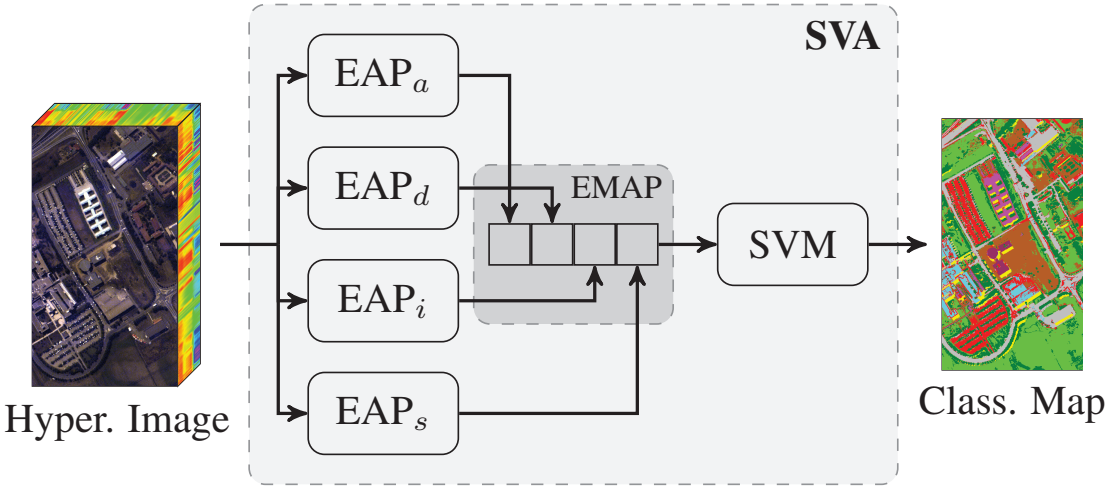
Figure 3.2: Examples of Extended Attribute Profiles computed on the first independent component of the ROSIS data set. Image at the top: first independent component. The four following rows correspond respectively to: area, diagonal of the bounding box, moment of inertia of Hu, standard deviation profile.

this issue relies on the computation of EAPs with different kinds of attributes. However, this leads to the problem of properly exploiting, in the analysis, the different information extracted by the computed EAPs.
A straightforward approach would be to simply consider two or three filters, together (for example, the two attribute filters providing the best classification accuracy in terms of percentage). In order to exploit the information of all the filters, we then proposed the *Stacked Vector Approach* (SVA), which combines the EAPs by concatenating them in a single vector of features (also called Extended Multi-Attribute Profile, EMAP [21]), see Fig. 3.3 (a).
However, even if complementary information can be extracted by considering different attributes, great redundancy is present in the features extracted. Thus, it is advisable that a classification algorithm with excellent penalization capability is used for classifying the features in order to handle the increased dimensionality.
Another approach is the *Fusion Approach* (FA) that is based on the separate classification of each EAP and on the fusion of the results obtained by the independent classifiers in order to generate the final decision map, see Fig. 3.3 (b). In comparison to the SVA, the FA keeps low the dimensionality of the data and increases the robustness of the results, especially if the different EAPs generate complementary errors.
In this work, an SVM classifier is considered with the One Against One (OAO) multiclass strategy. The fusion rule considered when combining the results of the single classifiers relies on the sum the votes of the classifiers applied to the four MPs, assigning each pixel to a class, according to the majority voting scheme. The illustration of this method is shown in greater detail by Fig. 3.4. Obviously, other decision criteria can be applied. The proposed approach represents however a good trade-off between the simplicity of the fusion rule and the result granted by the fusion technique.

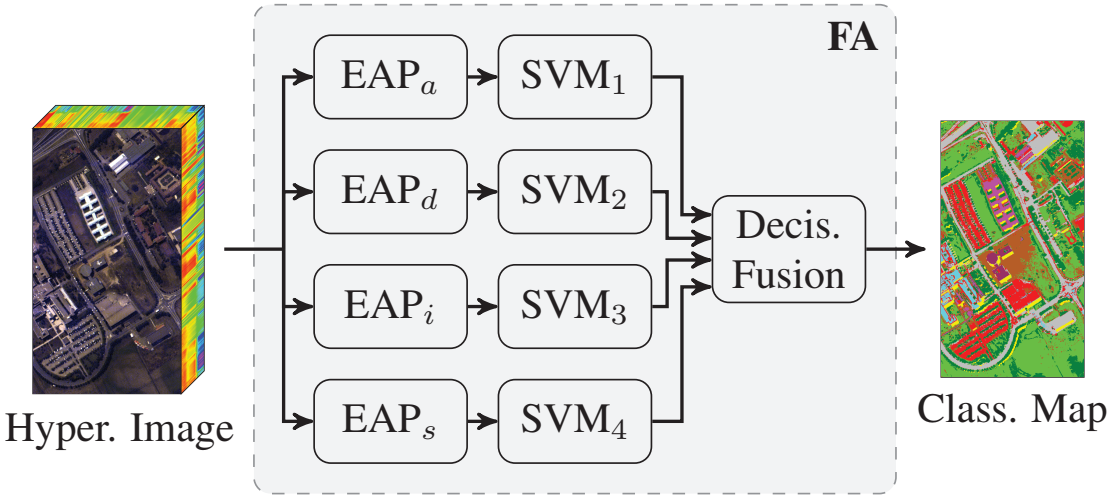
3.4 Experimental Analysis

The experimental analysis was carried out on two hyperspectral images acquired over the city of Pavia (Italy) by the ROSIS-03 (Reflective Optics Systems Imaging Spectrometer) hyperspectral sensor. The two images have geometrical resolution of 1.3 m. The first one shows the university campus (610×340 pixels) while the second one was acquired on the city center (1096×489 pixels). In the following we will refer to the two data sets as "University" and "Center" respectively. The original data are composed of 115 spectral bands, ranging from 0.43 to 0.86 μm with a band of 4 nm. However, noisy bands were previously discarded leading to 103 and 102 channels for the two images respectively. Nine thematic land-cover classes were identified in the university campus: Trees, Asphalt, Bitumen, Gravel, Metal sheets, Shadows, Self-blocking Bricks, Meadows, and Bare soil. For this data set, a total of 3921 and 42776 pixels were available as training and test sets respectively. In the center area the thematic classes found were: Water, Tree, Meadow, Self-blocking Bricks, Soil, Asphalt, Bitumen, Tile, and Shadow. The training and test sets for this data set were composed of 5536 and 103539 samples, respectively. The true color representation of the images and the test sets taken as reference are shown in Fig. 3.5. A detailed description of the two data sets can be found in the Appendix A.

In the analysis carried out, all the samples of the training set were used for the University data set while for the Center data sets, only 50 samples (randomly chosen from the full training set for each class) were considered. All the experiments conducted on the latter data set were run ten times with a set of different training samples each time. From both the two hyperspectral images four components extracted by PCA and ICA were considered. The first four PCs contain more than 99% of the total variance of the data for both the data sets. The components were rescaled to the range [0,1000] and converted to integer in order to be



(a)



(b)

Figure 3.3: Proposed approaches for dealing with multiple EAPs (for further information on the definition of the EAPs please refer to [21]). (a) Stacked Vector Approach (SVA) and (b) Fusion Approach (FA).

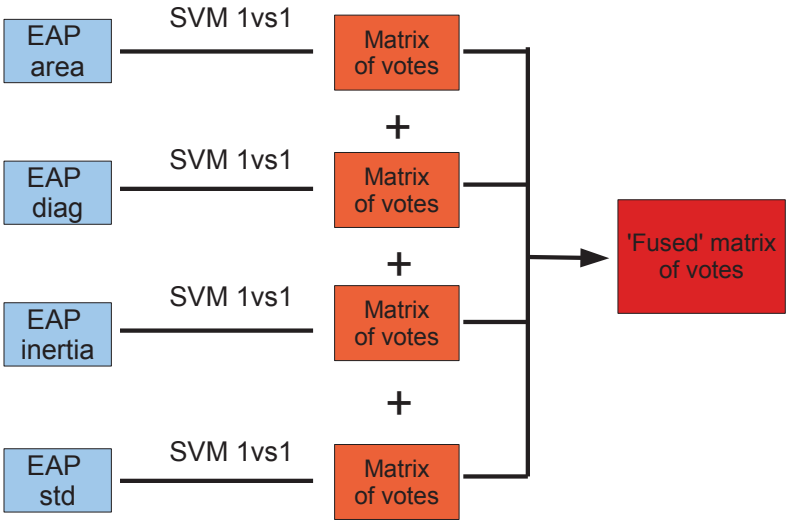


Figure 3.4: Fusion Approach, based on 1vs1 SVM classification. Each attribute filter is classified separately; at the end of the process the matrices of votes are summed in order to compensate the results.

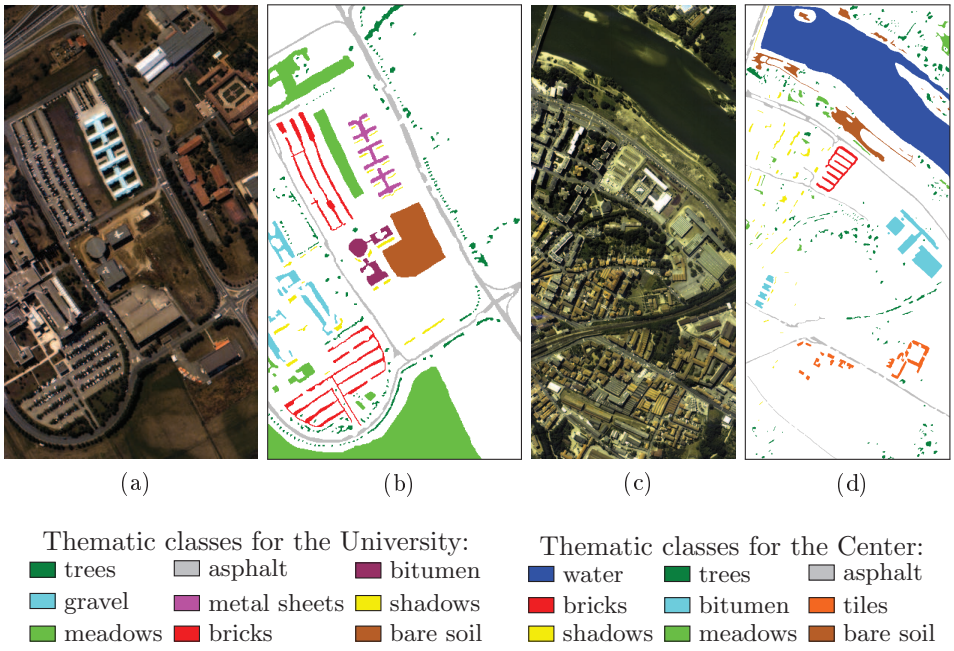


Figure 3.5: ROSIS Pavia data sets: True color representation and Test set for (a,b) University and (c,d) Center.

processed by the attribute filters. Four EAPs were computed by considering four different attributes on the components extracted by PCA and ICA:

i) a, area of the regions ($\lambda_a = [100\ 500\ 1000\ 5000]$);

ii) d, length of the diagonal of the box bounding the region ($\lambda_d = [10\ 25\ 50\ 100]$);

iii) i, first moment invariant of Hu, moment of inertia [91] ($\lambda_i = [0.2\ 0.3\ 0.4\ 0.5]$);

iv) s, standard deviation of the gray-level values of the pixels in the regions ($\lambda_s = [20\ 30\ 40\ 50]$).

The values of λ were selected by a preliminary analysis on the data in order to cover uniformly the range of values assumed by the connected components in the investigated images.

The area and the length of the diagonal of the bounding box extract information on the scale of the objects. The moment of inertia and the standard deviation are not dependent on the size dimension but they are related to the geometry of the objects and the homogeneity of the intensity values of the pixels, respectively. Each EAP is 36-dimensional, i.e., it is composed of four APs with 9 levels computed on each component extracted. In the sequel, the notation EAP_{attr} denotes the EAP built with the *attr* attribute. The classification maps are obtained by analyzing the features extracted by the extended profiles with an SVM classifier with RBF kernel. The model selection in the training phase of the classifier was based on a gradient descent method, which proved to be computationally less demanding than the exhaustive investigation of the parameters on a grid approach, giving comparable results [92]. Gradient descent SVM was chosen due to its computational efficiency and good results in terms of classification accuracy. Moreover, the possibility to implement a 1vs1 approach led to the possibility of development of multi-filter Fusion Approach.

The thematic accuracies of the obtained maps (which are presented in Tables 3.1 and 3.2) were assessed by computing the Overall Accuracy (OA), the Average Accuracy (AA) and the Kappa coefficient (K) on the available reference data. *The AA is computed by averaging the values of Producer Accuracy for the thematic classes.* The statistical significance of the classification maps obtained by PCA and ICA and the same morphological processing was evaluated with the McNemar's test. All the results were statistically significant.
The obtained results are reported in Table 3.1. It is clear as, in most of the cases, by including the features extracted by the EAPs in the analysis resulted in higher accuracies (up almost 17% of OA) than those obtained by considering only the spectral features. The ICA proved to extract more informative components from the data, leading to better results than those generated by the PCA in all the experiments. When considering the contribution of the single EAPs, the EAPs built with area and the moment of inertia attributes performed the best with the PCA and ICA, respectively. This proves how it can be difficult to select *a priori* the most suitable attribute on the data. In these experiments, considering all the EAPs together, in the SVA architecture, with the ICA gives excellent results in terms of classification accuracies. At the author best knowledge, these accuracies are higher than all those reported in the literature for this data set without post-processing (see for example [46, 84, 85, 93], where the best results obtained in terms of κ coefficient correspond to 91.48% and in terms of Average Class Accuracy to 94.39%). In contrast, the SVA approach led to low accuracies for the PCA. This can be due to the high variation in terms of accuracy showed by the single EAPs (more than 20% of OA), which affects the overall performances of this approach. The FA is performing well in average and has a robust behavior since in all the experiments the accuracies obtained, when compared to those of the single EAPs, are slightly lower than the best case (less than 2% of OA) and better than all the others. Considering the ICA instead

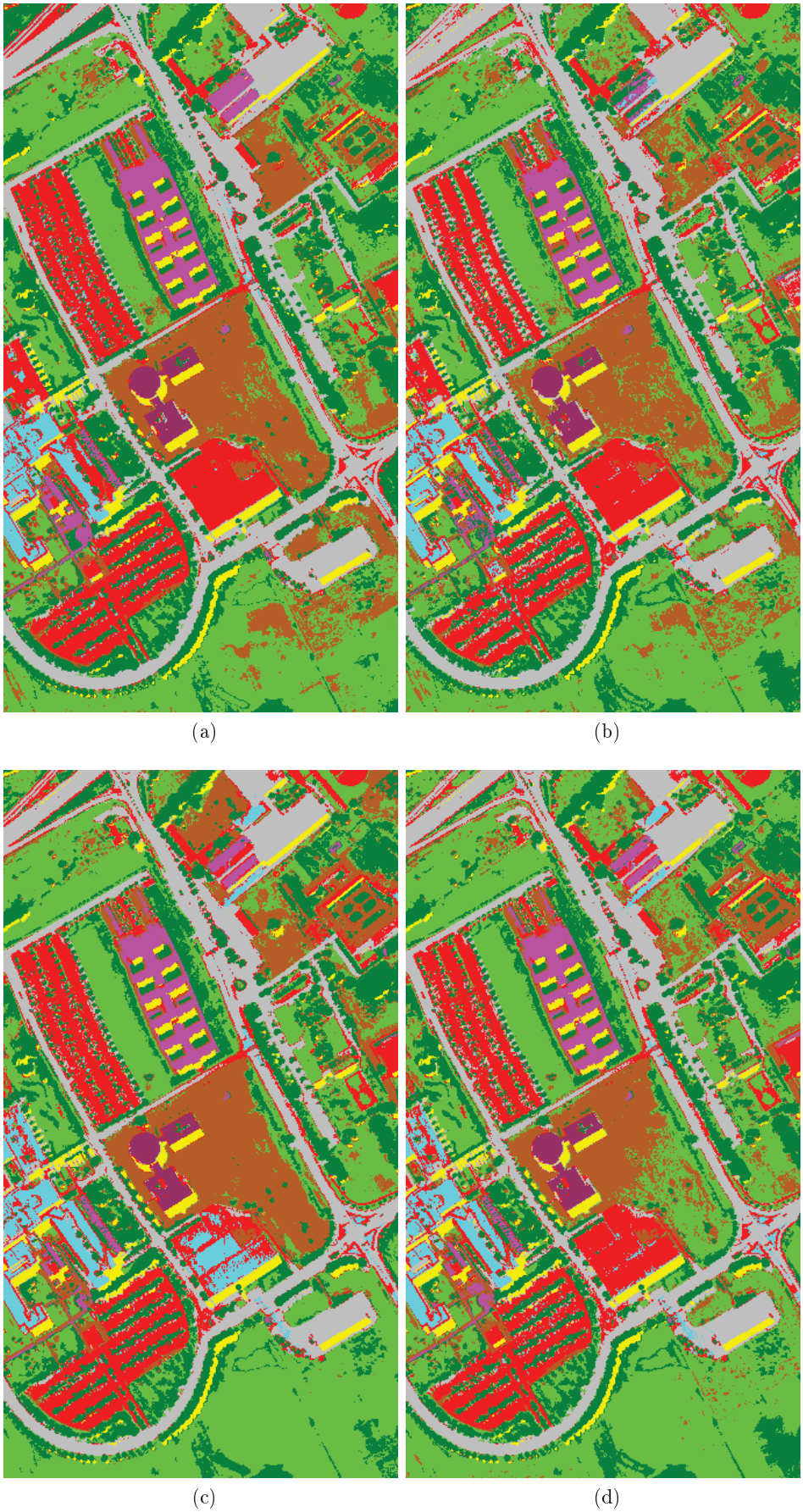
of the PCA led to a significant improvement of the class accuracy up to more than 30% and 20% (class Gravel) for the SVA and FA approaches, respectively. The improved accuracies obtained by the proposed technique are also confirmed by the higher precision shown in the map obtained when considering the ICA and all the EAPs together (see Fig. 3.6c). The results presented in Table 3.1 and Fig. 3.6 were also presented in [89].
Table 3.2 reports the thematic accuracies obtained on the Center data set. Similar considerations as for the University data set can be drawn. For this data set also, it is evident the importance of including the spatial information, which led to an increase in terms of accuracy with respect to considering the original hyperspectral data or the components obtained from the dimensionality reduction technique. The best overall accuracy obtained by using the EAPs, is higher of about 2% than those obtained by the original spectral features and the first components. Considering the PCA and ICA transformations, the latter leads to the best results in most of the cases (except for the single components extracted and for the EAP_s). When looking at the performances obtained by considering the spatial features extracted by the EAPs, one can see that the EAP with area attribute outperformed the other single EAPs with PCA, while when considering the ICA the choice of the standard deviation performed the best among the single EAPs. Moreover, when considering the SVA strategy resulted in the best accuracies with the ICA preprocessing (which is slightly worse than the best EAP). Again, the FA led to results over the average of the accuracies obtained by the single EAPs. Several combinations of filters were also tested. The results, shown in Tables 3.3 and 3.4, suggest that considering only the two or three filters providing the best results do not provide any significant improvement in terms of classification accuracy, if compared to the multi-filter approaches proposed in this work.

3.5 Conclusion

In this chapter, we have combined spectral feature reduction with spatial feature extraction to add contextual information during the classification process and further improve the classification results which can be obtained by using Independent Component Analysis.
In greater details, from the hyperspectral image some independent components are extracted, and different attribute profiles are computed for each one, leading to extended attribute profiles. The features obtained by the morphological processing are then classified with an SVM classifier. Two approaches were proposed for considering the features extracted by the different EAPs, one based on the concatenation of the EAPs and one based on the fusion of the classification results obtained on the single EAPs. The experimental results obtained on real hyperspectral data sets proved that the preprocessing of the hyperspectral data carried out with ICA is more suitable than the PCA for modeling the different sources of information present in the scene. Moreover, from the experiments and results, it was evident how important the spatial features extracted by the EAPs are for classification. The concatenation of the different EAPs gave excellent results in terms of classification accuracies (with respect to other works present in the literature on these data sets). This approach did not perform well only in one case with the PCA, i.e., when the single EAPs led to results significantly different one to the other. (range of difference in the overall accuracies greater than 20%). However, this effect did not occur with the ICA, where the obtained results were more uniform and all statistical significant (according to the McNemar's test). The approach based on the fusion of the classification results with the majority voting strategy proved to have a robust behavior leading to accuracies slightly lower than those of the best case obtained with a single EAPs but better than all the others. The proposed technique proved the importance of including the spatial information in the analysis of high resolution hyperspectral images. Even the increase of the number of features produced by considering the EAPs was not an issue due

Table 3.1: ROSIS University data set. Classification accuracies obtained according to the described scheme. The first 4 columns represent the results of SVM applied to the single EAPs. *All* means that the SVM is applied to the data obtained considering the outputs of the four filters, together. *Fusion* is obtained with the sum of the votes of the first four classifiers. The best results obtained for each dimensionality reduction method are underlined. The best results by comparing PCA vs. ICA are in bold.

	Original	4 Comp.	Area	Diagonal	Inertia	Std	All	Fusion
Feat.	103	4	36	36	36	36	144	(144)
		Principal Component Analysis						
OA (%)	77.89	72.92	90.00	85.42	69.80	86.56	77.81	89.21
κ (%)	72.34	66.25	87.06	81.24	63.22	82.82	71.08	86.06
AA (%)	85.78	81.55	92.04	89.55	82.48	**91.15**	86.84	92.04
Asphalt	78.29%	79.51%	**95.23%**	89.26%	83.32%	**91.57%**	88.34%	**93.45%**
Meadow	67.79%	61.72%	85.91%	80.84%	51.09%	79.91%	60.60%	85.06%
Gravel	60.21%	51.50%	57.12%	62.65%	54.69%	**69.56%**	59.27%	64.32%
Trees	98.47%	97.91%	**99.71%**	**99.22%**	95.46%	**98.89%**	98.76%	**99.41%**
Metal Sheet	99.48%	99.48%	99.70%	99.70%	99.26%	99.70%	99.70%	99.78%
Bare soil	86.68%	73.29%	**92.62%**	**81.05%**	77.73%	**91.85%**	86.56%	**89.76%**
Bitumen	89.70%	81.81%	**99.25%**	97.37%	96.84%	97.82%	98.65%	99.02%
Bricks	91.50%	88.86%	**98.89%**	95.87%	89.84%	91.12%	92.94%	97.61%
Shadow	99.89%	99.89%	**99.89%**	**100%**	94.09%	99.89%	96.73%	**100%**
		Independent Component Analysis						
OA (%)	77.89	74.64	**91.26**	**87.94**	**93.57**	**87.69**	**94.47**	**91.69**
κ (%)	72.34	68.22	**88.55**	**84.31**	**91.63**	**84.14**	**92.80**	**89.13**
AA (%)	85.78	77.18	**92.36**	**91.72**	**95.73**	90.92	**96.58**	**94.11**
Asphalt	78.29%	80.64%	90.53%	**90.54%**	**92.61%**	90.05%	**92.84%**	92.35%
Meadow	67.79%	64.63%	**91.37%**	**85.30%**	**90.96%**	**83.68%**	**91.92%**	**90.42%**
Gravel	60.21%	58.17%	**69.03%**	**85.23%**	**91.09%**	66.89%	**90.00%**	**85.71%**
Trees	98.47%	97.68%	99.58%	98.86%	**98.04%**	97.52%	**99.54%**	99.38%
Metal Sheet	99.48%	99.63%	**99.85%**	**99.85%**	**99.93%**	**99.85%**	**99.93%**	**99.85%**
Bare Soil	86.68%	72.80%	84.91%	74.21%	**95.70%**	88.79%	**97.91%**	83.08%
Bitumen	89.70%	82.93%	98.95%	**97.72%**	**99.32%**	**97.82%**	**99.10%**	**99.02%**
Bricks	91.50%	88.73%	98.75%	**97.80%**	**99.00%**	**94.43%**	**99.19%**	**98.48%**
Shadow	99.89%	99.58%	98.20%	95.78%	**95.56%**	99.37%	**98.84%**	99.47%



(a) (b)

(c) (d)

Figure 3.6: ROSIS Pavia University data set. Classification maps obtained by: (a) PCA with area attribute (EAP_a), (b) PCA with FA, (c) ICA with SVA, and (d) ICA with FA.

to the robustness of the SVM classifier. The results obtained by the proposed method has shown to be the most accurate, when compared to the spectral-spatial classification techniques recently proposed in the literature [46,84,85,93]. Moreover, since the use of attribute filters allows to obtain a representation of the image based on complementary characteristics, there is a high generalization capability in the proposed method, which is expected to show interesting results also in a different context with respect to the one analyzed in this chapter (urban area with very high spatial resolution).

As a matter of fact, spectral feature reduction and spatial feature extraction have proven to be suitable methods to improve the classification accuracy of hyperspectral images. However, a problem which was not address is the influence of low spatial resolution, which lead to the presence of mixtures of materials in the scenes. The second part of this thesis will explore the possibility to use sub-pixel information for a better representation of hyperspectral data affected by such problem.

Table 3.2: ROSIS Center data set. Classification accuracies obtained by averaging 10 trials with 50 training samples per class randomly chosen for each trial. The first 4 columns represent the results of SVM applied to the single EAPs. *All* means that the SVM is applied to the data obtained considering the outputs of the four filters, together. *Fusion* is obtained with the sum of the votes of the first four classifiers.

	Original	4 comp	Area	Diagonal	Inertia	Std	All	Fusion
Feat.	102	4	36	36	36	36	144	(144)
		Principal Component Analysis						
OA (%)	96.25 ± 0.25	96.24 ± 0.24	98.40 ± 0.20	97.83 ± 0.24	97.81 ± 0.25	**98.48**± 0.14	98.43± 0.25	98.39 ± 0.15
κ (%)	93.59 ± 0.42	93.56 ± 0.41	97.26 ± 0.34	96.27 ± 0.41	96.24 ± 0.43	**97.39** ± 0.23	97.31± 0.42	97.24 ± 0.25
AA (%)	92.80 ± 0.38	92.65 ± 0.35	97.18 ± 0.32	96.18 ± 0.42	**96.23** ± 0.33	96.93 ± 0.38	**97.26** ± 0.38	97.21 ± 0.25
Class 1	99.34%	99.53%	99.70%	**99.66%**	99.63%	**99.99%**	99.75%	99.70%
Class 2	85.67%	85.88%	90.46%	89.92%	88.38%	**88.71%**	90.58%	90.50%
Class 3	92.85%	92.50%	94.28%	**94.83%**	**94.41%**	**93.01%**	**95.32%**	95.32%
Class 4	86.83%	86.68%	**99.55%**	97.42%	**98.71%**	**99.57%**	**99.57%**	99.51%
Class 5	91.90%	91.92%	98.59%	**98.31%**	**97.47%**	**99.66%**	98.60%	98.60%
Class 6	93.69%	93.90%	**97.77%**	96.13%	94.87%	**97.11%**	97.14%	97.00%
Class 7	86.66%	84.27%	**94.66%**	89.74%	92.98%	**95.20%**	94.81%	94.74%
Class 8	99.34%	99.24%	99.59%	99.61%	99.61%	**99.19%**	99.62%	99.54%
Class 9	99.92%	99.92%	**99.99%**	**100%**	**99.98%**	99.94%	**99.98%**	**99.98%**
		Independent Component Analysis						
OA (%)	96.25 ± 0.25	96.47 ± 0.24	**98.59** ± 0.16	**98.18** ± 0.22	**97.91** ± 0.32	97.97 ± 0.28	**98.69** ± 0.19	**98.47** ± 0.19
κ (%)	93.59 ± 0.40	93.94 ± 0.41	**97.58** ± 0.26	**96.87** ± 0.38	**96.42** ± 0.54	96.51 ± 0.46	**97.75** ± 0.32	**97.38** ± 0.32
AA (%)	92.80 ± 0.38	92.39 ± 0.34	**97.50** ± 0.25	**96.76** ± 0.27	96.15 ± 0.47	96.18 ± 0.37	**97.58** ± 0.33	**97.26** ± 0.35
Class 1	99.34%	99.81%	**99.74%**	99.66%	**99.64%**	99.74%	**99.82**	**99.73%**
Class 2	85.67%	86.51%	**92.74%**	**91.79%**	**88.94%**	85.31%	**92.15**	**90.58%**
Class 3	92.85%	90.93%	**94.99%**	94.49%	92.53%	91.17%	94.62	**95.35%**
Class 4	86.83%	83.44%	99.32%	**98.02%**	97.36%	99.46%	99.50	**99.56%**
Class 5	91.90%	90.74%	**99.39%**	98.14%	95.64%	98.89%	**98.89%**	**98.60%**
Class 6	93.69%	94.54%	97.33%	**96.62%**	**96.79%**	96.96%	**97.50%**	**97.14%**
Class 7	86.66%	86.44%	94.50%	**92.63%**	**94.84%**	94.55%	**95.96%**	**94.82%**
Class 8	99.34%	99.25%	**99.67%**	**99.67%**	**99.67%**	99.58%	**99.70%**	**99.62%**
Class 9	99.92%	99.87%	99.78%	99.80%	99.97%	**99.97%**	99.97%	**99.98%**

Table 3.3: ROSIS University data set. Classification accuracies obtained with different combination of filters

	Principal Component Analysis		
	Area+Std	Area+Diago	Area+Diago+Std
OA (%)	90.25%	88.87%	89.46%
κ (%)	87.33%	85.65%	86.33%
AA (%)	91.98%	91.58%	91.84%
	Independent Component Analysis		
	Area+Diago	Area+Inertia	Area+Diago+Std
OA (%)	90.86%	95.00%	94.09%
κ (%)	87.98%	93.47%	92.30%
AA (%)	92.86%	96.77%	96.31%

Table 3.4: ROSIS Center data set. Classification accuracies obtained with different combination of filters

	Principal Component Analysis		
	Area+Std	Area+Diago	Area+Diago+Std
OA (%)	98.52 $\pm$ 0.19	98.32$\pm$ 0.21	98.45 $\pm$ 0.21
κ (%)	97.46 $\pm$ 0.30	97.12$\pm$ 0.36	97.34 $\pm$ 0.35
AA (%)	97.24 $\pm$ 0.32	97.10$\pm$ 0.34	97.25 $\pm$ 0.34
	Independent Component Analysis		
	Area+Diago	Area+Std	Area+Diago+Std
OA (%)	98.54 $\pm$ 0.15	98.66 $\pm$ 0.17	98.65 $\pm$ 0.15
κ (%)	97.50 $\pm$ 0.26	97.70 $\pm$ 0.28	97.68 $\pm$ 0.25
AA (%)	97.36 $\pm$ 0.23	97.50 $\pm$ 0.32	97.52 $\pm$ 0.27

Part II

Exploitation of sub-pixel information for dimensionality reduction

Chapter 4

Spectral Unmixing for Dimensionality Reduction

Abstract

In this chapter, new feature reduction chains making use of sub-pixel information are proposed. Feature reduction techniques have proven to be useful to solve the curse of dimensionality and to improve the performances of supervised classification algorithms. However, they do not handle the presence of mixed pixels, that are pixels containing more than a single land cover class, caused by low spatial resolution of satellite images. Mixed pixels are a major issue related to satellite hyperspectral data, and can severely affect classification accuracy. In order to address the problem, the possibility to exploit sub-pixel information provided by spectral unmixing methods is investigated in this chapter.

Contents

4.1 Introduction

The first part of this thesis was devoted to the investigation of classical dimensionality reduction techniques, and the possibility to use them for addressing issues affecting supervised classification of hyperspectral images. The techniques presented in the first three chapters adequately address the curse of dimensionality, and can be usefully proposed to alleviate problems connected to the unfavorable ratio between number of samples used for training a classifier and the dimensionality of the analyzed data.
However, the other major issue related to hyperspectral images, which is the low spatial resolution affecting satellite images, due to physical constraints and to the trade-off which is generally observed between spatial and spectral resolution, is not addressed. In the second part of this thesis, the issue related to low spatial resolution are addressed. Figure 4.1 illustrates the problem connected with low spatial resolution. Every pixel of the image is characterized by a single spectral signature; This signature however is often composed by the mixture of several material spectral signatures [94]. Algorithms which do not consider such a limitation could be severely affected in terms of classification accuracy.
The objective of this chapter is to propose methods performing dimensionality reduction, by exploiting sub-pixel information contained in the image. These methods could provide two main advantages, that are the improvement of discrimination capability when analyzing highly mixed scene, as often happens with medium spatial resolution satellite images, and the possibility to obtain features with a physical meaning.
In many studies, hyperspectral analysis techniques are divided into full-pixel and mixed-pixel classification techniques [14, 93, 94], where each pixel vector defines a *spectral signature* or *fingerprint* that uniquely characterizes the underlying materials at each site in a scene. All the feature reduction/extraction techniques presented in the first part of this thesis are full pixels techniques. Full-pixel classification techniques assume that each spectral signature comprises the response of one single underlying material. Often, however, this is not a realistic assumption [5]. There are several reasons that may cause the appearance of mixed pixels:

- If the spatial resolution of the sensor is not fine enough to separate different pure signature classes at a macroscopic level, several signatures can contribute to the reflectance value of a single pixel, and the resulting spectral signature will be a mixture of the individual pure spectra, often called *endmembers* in hyperspectral imaging terminology [95].

- The Point Spread Function of the sensors, also in the case of advanced technology, makes happen that a pixels will have contribution from its surrounding areas [96].

- When geological or mineral covered scenes are analyzed, there will be a probability of mixtures at a microscopic level, also called intimate mixtures [97].

The presence of mixed pixels is not a minor issue when performing supervised classification [98]. If a pixel is a mixture of two (or more) land cover classes, its statistical characteristics will be different from the classes which it includes. The classification accuracy can be therefore severely affected in this case, if using traditional full pixel methods for dimensionality reduction, as PCA or ICA shown in the first part.
In this chapter, we explore an alternative strategy focused on the use of spectral unmixing for feature extraction prior to classification. The use of spectral unmixing, and sub-pixel scale information, allows to better represent mixtures of classes.
The analysis of sub-pixels information to obtain fractional abundances maps of land cover classes was extensively investigated since the first research carried out on spectral unmixing. Spectral unmixing has been used in a large field of applications, such as cloud screening [99],

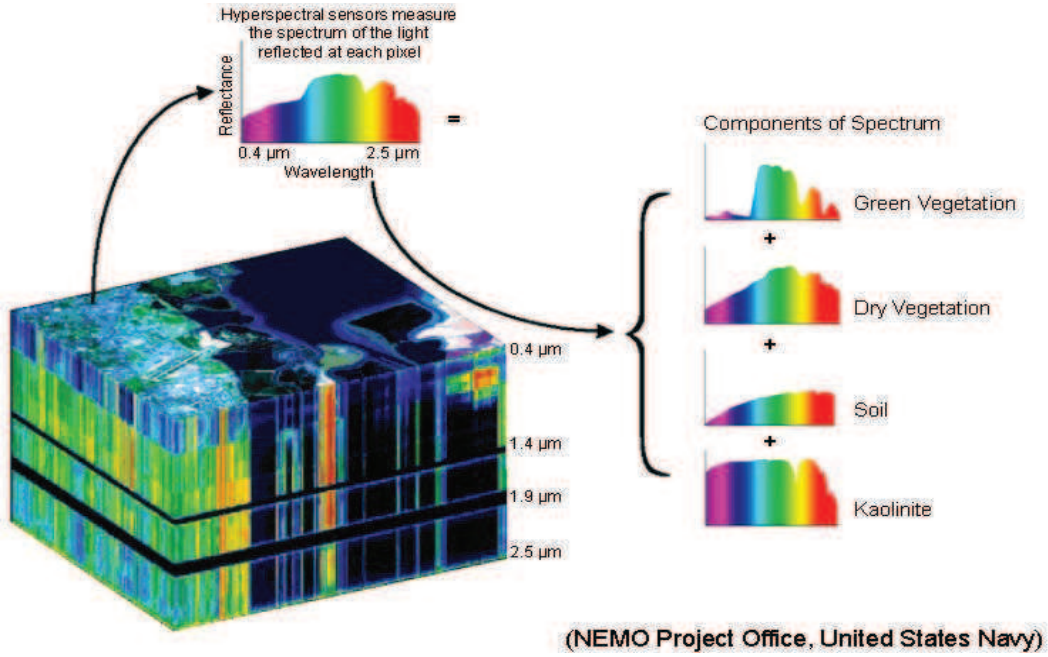


Figure 4.1: Illustration of the mixed pixels problem. When the spatial resolution is not fine fine enough, mixture of materials can be found in the same pixel. Also in case of high spatial resolution, mixed pixels can be found in the image

resolution enhancement of hyperspectral images [100], mapping forest heterogeneity [101] and, more recently, also for hyperspectral image compression [102].
However, the possibility to exploit sub-pixel information to obtain dimensionality reduction techniques and as a possible pre-processing step for supervised classification was only little touched within these investigations. Previous efforts in this direction were presented in [16, 17], but the analysis of whether spectral unmixing can replace standard feature extraction transformations remains an unexplored topic. Although classification techniques often neglect the impact of mixed pixels in the provision of a set of final class labels, widely used benchmark data sets in the literature –e.g. the AVIRIS Indian Pines scene– are known to be dominated by mixed pixels, even if the associated ground-truth information is only available in full-pixel form. Hence, the use of spectral unmixing presents distinctive features with regards to other approaches such as PCA, MNF or ICA. First, it provides additional information for classification in hyperspectral analysis scenarios with moderate spatial resolution, since the sub-pixel composition of training samples can be used as part of the learning process of the classifier. Second, the components estimated by spectral unmixing can be physically explained as the abundances of spectral endmembers. Third, spectral unmixing does not penalize classes which are not relevant in terms of variance or SNR. Here, starting from the initial background idea, we design different unmixing processing chains with the goal of addressing three specific research questions:

1. Is spectral unmixing a feasible strategy for feature extraction prior to classification?

2. Does the inclusion of spatial information at the endmember extraction stage lead to better classification results?

3. Is it really necessary to estimate pure spectral endmembers for classification purposes?

The objective of this chapter is to adequately answer to the above mentioned questions. The remainder of this chapter is structured as follows. Section 4.2 provides an overall in-

troduction about spectral unmixing. Section 4.3 describes the considered spectral unmixing chains. Section 4.4 presents different experiments specifically designed to address the research questions above and to provide a comparison between the proposed unmixing-based strategy and other feature extraction approaches presented in the literature. Section 4.5 concludes with some remarks and future research avenues.

4.2 Spectral Unmixing

Spectral unmixing is the procedure by which the measured spectrum of a mixed pixel is decomposed into a collection of constituent spectra, or endmembers, and a set of corresponding fractions, or abundances, that indicate the proportion of each endmember present in the pixel. Endmembers normally correspond to familiar macroscopic objects in the scene, such as water, soil, metal, or any natural or man-made material. Unmixing provides a capability that is important in numerous tactical scenarios in which subpixel detail is valuable.
Mixing models attempt to represent the underlying physics that are the foundation of hyperspectral phenomenology, and unmixing algorithms use these models to perform the inverse operation, attempting to recover the endmembers and their associated fractional abundances from the mixed-pixel spectrum.
Two main models can be used to descripe a hyperspectral scene by mean of spectral unmixing. The first possibility is the *linear unmixing*, which assumes that the different land cover classes have linear relationship and their reflectance is directly proportional to the area occupied within the pixel. The second possibility is to use a non-linear model, assuming that the materials are not lying side by side, but they are homogeneously distributed within the pixels and mixed at a microscopic level.
In this work, the focus is posed on the linear mixing model. When performing supervised classification, macro-land cover classes are analyzed. In such a case, the linear spectral unmixing approximation can be used without loss of important information. Moreover, the non linear spectral model is much more expensive from a computational point of view. In the following, after describing from a mathematical point of view the linear unmixing problem, a general overview of the different methods proposed to retrieve endmembers and estimate their abundances is provided.

4.2.1 Mathematical definition

Let us denote a remotely sensed hyperspectral scene with n bands by $\mathbf{I}$, in which each pixel is represented by a vector $\mathbf{X} = [x_1, x_2, \cdots, x_n] \in \Re^n$, where $\Re$ denotes the set of real numbers in which the pixel's spectral response x_k at sensor channels $k = 1, \ldots, n$ is included. Under the linear mixture model assumption, each pixel vector can be modeled using:

$$\mathbf{X} = \sum_{z=1}^{p} \Phi_z \cdot \mathbf{E}_z + \mathbf{n}, \tag{4.1}$$

where $\mathbf{E}_z$ denotes the spectral response of endmember z, Φ_z is a scalar value designating the fractional abundance of the endmember z at the pixel $\mathbf{X}$, p is the total number of endmembers, and $\mathbf{n}$ is a noise vector. Two physical constrains can be imposed into the model described in (6.1) [103]:

$$\Phi_z \geq 0 \tag{4.2}$$

$$\sum_{z=1}^{p} \Phi_z = 1. \tag{4.3}$$

The first equation represent the abundance non-negativity constraint, i.e., no negative abundances are possible, while the second one is the abundance sum-to-one constraint, i.e., the sum of all the materials fractions within a pixel must have the unity as result of the sum.

4.2.2 Endmember extraction

When performing spectral unmixing, there are two key tasks: endmember determination and abundances quantification. The first task in linear spectral mixture analysis is to find an appropriate suite of pure spectral signatures (endmembers), which are then used to model at-sensor pixel spectra through a linear combination of endmember signatures. The selection of endmembers can be performed in two ways: 1) by deriving them directly from the image (image endmembers) or 2) from field or laboratory spectra of known target materials (library endmembers); see [104] for a comparison between the two. The risk in using library endmembers is that these spectra are rarely acquired under the same conditions as the airborne data. Image endmembers have the advantage of being collected at the same scale as the data and can, thus, be more easily associated with features on the scene [94]. A number of algorithms have been developed over the past decade to accomplish the task of finding appropriate image endmembers for spectral mixture analysis. It should be taken into account that the presence of pure class pixels in the image data depends on available sensor spatial resolution. As a result, there may be cases where it is not possible for a certain algorithm to find such pure pixels in a scene. In those situations, the fractional components found for the mixed pixels are usually expressed in terms of other mixed pixels (the endmembers identified by the algorithm) and not in terms of pure classes.
The data sets analyzed in this work are common benchmarks for supervised classification. These images are composed by both pure and mixed pixels, also in case of highly mixed scenes, like the AVIRIS Indian Pine data set, and pixels occupied by a single land cover class can be easily found. For such a reason, the use of algorithms which search for the pure components directly within the image were preferred to the use of spectral libraries.
During the last decade, several algorithms have been proposed for the purpose of autonomous/ supervised endmember selection from hyperspectral scenes. Besides the large number of techniques proposed, there are two main groups of algorithms: geometrical and statistical based algorithms. Geometric endmember determination techniques exploit the parallelism between mixing models and the geometric orientation of hyperspectral data in multidimensional spaces. Starting from the seminal work of Craig (minimum volume transform or MVT [105]), try to find the minimum-volume simplex, i.e. the one that embraces the data cloud as tightly as possible, a great number of techniques were proposed in the last years, such as N-FINDR, Vertex Component Analysis, Orthogonal Subspace Projection. If a spectral unmixing algorithm processes a mixed pixel by using statistical representations, then the algorithm is statistical. The representations can be analytical expressions that represent probability density function (parametric). An example of this family is the stochastic mixing model [106], in which each endmember distribution has Gaussian form. Also in this case, a large number of algorithms were proposed. A detailed description of these algorithms can be found in [107,108]. However, also in these mentioned remarkable works, the impossibility to have a very detailed comparison of all these methods was pointed out, because of the huge

number of algorithms proposed in the last years and of the difficulty to evaluate the methods on several data sets, because of the lack of knowledge about the scene at a sub-pixel level. For such a reason, in this chapter several techniques were tested. The geometric based techniques were preferred, due to their good behavior in terms of computational burden.

4.2.3 Abundances determination

Once the endmembers are extracted from the image, the abundance fractions of the elements within each pixel should be determined. In the following, we will briefly review the main technique proposed for abundances determination.

4.2.3.1 Linear Spectral Unmixing

The easiest way to estimate the abundances within a remote sensing image is through a simple inversion. If we consider the noise vector $\mathbf{n}$ equal to zero in Equation 4.1, the abundances of each endmember spectrum can be easily found once the matrix $\mathbf{E}_z$ is known. This technique is optimal for the computational burden, since it does not need any optimization step. However, it does not assure the physical constrains 4.2. When using unconstrained linear spectral unmixing, abundances lower than zeros or greater than unity will be likely to appear in the results.

4.2.3.2 Fully Constrained Linear Spectral Unmixing

Several algorithms have been developed to handle the linear mixing model according with the required physical constrains of abundance fractions, which are non negativity (all the abundances must be greater than or equal to zero) and full additivity (the sum of the endmember abundances within a pixel should be equal to one). Due to the efficiency from a computational point of view, a common choice is to use a fully constrained least squares (FCLS) algorithm, which satisfies both abundance constraints and is optimal in terms of least squares error [109]. In concrete processing, the main problem is that the FCLS does not have a closed-form mathematical solution due to the nonnegativity constraints; thus, a numerical solution is always required. To calculate the FCLS solution, the non negativity constraint is considered first. The idea is to minimize the LSE by estimating the nonnegative abundance values, which is mathematically expressed as

$$\min(\mathbf{z} - \mathbf{Sa})^{\mathbf{T}}(\mathbf{z} - \mathbf{Sa}). \tag{4.4}$$

By using Lagrange multipliers, a Lagrangian J is defined as

$$J = \frac{1}{2}(\mathbf{z} - \mathbf{Sa})^{\mathrm{T}}(\mathbf{z} - \mathbf{Sa}) + \lambda^{\mathrm{T}}(\mathbf{a} - \mathbf{c}) \tag{4.5}$$

where $\mathbf{a} = \mathbf{c}$, each member of the unknown constant L×1 vector $\mathbf{c}$ is nonnegative to enforce the non negativity constraint, and λ is the Lagrange multiplier denoted by an L×1 vector. The defined equation 4.4 allows the use of Lagrange multipliers because the non negativity constraint has been substituted by equality constraints with the unknown vector $\mathbf{c}$. To calculate the estimate of $\mathbf{a}$, we take the partial derivative of J with respect to $\mathbf{a}$. Equation 4.5 contains two unknown parameters, *i.e.*, the abundance estimates and the Lagrange multipliers. Solving for these unknown parameters results in

$$\hat{\mathbf{a}} = (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}S^{T}\mathbf{z} - (\mathbf{S}^{T}\mathbf{S})^{-1}\lambda \tag{4.6}$$

$$\lambda = \mathbf{S}^{\mathrm{T}}(\mathbf{z} - \mathbf{S}\hat{\mathbf{a}}). \tag{4.7}$$

By iterating through the equations 4.6, the numerical solution is provided for the non negativity constraint. To begin this iterative method, we set all the Lagrange multipliers to zero and calculate the abundance. Note that this initial calculation is the unconstrained least squares solution for the abundance values. From this solution, we find those abundance values which are greater than zero and put them into the passive set P. The remaining non positive abundance values are placed in the active set R. Equations 4.6 are iterated until all Lagrange multipliers in the passive set are zero and all Lagrange multipliers in the active set are either zero or negative. At this point, the Kuhn-Tucker conditions have been met, and an optimal solution for the abundance values has been found [110]. It should be noted that this solution only accounts for the non negativity constraint. To handle the sum-to-one constraint, an easy modification of the aforementioned algorithm was developed to retain the optimality guaranteed under the Kuhn-Tucker conditions for numerical optimization on a finite computing machine. In the modification, the endmember matrix and pixel signatures are extended such that

$$\widetilde{\mathbf{S}} = \begin{bmatrix} \delta\mathbf{S} \\ 1^{T} \end{bmatrix} \tag{4.8}$$

is the new endmember matrix and

$$\widetilde{\mathbf{a}} = \begin{bmatrix} \delta\mathbf{z} \\ 1 \end{bmatrix} \tag{4.9}$$

is the new pixel signature, where δ is a constant (typically, 1×10^{-5}). The variable δ controls how tightly the solution will sum to one so that the smaller values provide a better solution, but it may need a longer convergence time. The new endmember matrix and pixel signature are then used in 4.8 and 4.9 to obtain an abundance solution that subjects to both the non negativity and sum-to-one constraints simultaneously.
The solution obtained by the FCLS algorithm is the optimal one. By the FCLS algorithm, the abundance of each endmember in each pixel can be obtained.

4.2.3.3 Partial Unmixing

Linear Spectral Unmixing and Fully Constrained LSU are the most widely used techniques to determine endmember abundances. However, in case not all the endmembers contained within a data set are recovered, or if the set of pure spectral signatures is different from the set of interest, they could lead to misleading results. In such a case, partial unmixing, which tries to determine abundances of each of the endmembers separately, could be more appropriate. In this work, we will consider the Mixture Tuned Match Filter (MTMF), also known as Constrained Energy Minimization (CEM). CEM is one of a number of inner product measures that are used with hyperspectral data. That is to say, it calculates a weighted sum, η say, of the components of a multivariate, remotely sensed signal

$$\nu = w_1x_1 + w_2x_2 + \ldots + w_nx_n = \mathbf{w^{T}x} \tag{4.10}$$

where $\mathbf{x}$ is the n-dimensional signal of a multispectral pixel. Unconstrained spectral unmixing (i.e., where we do not require the sum of the estimated cover fractions to add to

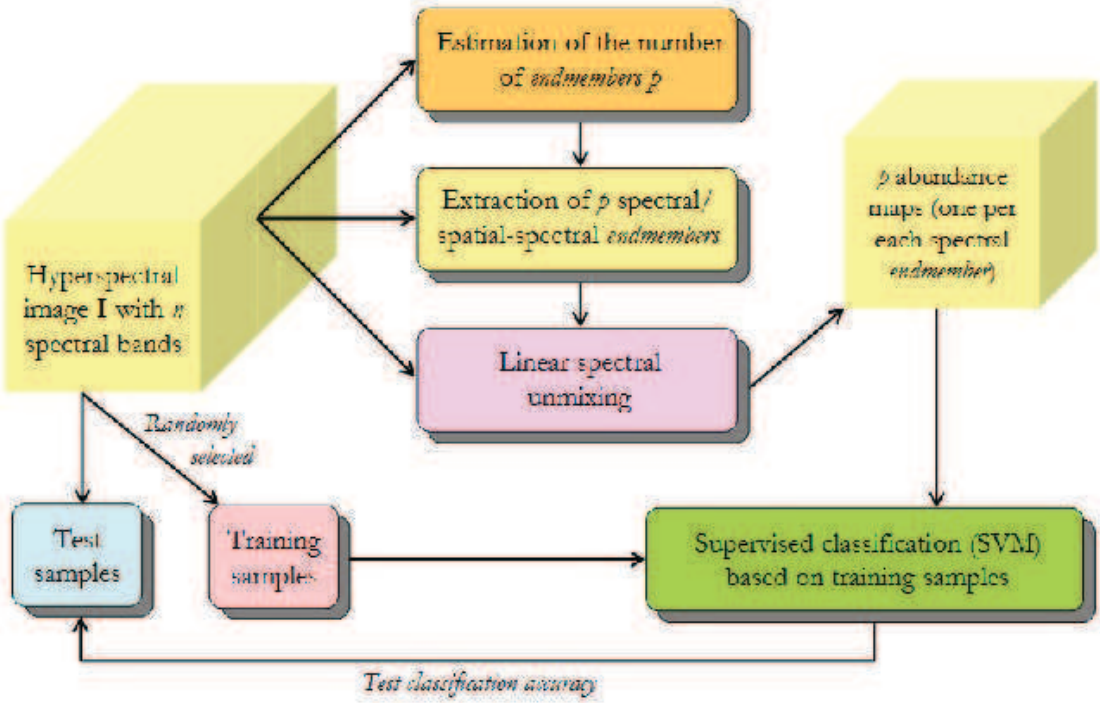


Figure 4.2: Unmixing-based feature extraction chains #1 (spectral endmembers) and #2 (spatial-spectral endmembers).

unity) is another example of such a measure. If a constant term is included, we have a slightly more general measure to characterize surface cover,

$$\eta = \eta_0 + \mathbf{w^T x} \tag{4.11}$$

Unit-sum constrained spectral unmixing has this form [111], as does abundance estimation based on regression of fractional abundance on the signal [112]. In the CEM matched filter, the weight vector $\mathbf{w}$ is given by

$$\mathbf{w} = (d^T R^{-1} d)^{-1} d \tag{4.12}$$

where d is a particular signal of interest and R is the matrix

$$R = \frac{1}{m} \sum_{1}^{m} \mathbf{x}\mathbf{x}^T \tag{4.13}$$

$(x_1, x_2, \ldots, x_m)$ is the complete set of spectral signals for the image, m pixels in all. (Henceforth, brackets $\{\}$ will be used to denote an average over all image pixels, so that for instance $R = \{xx^T\}$) i). The CEM minimizes the mean square value of η while forcing it to take the value 1 for pixels matching the signal of interest. The idea is that its values are generally as small as possible, except for pixels where the constituent of interest dominates. The treatment of this method by Nielsen [113] is particularly elegant. The studies in [114] have shown that the resulting value of η can correlate well with the proportion of the cover type present. Thus, it not only takes a low value when the cover type is absent, and the value of 1 when it covers a pixel, but the intermediate values are monotonically related to the true cover proportions. In [113] examples are given on AVIRIS data of "abundance" images obtained after scaling the CEM results from 0 to 1. The formulation provided by Settle in [115] address the two important questions of how well does the CEM approach correspond with the fractional abundance of the cover type in a general mixed pixel, and to what extent can CEM be used

to provide a true partial unmixing of a general mixed pixel, rather than an ad hoc measure of similarity, deriving conditions for this to occur.

4.3 Unmixing-based Feature Extraction

As pointed out in the previous sections, the idea to exploit sub-pixel information through spectral unmixing seems to be a very interesting possibility as an alternative dimensionality reduction method. However, several major issues which should be carefully addressed, hold in the framework. First and probably most important, the choice of the algorithms to be used to realize the different chains. Section 4.2 showed presented a large number of techniques proposed both for retrieve the pure spectral compoents of the image, and to estimate their fractional abundances. Another very important issue is the number of features which need to be computed: On one hand, several techniques were proposed in the last years to estimate the dimensionality of hyperspectral data, which could be used to solve the problem. On the other hand, if a supervised classification is to be performed, there is also the possibility to take advantage of the prior information about land cover classes, and their number.
With these issues in mind, we have explored in this work a number of different feature reduction chains based on spectral unmixing, which could adequately address the problems just mentioned.

4.3.1 Unmixing Chain #1

In this subsection we describe our first approach to design an unmixing-based feature extraction chain which can be summarized by the flowchart in Fig. 4.2. First, we estimate the number of endmembers, p, directly from the original n-dimensional hyperspectral image $\mathbf{I}$. For this purpose, we use in this work two standard techniques widely used in the literature such as the HySime method [116] and the virtual dimensionality (VD) concept [117]. Once the number of endmembers p has been estimated, we apply an automatic algorithm to extract a set of endmembers from the original hyperspectral image [118]. Here, we use an orthogonal subspace projection (OSP) technique [119] which has been shown in previous works to provide a very good trade-off between the signature purity of the extracted endmembers and the computational time to obtain them. Preliminary experiments conducted with other endmember extraction techniques, such as Vertex Component Analysis (VCA) [120] and N-FINDR [121], have shown very similar results in terms of classification accuracy. Finally, linear spectral unmixing (either unconstrained or constrained) can be used to estimate the abundance of each endmember in each pixel of the scene, providing a set of p abundance maps that are used to train a standard SVM classifier.

4.3.2 Unmixing Chain #2

In this subsection we introduce a variation of the unmixing-based feature extraction chain which includes spatial preprocessing prior to endmember extraction in order to guide the endmember searching process to those areas which are more spatially homogeneous. This approach is represented in Fig. 4.2. The method computes, for each pixel vector, a scalar spatially derived factor that relates to the spectral similarity of pixels lying within a certain spatial neighborhood. This scalar value is then used to weigh the importance of the spectral information associated to each pixel in terms of its spatial context. Two basics of the method are that: i) no modification of existing image spectral-based endmember extraction methods is necessary in order to apply the proposed approach. ii) The preprocessing method enhances the search for image spectral endmembers in spatially homogeneous areas. A complete description of the spatial pre-processing strategy can be found in [122]. As in the previous

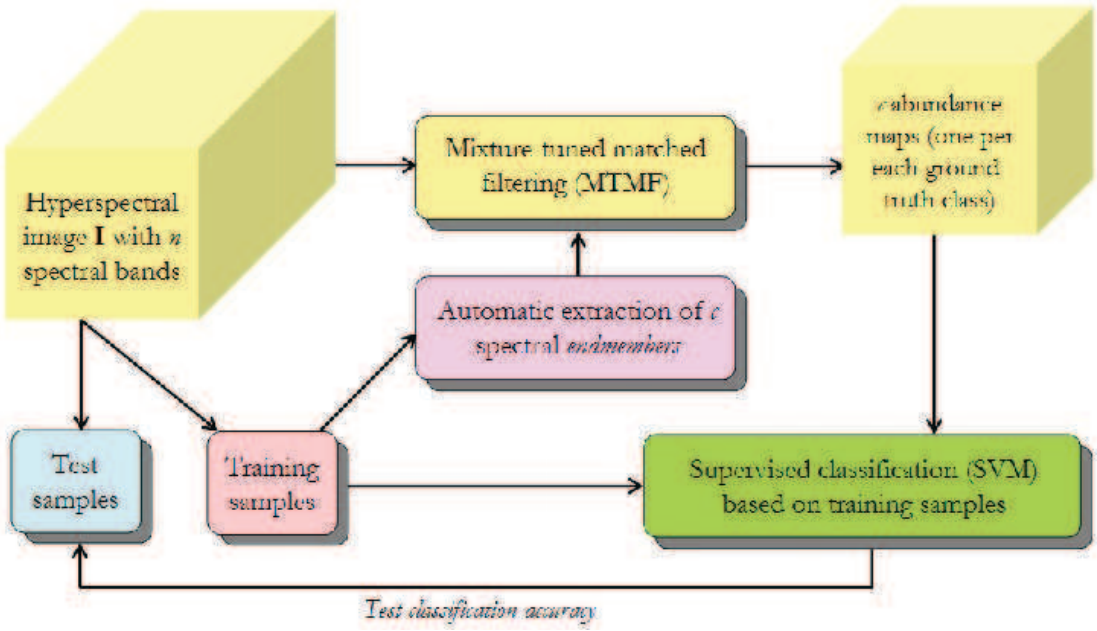


Figure 4.3: Unmixing-based feature extraction chain #3.

chain, the features resulting from the proposed (spatially enhanced) unmixing process are used to train an SVM classifier with a few randomly selected labeled samples. The classifier is then tested using the remaining labeled samples.

4.3.3 Unmixing Chain #3

Our main motivation for introducing a third unmixing-based feature extraction chain is the fact that the estimation of the number of endmembers p in the original image is a very challenging issue. Fig. 4.3 describes a new chain in which the endmembers are extracted from the set of available (labeled) training samples instead of from the original image. This chain introduces two important variations: 1) first, as a simplification to the challenging estimation problem, the number of endmembers to be extracted is set as the total number of different classes, c, in the training set; and 2) the endmember searching process is conducted only on the training set, which reduces computational complexity. However, the number of endmembers in the original image, p, is probably different than c, the number of labeled classes. Therefore, in order to unmix the original image we need to address a *partial unmixing* problem (in which not all endmembers may be available *a priori*). In this work it was considered the constrained energy minimization (CEM) approach [123], which combines linear spectral unmixing and statistical matched filtering. From matched filtering, it inherits the ability to map a single known target without knowing the other background endmember signatures. From spectral mixture modeling, it inherits the leverage arising from the mixed pixel model and the constraints on feasibility.

4.3.4 Unmixing Chain #4

The fourth unmixing chain tested in our experiments represents a slight variation of the unmixing chain #3 in which the spectral signatures used for unmixing purposes are not obtained via endmember extraction but through averaging of the spectral signatures associated to each labeled class in the training set. To keep the number of estimated components low, only one component is allowed for each class. This averaging strategy produces c signatures,

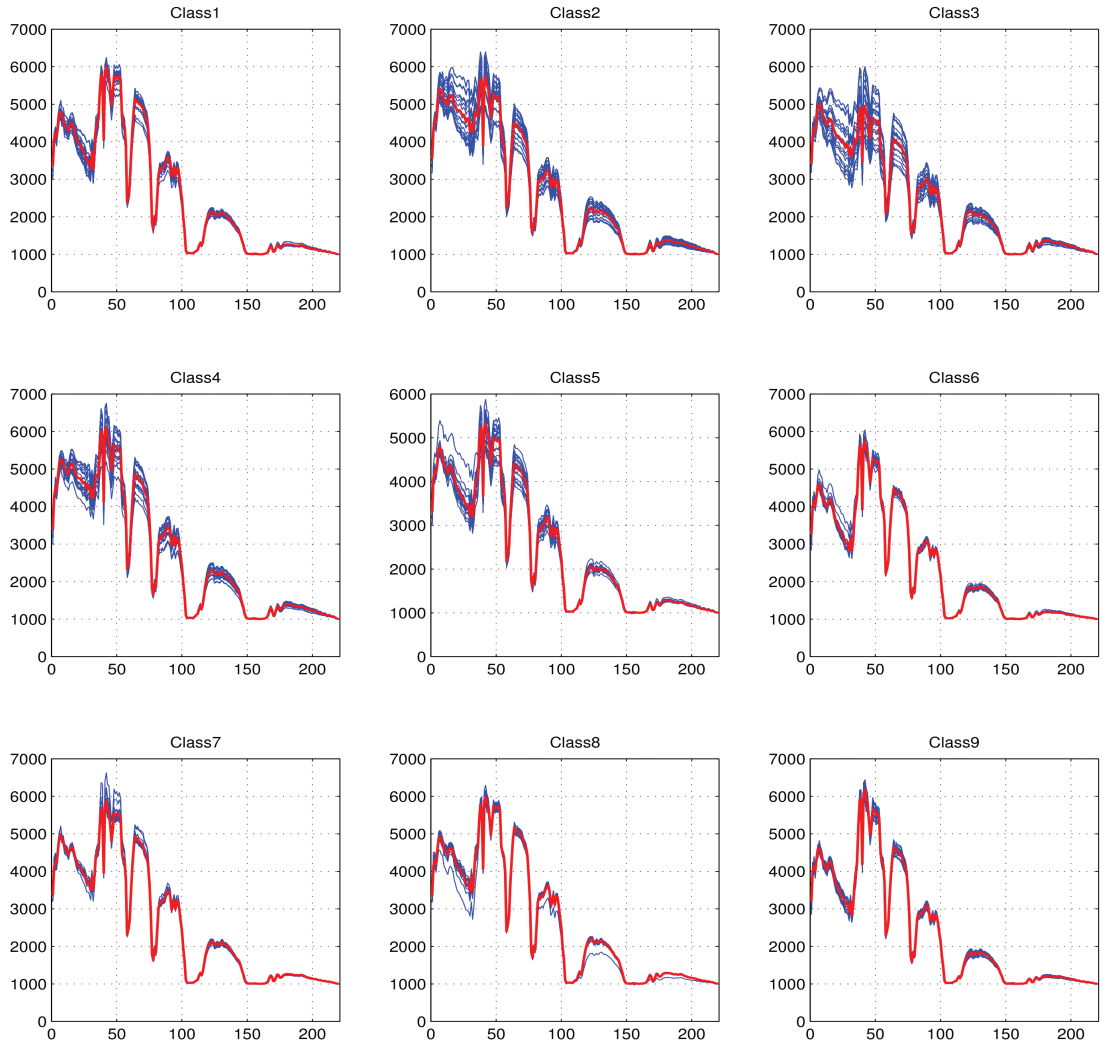


Figure 4.4: Example of averaged training spectra, Indian Pine data set, Classes 1 to 9. Blue lines: training samples. Red line: average spectrum.

each representative of a labeled class, which are then used to partially unmix the original hyperspectral scene using MTMF. Fig. 4.4 shows an example of spectra extracted from the Indian Pine data set.

4.4 Experimental Results

The experimental tests were conducted on two real hyperspectral data with medium spatial resolution. The four unmixing chains proposed in Section 4.3 have been compared with three classical algorithms for dimensionality reduction, PCA, ICA and MNF.

4.4.1 Hyperspectral Data

The first data set used in our experiments was collected by the AVIRIS sensor over the Indian Pines region. The scene[1] comprises 145 lines by 145 samples and 220 spectral

[1] Available online: http://dynamo.ecn.purdue.edu/biehl/MultiSpec

channels in the wavelength range from 0.4 to 2.5 μm, nominal spectral resolution of 10 nm, and spatial resolution of 20 meters by pixel. After removing noisy and water absorption bands, 202 channels were left. The ground-truth contains 16 land cover classes. The number of pixels in the smallest class is 20, while the number of pixels in the largest class is 2468.

The second data set was collected by the AVIRIS sensor over the Kennedy Space Center[2], Florida, on March 1996. The portion of this scene used in our experiments has dimensions of 292×383 pixels. After removing water absorption and low SNR bands, 176 bands were used for the analysis. The spatial resolution is 20 meters by pixel. 12 ground-truth classes where available, where the number of pixels in the smallest class is 134 while the number of pixels in the largest class is 761.

4.4.2 Experiments

Experiment 1. Use of unmixing as a feature extraction strategy

In this experiment, we use the AVIRIS Indian Pines and Kennedy Space Center data sets to analyze the impact of imposing non-negativity and sum-to-one constraints in abundance estimation prior to classification. For the AVIRIS Indian Pines image, we construct ten small training sets by randomly selecting 5%, 10% and 15% of the ground-truth pixels. For the AVIRIS Kennedy Space Center, since the size of the smaller classes is bigger, we decided to reduce the training sets even more and selected 1%, 3% and 5% of the available ground-truth pixels. Then, the three considered types of input features (original, reduced and unmixing-based) are built for the selected training samples and used to train an SVM classifier in which two types of kernels: polynomial and Gaussian are used. The SVM was trained with each of these training subsets and then evaluated with the remaining test set. Each experiment was repeated ten times, and the mean and standard deviation accuracy values were reported. Kernel parameters were optimized by a grid search procedure, and the optimal parameters were selected using 10-fold cross-validation. The LIBSVM library[3] was used for experiments.

Table 4.1 summarizes the overall classification accuracies obtained after applying the considered SVM classification system (with polynomial and Gaussian kernels) to the features extracted after applying the unmixing chain #1 (see Fig. 4.2) to the AVIRIS scenes. The dimensionality of the input data, as estimated by a consensus between the HySime and the VD concept, was $p = 18$ for the Indian Pines scene, and $p = 15$ for the Kennedy Space Center scene. The chain #1 was implemented using two different linear spectral unmixing algorithms [103]: unconstrained and fully constrained; due to better accuracy and faster computation, only results for the unconstrained case are presented. The better results obtained with the unconstrained linear spectral unmixing can be explained with the lower number of constrains imposed to the components. Since the sum of all the features values for a single pixel should be equal to one, the information of a feature is somehow dependent from the information provided by the other components, thus decreasing the overall amount of useful data. The results after applying the classification system to the original spectral features, and to those extracted using traditional unsupervised methods proposed in the literature (PCA, MNF and ICA) are also reported.

As shown by Table 4.1, the classification accuracy is correlated with the training set size (the larger the training set, the higher the classification accuracy). The good generalization ability exhibited by SVMs is demonstrated by the classification results reported for the original spectral information, even with very limited training sets. The fact that MNF is more effective than PCA and ICA for feature extraction purposes is also remarkable, since

[2]Available online: http://www.csr.utexas.edu/hyperspectral/data/KSC/

[3]http://www.csie.ntu.edu.tw/cjlin/libsvm/

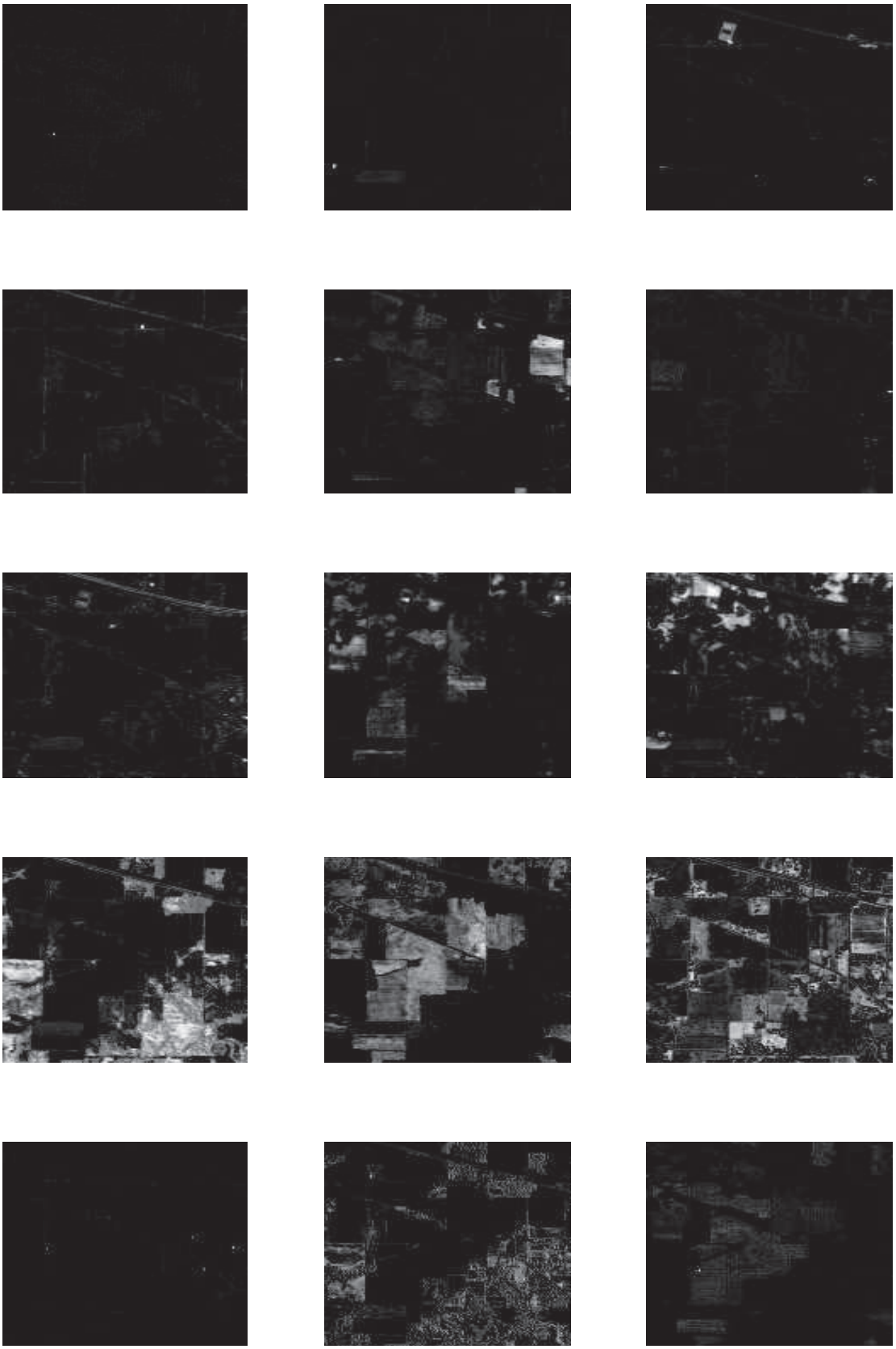


Figure 4.5: AVIRIS Indian Pine data set. First fifteen components extracted by using the chain #1, fully constrained estimation of the abundances.

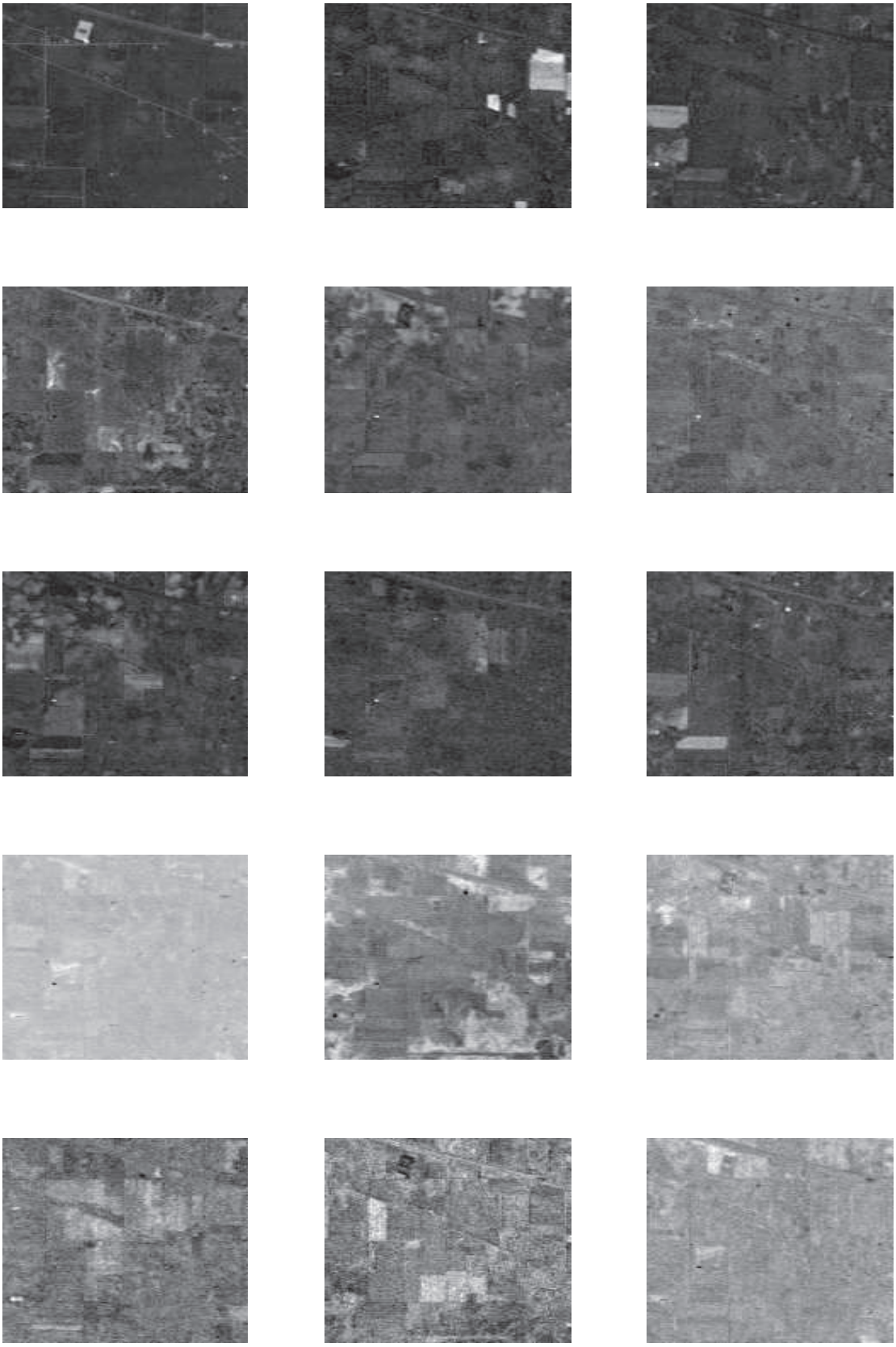


Figure 4.6: AVIRIS Indian Pine data set. First fifteen components extracted by using the chain #4, partial unmixing estimation of the abundances.

Table 4.1: Classification accuracies (percentage) and standard deviation obtained after applying the considered SVM classification system (with Gaussian and polynomial kernels) to three different types of features (original, reduced and unmixing-based) extracted from the AVIRIS Indian Pines and Kennedy Space Center scenes(ten randomly chosen training sets)

Image	**AVIRIS Indian Pines**						
Type of	# of	**Polynomial kernel**			**Gaussian kernel**		
feature	feat.	5%	10%	15%	5%	10%	15%
Original	202	75.23±1.23	81.55±0.86	83.58±0.78	76.85±0.45	83.58±0.35	86.72±0.45
PCA	18	77.07±1.46	81.66±0.88	83.11±0.52	78.01±1.18	82.94±0.51	84.65±0.79
MNF	18	82.97±1.93	87.41±0.31	88.38±0.57	84.73±0.77	88.72±0.45	90.69±0.29
ICA	18	76.63±1.27	81.00±0.71	82.94±0.36	76.92±0.72	81.27±0.61	82.95±0.71
Ch. #1	18	74.56±1.04	79.20±1.12	80.97±0.50	75.58±0.85	80.79±0.36	82.31±0.59
Ch. #2	18	71.93±0.96	77.58±0.92	79.31±0.33	73.25±1.00	78.68±0.63	81.07±0.30
Ch. #3	15	81.32±0.84	85.56±0.84	86.83±0.55	82.52±0.56	87.20±0.61	88.83±0.63
Ch. #4	15	82.36±1.09	86.87±0.59	87.97±0.57	83.38±1.05	88.62±0.47	90.25±0.64
Image	**AVIRIS Kennedy Space Center**						
Type of	# of	**Polynomial kernel**			**Gaussian kernel**		
feature	feat.	1%	3%	5%	1%	3%	5%
Original	176	70.97±3.32	82.53±1.63	85.71±1.40	72.26±2.42	82.91±1.38	85.50±1.35
PCA	15	73.52±3.69	83.26±1.26	86.11±1.16	74.66±2.94	82.54±1.70	86.28±1.46
MNF	15	77.01±3.77	86.85±2.19	89.59±1.89	77.94±3.48	87.43±2.11	90.01±1.52
ICA	15	70.09±2.91	80.28±1.73	84.59±1.50	70.39±1.58	80.79±1.60	84.58±1.58
Ch. #1	15	69.41±2.64	78.62±1.58	82.84±1.17	69.02±5.40	79.08±1.46	83.53±1.25
Ch. #2	15	67.91±3.98	78.61±3.56	84.26±1.41	68.56±4.70	83.86±1.89	83.86±1.22
Ch. #3	12	74.28±3.23	85.37±1.30	87.88±1.57	75.02±4.13	84.92±1.97	88.47±1.38
Ch. #4	12	76.10±2.49	86.38±1.40	87.84±1.28	77.53±2.58	86.57±0.97	87.72±1.13

Table 4.2: Statistical differences evaluated using McNemar's test (polynomial kernel). The table presents the value of Z: Differences are considered as significant at 95% of confidence if $|Z|>1.96$.

	Indian Pine		Kennedy SC	
	5%		**1%**	
	Chain #3	Chain #4	Chain #3	Chain #4
PCA	-9.52 (0/0/10)	-11.88 (0/0/10)	-2.09 (1/2/7)	-5.73 (0/2/8)
MNF	5.22 (8/1/1)	2.05 (6/2/2)	3.08 (6/3/1)	-0.58 (3/4/3)
ICA	-10.45 (0/0/10)	-12.85 (0/0/10)	-7.14(0/0/10)	-8.82(0/0/10)
	10%		**3%**	
	Chain #3	Chain #4	Chain #3	Chain #4
PCA	-9.24 (0/0/10)	-12.40 (0/0/10)	-2.61(1/2/7)	-5.14 (0/0/10)
MNF	6.13 (10/0/0)	1.72 (5/4/1)	3.48 (6/2/2)	0.78 (5/1/4)
ICA	-17.37(0/0/10)	-20.26(0/0/10)	-7.14 (0/0/10)	-8.82 (0/0/10)
	15%		**5%**	
	Chain #3	Chain #4	Chain #3	Chain #4
PCA	-8.86(0/0/10)	-11.40 (0/0/10)	-3.78 (0/1/9)	-2.56 (0/4/6)
MNF	5.23 (10/0/0)	1.35 (3/7/0)	2.42 (6/2/2)	4.04 (7/3/0)
ICA	-9.28 (0/0/10)	-11.87 (0/0/10)	-5.08 (0/1/9)	-5.29 (0/0/10)

the MNF has been more widely used in the context of spectral unmixing rather than classification (e.g. as a dimensionality reduction method prior to application of an endmember extraction method). Most importantly, Table 4.1 also reveals that the use of unmixing chain #1 as feature extraction strategy cannot improve the classification results provided by PCA, MNF, ICA or the original spectral information. This is because endmember extraction is generally sensitive to outliers and anomalies, hence a strategy for directing the endmember searching process to spatially homogeneous areas could improve the final classification results. The influence of spectral anomalies on the chain #1 can be seen by looking at Fig. 4.5. A number of components (which represents the fractional abundances extracted with OSP) provides useful information about a small part of the image, while neglecting all the rest. This is caused by the fact that OSP estimate as endmember a pixel with an extreme value of reflectance, which could be caused by a problem of the spectrometer or by an object with very high reflectance power. The abundance of such a pixel will be maximum only in the pixel where this anomaly is located, while does not provide any information about the rest of the image (it can be noticed that at least four components of the overall 18 estimated in this experiment for the AVIRIS data set suffer of this drawback, causing a major loss of information for the supervised classification).

Experiment 2. Impact of including spatial information at the endmember extraction stage

In this experiment we apply the unmixing chain #2 for feature extraction prior to classification. As shown by Table 4.1, spatial preprocessing prior to endmember extraction can not

lead to improved classification results with regards to the chain #1 and the original spectral information. This is due to the spectral similarity of the most spatially representative classes in our considered scenes. For instance, in the AVIRIS Indian Pines scene the corns and soybeans were very early in their growth cycle at the time of data collection, which resulted in low coverage of the soil ($\approx 5\%$) [2]. Given this low canopy ground cover, the variation in spectral response among different classes is very low and spatial information cannot significantly increase discrimination between different classes. In order to address this issue, a possible solution is to conduct the endmember extraction process in supervised fashion, taking advantage of the information contained in the available labeled samples in order to guarantee that a highly representative endmember is selected per each class.

Experiment 3. Impact of endmember purity on the final classification results

In a supervised endmember extraction framework, our first experiment is based on applying the unmixing chain #3 to select endmembers only from the available training samples. Apart from reducing computational complexity (which in this case involves a search for c endmembers in the pixels belonging to the training set), Table 4.1 reveals that this strategy improves the classification results reported for the chains #1 and #2. However, in order to make sure that only one endmember per labeled class is used for unmixing purposes, we also apply unmixing chain #4 in which spectral averaging of the available training pixels in each class is conducted in order to produce a final set of c spectral signatures. Despite averaging of endmembers can lead to degradation of spectral purity, it can also reduce the effects of noise and/or average out the subtle spectral variability of a given class, thus obtaining a more representative endmember for the class as a whole. This is illustrated by the classification results for unmixing chain #4 in Table 4.1, which outperform those reported for most other tested methods except the MNF. This indicates that, in a supervised unmixing scenario, the use of spectrally pure signatures is not as important as the choice of signatures which are representative of the available training samples. The improvement of the quality of information provided by chain #4 can be seen by looking at Fig. 4.6. All the single components provide an information about a spatially representative part of the image. In the case of chain #3, this is due to the fact that, considering only the training samples as possible endmembers for the OSP algorithm, most of the outliers are automatically discarded before the endmember selection process. Chain #4 further improves the results since every components gives specific information about the spatial coverage of each class at a sub-pixel level, allowing to better represent mixture of classes which may be caused by low spatial resolution.

Table II shows the statistical differences (average value of ten comparisons) between traditional dimensionality reduction methods and the unmixing chains #3 and #4, computed using McNemar's test [124] for the case of the polynomial kernel. The differences are statistically significant at a confidence level of 95% if $|Z| > 1.96$. For each couple of compared feature extraction chains, we report also how many times each chain wins/ties/loses after comparing the thematics maps obtained using the same training set. If the value of Z reported for each entry of Table II is positive and larger than 1.96, the first compared chain wins. By convention, the comparison is always performed with the first chain in a line of Table II and the second chain in a column of Table II. It can be noticed that unmixing chains #3 and #4 always perform significantly better than PCA and ICA. MNF performs better than chain #3, while the differences with chain #4 are in general not significant.

To conclude this section, Fig. 4.7 displays the best classification results (out of 10 runs) obtained after applying the SVM –trained with 10% of the available training samples– to each feature extraction strategy considered for the AVIRIS Indian Pines scene. As shown by Fig. 4.7, both the MNF in Fig. 4.7(d) and the chain #4 in Fig. 4.7(h) provide the best classification scores, with less confusion in heavily mixed classes.

(a) Ground-truth (b) Original image (84.27%)

(c) PCA (83.33%) (d) MNF (89.41%)

(e) Chain #1 (81.29%) (f) Chain #2 (79.64%)

(g) Chain #3 (87.99%) (h) Chain #4 (89.26%)

Figure 4.7: Best classification results for AVIRIS Indian Pines (using SVM classifier with Gaussian kernel, trained with 10% of the available samples per class).

4.5 Conclusions and Future Research Lines

In this chapter, we have investigated several strategies to extract relevant features from hyperspectral scenes prior to classification. For classification scenarios using SVMs trained with relatively small subsets of labeled samples, our experimental results reveal that the MNF greatly improves accuracies when compared to the more well known PCA and ICA transformations, used as an unsupervised feature reduction tool prior to classification. Due to the reduced dimensionality, classification using both the MNF and PCA subspaces generally improved the overall accuracy when compared to using all the original pixel's spectral curve. Results indicate that the proposed unmixing-based feature extraction chains can provide an alternative strategy to the PCA or MNF by incorporating information about the (possibly) mixed nature of the training samples during the learning stage, with the potential advantage of improved interpretability of features due the physical nature of the extracted abundance maps. Although final classification accuracies are likely to be dependent on the particular data set considered, the chains tested suggest higher accuracies with respect to traditional methods, such as PCA and ICA, and comparable accuracies related to MNF.

Further research is needed to define an optimality criterion to design unmixing chains as a feature reduction tool for classification purposes. A start point might be the chain #4 which indicates that, in the context of a supervised unmixing scenario, the use of spectrally pure signatures is not as important as the choice of signatures which are highly representative of the available training samples. Further lines of research should be the attempt to mitigate the sensitivity of spectral unmixing based chains to outliers and pixels with extreme value of reflectance, and a careful investigation of the influence of the number of features extracted. These issues will be the focus of Chapter 5.

Chapter 5

Investigation on the influence of the number of features

Abstract

Previous chapters have proven the importance of dimensionality reduction for hyperspectral images classification. A number of techniques were proposed, exploiting statistical properties and sub-pixel information of the image. Independently from the method selected to compute the dimensionality reduction, the number of components retained before the classification process can have a large influence on the final results. In this chapter, a thorough analysis of the issue is carried out. A number of feature reduction techniques, both unsupervised and supervised, are analyzed, and tests are performed on four hyperspectral data sets, in order to obtain results as general as possible. Moreover, a new technique based on unsupervised clustering and spectral unmixing was proposed, showing low sensitivity to pixels with extreme values of reflectance.

Contents

5.1 Introduction

In the context of supervised classification, the good generalization capability of machine learning techniques such as the support vector machine (SVM) [86] can still be enhanced by an adequate selection of the number of features to be used for classification purposes [125], especially if limited training sets are available *a priori*. The first four chapters of this thesis have discussed the use of dimensionality reduction techniques, which can help mitigating the problems connected to the unfavorable ratio between the (large) number of spectral bands and the (limited) number of training samples available *a priori*, which results in the Hughes phenomenon [126].

However, there are several issues still holding. Chapter 4 pointed out the importance of considering sub-pixel information in order to improve the average class accuracies of classes with very few training samples. This issue was investigated in the context of a set of newly developed feature extraction techniques based on spectral unmixing concepts [127]. These chains are intended to take advantage of the linear spectral unmixing model [94] in the characterization of training samples during the classification process. Endmember extraction algorithms have however demonstrated to be sensitive to the presence or outliers/pixels with extreme values of reflectance, which are considered as pure spectra and used to build components which are spatially representative in a small portion of the image. The consideration of this type of components for classification leads to a decrease of the algorithm performances, since they provide a small amount of information about the scene. Secondly, the choice of the most adequate number of features needed for classification purposes remains an elusive task. Such number depends mainly on the characteristics of the feature extraction algorithm and on the size of the available training set, which opens the way for the incorporation of supervised techniques for feature extraction in addition to more classic unsupervised ones. A challenging problem in this context is how to estimate the dimensionality of a scene, in order to retain the best possible number of features [128].
A traditional approach which is generally used for the identification of dimensionality is based on a certain percent of explained variability (e.g., 95% or 99%) obtained after applying Principal Component Analysis, as the threshold giving a stopping criterion [129]. However, these thresholds are often crossed by the first two or three PCs, which is a much too low number for a realistic dimensionality. A possibility is, of course, to set up higher thresholds, but a choice of a specific threshold seems to be somewhat arbitrary. Even more importantly, by increasing the percent of explained variability in a continuous fashion, one can end up with an arbitrary dimension. Therefore, in effect, the decision about the dimensionality is transferred to the decision about the percent of explained variability, without any clear guidelines.
The hyperspectral imaging community has tried to tackle the problem by subspace estimation methods such as the hyperspectral subspace identification by minimum error (Hysime) algorithm [116] or the virtual dimensionality (VD) concept [117]. More recently, a method based on second moments which can be considered as alternative to VD was proposed in [130]. However, an investigation of whether the dimensionality estimates provided by those methods can be related with the number of features needed to be retained prior to classification purposes has never been conducted in the past. All these methods estimate the number of pure materials signatures of the image. This number however could be different from the number of features allowing to achieve the best results in terms of classification accuracy, when dimensionality reduction has to be performed. Moreover, whether different chains have similar trends of results (for example, the maximum classification accuracy is obtained by keeping the same number of components) is another unexplored topic. A preliminary attempt of comparing different unsupervised dimensionality reduction techniques was presented in [131]. More

work is however requested in order to extend the evaluation to a larger number of techniques and to consider the possibility to use prior information which is always available in case of supervised classification purposes.

By keeping in mind the results obtained in the last chapter, here two main contributions are proposed. First, in order to mitigate the problem shown by spectral unmixing based algorithms, which may be affected by outliers and pixels with extreme values of reflectance, a new feature reduction chain is proposed. This chain is based on unsupervised clustering, in order to obtain components with a more significant spatial coverage than in the case of spectral unmixing chains. As a second contribution of this chapter, the issue of how many (and what type of) features can be used effectively for SVM-based classification is investigated. For this purpose, different types of feature extraction strategies (both unsupervised and supervised) prior to SVM classification are considered and the ability of automatic techniques for dimensionality estimation (i.e., Hysime and VD) to determine the optimal number of features that need to be retained for classification purposes is investigated. The study is conducted using two different hyperspectral scenes collected by the Airborne Visible Infra-Red Imaging spectrometer (AVIRIS) operated by NASA's Jet Propulsion Laboratory [132].
The remainder of the chapter is organized as follows. In Section 5.2 the different feature extraction techniques considered in experiments are described, along with the algorithms considered to determine the dimensionality of a hyperspectral image. Section 5.3 presents the experiments carried out. Section 5.4 draws some conclusions and future lines of research.

5.2 Processing Chains

This section provides an overview of the techniques used to extract features from the original hyperspectral data, and to estimate the number of features containing useful information for classification purposes. A detailed mathematical description of these techniques is out of the scope of this work, since most of them are algorithms well known in the remote sensing literature, so here only a short description of the conceptual basis is given. Feature extraction techniques are divided into unsupervised approaches, if the algorithm is applied on the whole data cube, and supervised techniques, if the information associated with the training set of the data is somehow exploited during the pre-processing step.

5.2.1 Unsupervised feature extraction techniques

We consider in this work five unsupervised feature extraction techniques. Three of them are classical algorithms proposed in the literature (PCA, MNF and ICA); the two remaining are based on the exploitation of sub-pixelic information, in the version proposed in Chapter 4 and in a variant based on the use of clustering techniques.

- *Principal component analysis* (PCA), an orthogonal linear transformation which projects the data into new coordinate system, such that the greatest amount of variance of the original data is contained in the first principal components [14]. The resulting components are uncorrelated.

- *Minimum noise fraction* (MNF), which differs from PCA in the fact that MNF ranks the obtained components according to their signal-to-noise ratio [133].

- *Independent component analysis* (ICA), which tries to find components as statistically independent as possible, minimizing all the dependencies in the order up to fourth [6]. There are several strategies that can be adopted to define independence (e.g., minimization of mutual information, maximization of non-Gaussianity, etc.). In this work,

among several possible implementations, we have chosen JADE [39] which provides a good tradeoff between performance and computational complexity when used for dimensionality reduction of hyperspectral images.

- *Unsupervised mixture-tuned matched filtering* (MTMF_u), which relies on the classic MTMF [134] method –also known as constrained energy minimization (CEM) [123]– in order to estimate fractional abundances of spectral endmembers extracted from the original hyperspectral image using the orthogonal subspace projection (OSP) algorithm [135]. Our main reason for using MTMF instead of other popular techniques for abundance estimation such as fully constrained linear spectral unmixing (FCLSU) [103] is the possibility to perform *partial unmixing* [134] in case not all endmembers are properly estimated. In a study like the one proposed in this chapter, where the feature extraction methods are evaluated for different numbers of components, partial unmixing may be more suitable than traditional techniques for abundance estimation which assume the presence of all endmembers in advance.

- *K-Means MTMF*(KM-MTMF_u), which is based on the use of the K-Means algorithms to perform an unsupervised clustering of the hyperspectral image. First, K-Means clustering is performed, by searching for as many classes as the components we want to retain after the pre-processing steps. The centroids of each cluster are considered as the endmembers of the image, and then the feature reduction components are obtained by applying the MTMF for the abundances estimation. The main objective of this chain is to solve problems highlighted in Chapter 4 by endmember extraction based algorithms, which are sensible to outliers and pixels with extreme values of reflectance. By using an unsupervised clustering method, we expect to avoid the problem showed by spectral unmixing methods, which sometimes extract endmembers with very high value of reflectivity covering only few pixels.

5.2.2 Supervised feature extraction techniques

We consider five supervised feature extraction techniques. Two of them are techniques previously presented in the literature, which exploit the spectral information of the pixel in order to project the data into a feature space suitable for class separation. We investigate also chains #3 and chain #4 presented in chapter 4 and a variation based on the use of clustering techniques.

- *Supervised MTMF* (MTMF_s), which is equivalent to MTMF_u but assuming that the pure spectral components are searched by OSP in the training set instead of in the entire original image. Our assumption is that training samples may better represent the available land cover classes in the subsequent classification process [127].

- *Supervised K-Means MTMF* (KM-MTMF_s), as in the previous case, this is the supervised counterpart of the chain (KM-MTMF_u). It differs only in the fact that the clustering is performed only in the training samples, and not in the whole image.

- *Average MTMF* (AV-MTMF), which corresponds to the chain # 4 presented in Chapter 4. One endmember per class is obtained, by averaging the value of all the training samples known in advance. When using this chain, the number of components can not be varied, being always equal to the number of classes considered for classification.

- *Discriminant analysis for feature extraction* (DAFE), which tries to maximize the ratio of the between-class covariance matrix and the within-class covariance matrix as a criterion function to perform feature extraction. The transformation matrix obtained

after the projection should therefore enhance the separability between different classes [126].

- *Non-parametric weighted feature extraction* (NWFE), which is based on DAFE but focusing on samples near the eventual decision boundaries that best separate the classes. The main ideas of the NWFE are: 1) assigning different weights to every training sample in order to compute local means, and 2) defining non-parametric between-class and within-class scatter matrices to perform feature extraction [126].

Another technique proposed in the literature to project the data into a new subspace taking into consideration the spectral characteristic of the training samples is the Decision Boundary Feature Extraction (DBFE). In the following experiments the DBFE was not tested since it could not be applied in case of limited/very limited training sets. As a matter of fact, DBFE requires, in order to estimate the statistics used to project the data, a number of samples (for each class) bigger than the number of dimensions of the original data sets. As it will be shown in the sections, these requirement was not satisfied for most of the experiments carried out in this work.

5.2.3 Dimensionality estimation techniques

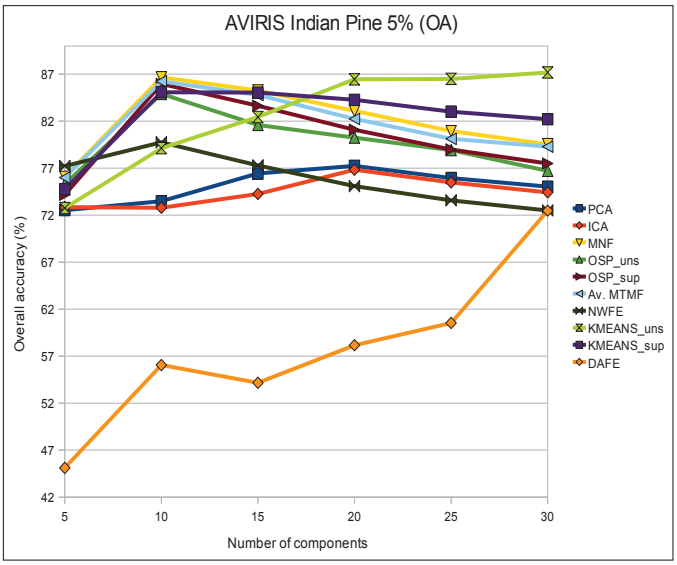
Recently, several techniques were proposed to estimate the dimensionality of a hyperspectral data set [116, 117, 130]. In order to have a comparison with the results obtained in our experiments, three dimensionality estimation techniques widely used in the literature were considered:

- *Hyperspectral signal subspace identification by minimum error* (HySime), which tries to estimate the number of pure materials by estimating the signal subspace in the hyperspectral image [116]. The method infers the data subspace by minimizing the sum of the projection error power with the noise power which are, respectively, decreasing and increasing functions of the subspace dimension.

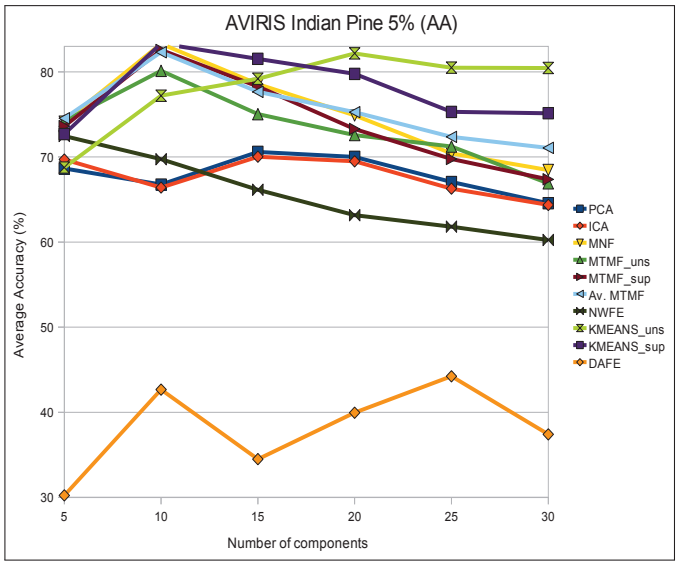
- *Virtual dimensionality* (VD), which assumes that if a signal source is present in the data, it should be possible to detect it in the relevant spectral band [117]. In essence, this implies that the eigenvalues of both the data-correlation and covariance matrices have to be computed. For a given false-alarm probability P_F the algorithm explores, based on the Neyman-Pearson detection theory, how many times the test fails for all spectral bands (that is, no signals are detected) and consequently determines the number of spectral signatures.

- *Principal component analysis-based dimensionality estimation* (PCADE), which simply retains a number of components which corresponds to the principal components spanning a percentage of the total variance of the data.

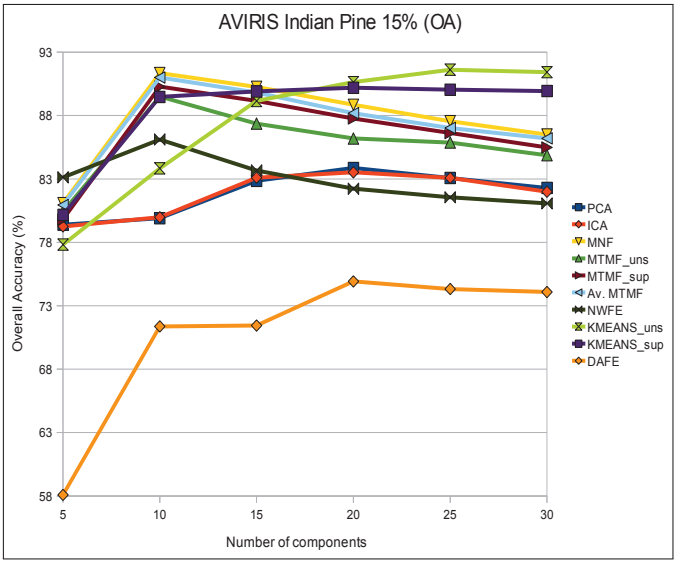
These techniques were not used to select a number of components in our tests, rather to have a term of comparison with the trend pointed out by analyzing a wide range of number of components retained. Another interesting issue investigated is if the dimensionality of an image, as estimated by these methods, could correspond to the most suitable number of features which should be computed in case of dimensionality reduction prior to supervised classification.



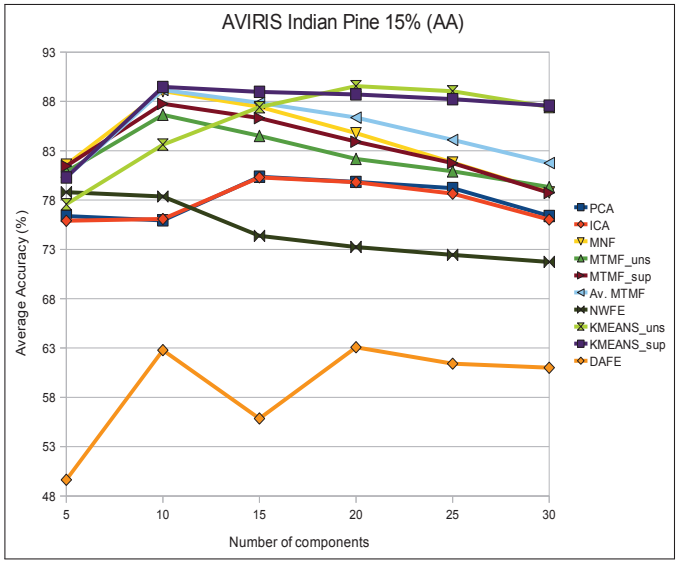
(a) Indian Pines 5%(overall accuracy plot)



(b) Indian Pines 5%(average accuracy plot)

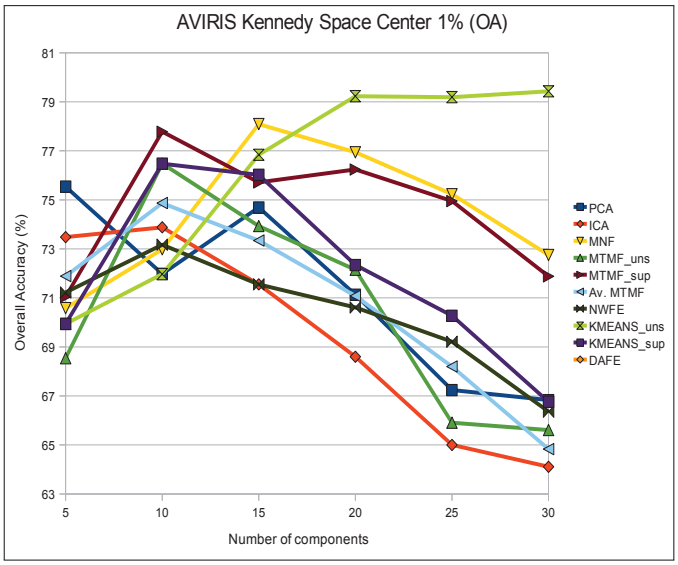


(c) Indian Pines 15%(overall accuracy plot)

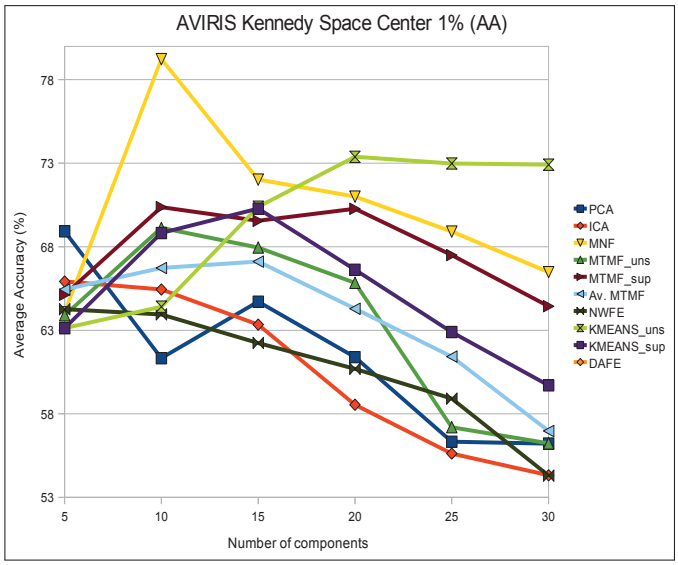


(d) Indian Pines 15%(average accuracy plot)

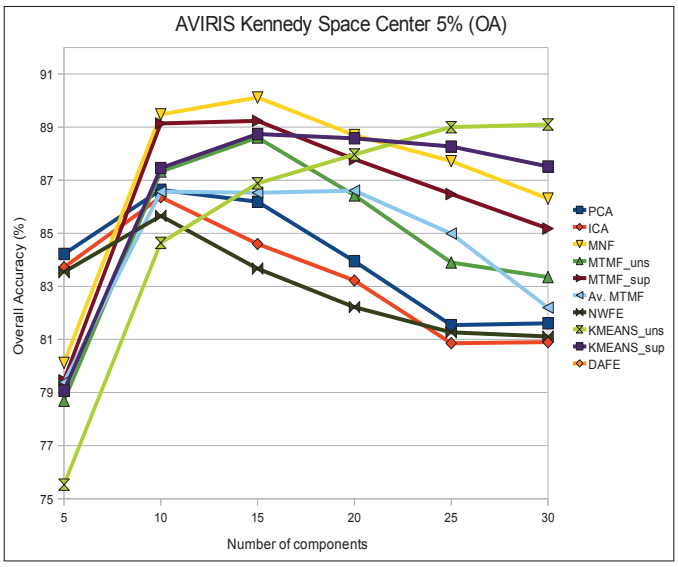
Figure 5.1: Overall and average accuracy plots obtained for the AVIRIS Indian Pines hyperspectral image with different types and number of features (using an SVM classifier with Gaussian kernel and trained respectively with 5% and 15% of the available samples per class).



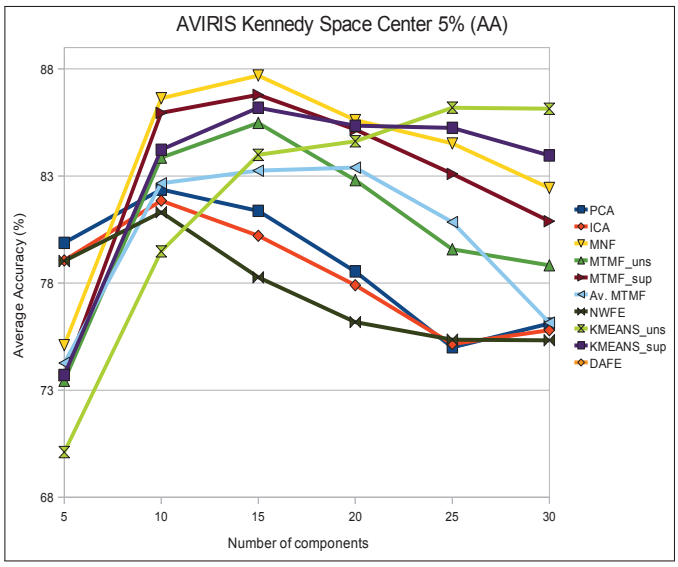
(a) Kennedy Space Center 1% (overall accuracy plot)



(b) Kennedy Space Center 1% (average accuracy plot)

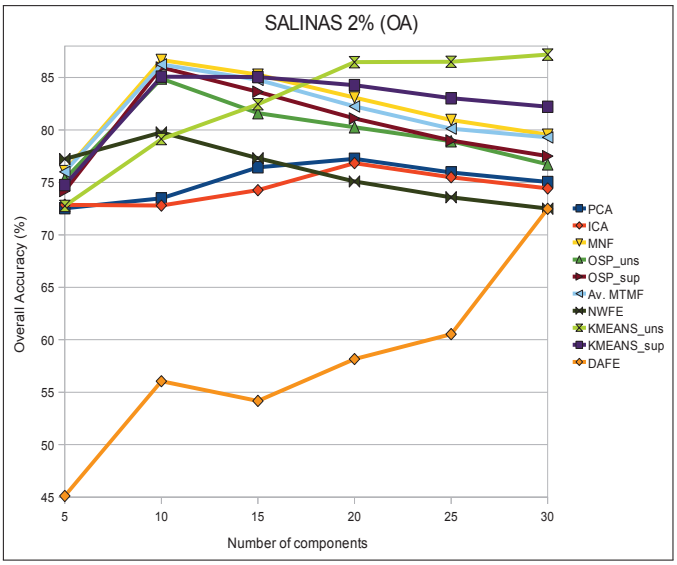


(c) Kennedy Space Center 5%(overall accuracy plot)

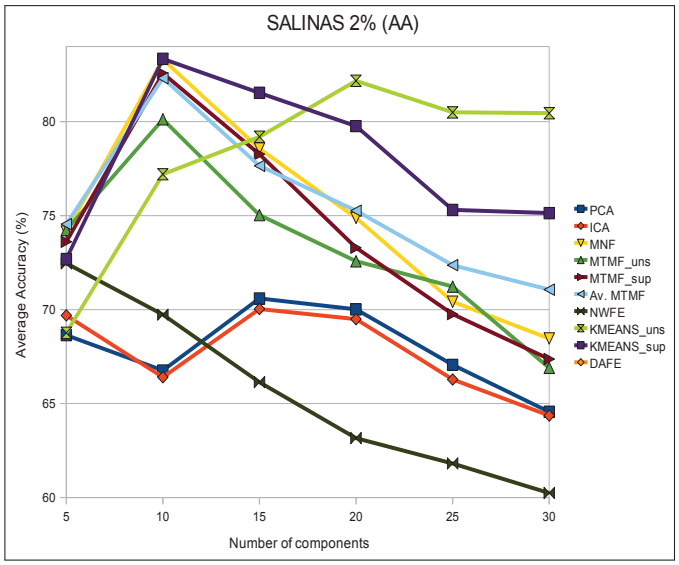


(d) Kennedy Space Center 5%(average accuracy plot)

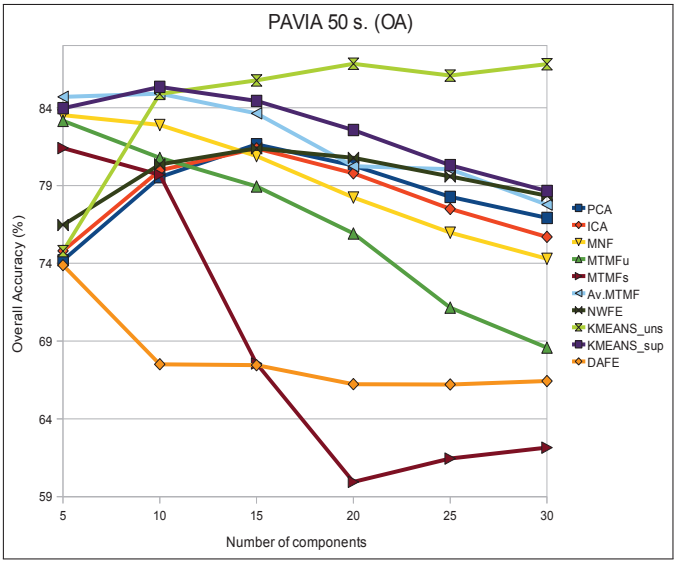
Figure 5.2: Overall and average accuracy plots obtained for the Kennedy Space Center hyperspectral image with different types and number of features (using an SVM classifier with Gaussian kernel and trained respectively with 1% and 5% of the available samples per class).



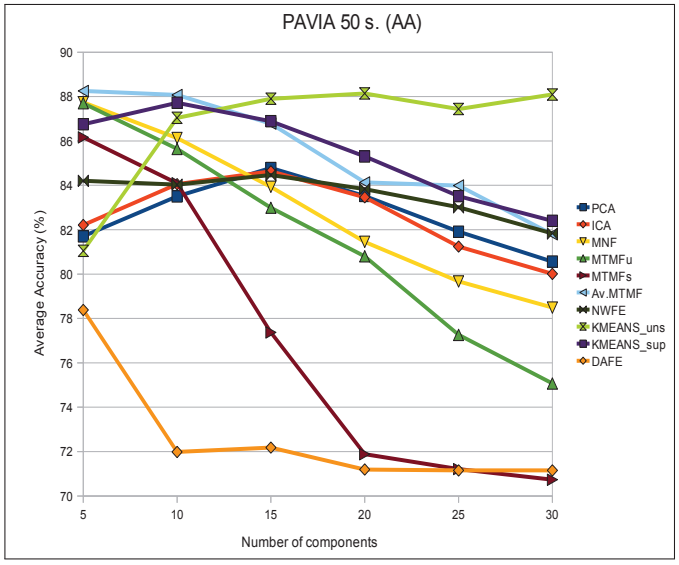
(a) Salinas 2% (overall accuracy plot)



(b) Salinas 2% (average accuracy plot)



(c) ROSIS University (overall accuracy plot)



(d) ROSIS University (average accuracy plot)

Figure 5.3: Overall and average accuracy plots obtained for the AVIRIS Salinas and ROSIS University hyperspectral images with different types and number of features (using an SVM classifier with Gaussian kernel and trained respectively with 5% of the available samples and 50 samples per class).

Table 5.1: Number of features estimated by different techniques from two different hyperspectral images collected by AVIRIS.

Dimensionality est.	Indian Pines	Kennedy	Salinas	ROSIS
HySime	12	21	20	60
VD ($P_F = 10^{-5}$)	29	26	19	14
VD ($P_F = 10^{-6}$)	28	24	19	13
VD ($P_F = 10^{-7}$)	27	22	18	12
PCADE (99% var.)	4	3	4	4
PCADE (99.9% var.)	30	10	21	28

5.3 Experimental Results

5.3.1 Experimental setup

In our experiments we have tried to compare a large number of data set, in order to obtain results as general as possible. Four different hyperspectral images were therefore analyzed in this work:

- *AVIRIS Indian Pines*, collected over the Indian Pines region in Northwestern Indiana on July 1992. The scene[1] comprises 145 lines by 145 samples and 220 spectral channels in the wavelength range from 0.4 to 2.5 μm, nominal spectral resolution of 10 nm, and spatial resolution of 20 meters by pixel. Several spectral bands were removed from the data set due to noise and water absorption, leaving a total of 202 radiance channels to be used in experiments. The ground-truth contains 16 land cover classes. The number of pixels in the smallest class is 20, while the number of pixels in the largest class is 2468.

- *AVIRIS Kennedy Space Center*, collected over the famous NASA Center in Florida[2] on March 1996. The portion of this scene used in our experiments has dimensions of 292×383 pixels. After removing water absorption and low SNR bands, 176 bands were used for the analysis. The spatial resolution is 20 meters by pixel. 12 ground-truth classes where available, where the number of pixels in the smallest class is 134 while the number of pixels in the largest class is 761.

- *AVIRIS Salinas*, collected over the Valley of Salinas, Southern California, in 1998. It contains 217×512 pixels and 224 spectral bands from 0.4 to 2.5 μm, with nominal spectral resolution of 10 nm. It was taken at low altitude with a pixel size of 3.7 m. The data include vegetables, bare soils, and vineyard fields.

- *ROSIS University*, collected over the Engineering School of the city of Pavia (Italy). The image is composed by 610×340 pixels and 103 spectral bands, in the wavelength range 0.4-0.86μm. The ROSIS is an airborne sensor, granting a higher spatial resolution with respect to satellite sensors like AVIRIS (the spatial resolution of this data set is 1.3 m/pixels).

For all the images, the feature extraction chains were applied prior to a supervised classification process performed by the SVM with Gaussian kernel. Kernel parameters were

[1] Available online: http://dynamo.ecn.purdue.edu/biehl/MultiSpec
[2] Available online: http://www.csr.utexas.edu/hyperspectral/data/KSC/

optimized by a grid search procedure, and the optimal parameters were selected using 10-fold cross-validation. The LIBSVM library[3] was used for experiments. For each feature extraction method, the number of selected components was varied from 5 to 30. The number of features to be retained was also computed by the HySime, VD and PCADE algorithms.
Since the main focus of the work was to evaluate the ability of the tested methods to perform with small training sets, we selected a very small number of samples to perform the classification. In the case of the Indian Pines data set we randomly selected 5% of each class and used them to build the training set. For the Kennedy Space Center data set –in which the size of the smaller classes is bigger– we decided to reduce the training sets even more and selected only 1% of the available ground-truth pixels for training purposes. For the other two data sets (AVIRIS Salinas and ROSIS University), we have selected respectively 2% of each class and 50 samples per class. However, in order to perform a preliminary investigation on the influence of the training set size on the accuracy of the algorithms, we have tested also larger training sets for the AVIRIS Indian Pine and the Kennedy Space Center. For these images, an additional training set containing respectively 15% and 5% of each class labelled samples was selected.
The overall and average classification accuracies were then computed over the remaining samples. Each experiment was repeated ten times in order to guarantee statistical consistence.

5.3.2 Analysis and discussion of results

The results shown in Table 5.2 and in Fig. 5.1, 5.2, suggest several considerations giving a better understanding of feature reduction techniques. First and probably most important, the use of supervised techniques for feature reduction is not always beneficial to improve the overall classification accuracy, especially in case of limited training sets and statistical based feature extraction approaches. For example, NWFE exhibits better results when compared to traditional unsupervised techniques such as PCA or ICA. However, DAFE exhibits poor results in the case of the Indian Pine data set (this method could not be computed for the Kennedy Space Center data set since the very small number of selected training samples did not allow us to estimate the statistics of the classes). The low performances obtained by DAFE should be therefore attributed to the very small size of the training set and, in the case of Indian Pine data set, to the fact that the land cover classes are spectrally very close thus making it very difficult to separate them by using spectral means and covariance matrices. Moreover, the importance of integrating the additional information provided by the training samples is strictly connected with the nature of the considered approach. This can be noticed when comparing the two proposed chains, MTMF and K-Means based, both unsupervised and supervised. In the former case (OSP+MTMF chain), the best results are provided by the supervised approach, since when considering only the training samples as possible endmembers which should be extracted with the OSP algorithm, the major drawback of the method (that is, the sensibility to outliers and pixels with extreme values of reflectance) is overcome. At the contrary, the K-Means based algorithm shows better results when used in unsupervised fashion. When trying to search a large number of cluster in a very small training set, several problems appear, such as the bad conditioning of matrices when computing the inverse (in the K-Means step), and the selection of very similar clusters, leading to redundant information which affects the classification performances.
Overall, the experiments carried out in this chapter show that the two techniques providing the best results are both unsupervised (KM_u and MNF). A closer look at the results reveals that K-Means based technique can be regarded as the most effective unsupervised feature extraction method, while $MTMF_s$ can be regarded as the best supervised feature extraction

[3] http://www.csie.ntu.edu.tw/cjlin/libsvm/

Table 5.2: Overall and average accuracies obtained in the best case observed in our experiments (with the number of features retained in such case indicated in the parentheses) for different feature extraction chains.

	Indian Pines 5%		**Indian Pines 15%**	
Feature extraction	OA	AA	OA	AA
PCA	77.25% (20)	70.59% (15)	83.86% (20)	80.37% (15)
ICA	76.84% (20)	70.03% (15)	83.52% (20)	80.30% (15)
MNF	86.67% (10)	83.31% (10)	91.35% (10)	89.04% (10)
MTMF_u	84.90% (10)	80.12% (10)	89.50% (10)	86.65% (10)
KM_u	87.18% (30)	82.17% (20)	91.61% (10)	89.55% (20)
MTMF_s	85.96% (10)	82.57% (10)	90.28% (10)	87.76% (10)
KM_s	85.08% (10)	83.34% (10)	90.19% (10)	89.47% (10)
DAFE	72.48% (30)	44.23% (25)	74.92% (20)	63.07% (20)
NWFE	79.76% (10)	72.46% (5)	86.11% (10)	78.80% (5)

	Kennedy SC 1%		**Kennedy SC 5%**	
Feature extraction	OA	AA	OA	AA
PCA	75.54% (5)	68.92% (5)	86.64% (10)	82.37% (10)
ICA	73.88% (10)	65.92% (5)	86.36% (10)	81.85% (10)
MNF	78.09% (15)	79.23% (10)	90.12% (15)	87.69% (15)
MTMF_u	76.48% (10)	69.13% (10)	88.61% (15)	85.48% (15)
KM_u	79.53% (30)	73.38% (10)	89.10% (30)	86.19% (25)
MTMF_s	77.78% (10)	70.80% (10)	89.24% (15)	86.79% (15)
KM_s	76.48% (10)	70.28% (15)	88.74% (15)	86.10% (15)
DAFE				
NWFE	73.17% (10)	64.26% (5)	85.66% (10)	81.31% (10)

	AVIRIS Salinas 2%		**ROSIS Univ. 50 samples**	
Feature extraction	OA	AA	OA	AA
PCA	91.93% (10)	95.48% (10)	81.65% (15)	84.78% (15)
ICA	91.72% (20)	95.33% (10)	81.39% (15)	84.63% (15)
MNF	93.71% (15)	96.60% (10)	83.52% (5)	87.74% (5)
MTMF_u	93.27% (15)	96.27% (10)	83.16% (5)	87.70% (5)
KM_u	92.83% (30)	96.18% (30)	86.83% (20)	88.15% (20)
MTMF_s	92.67% (15)	95.78% (10)	81.41% (5)	86.16% (10)
KM_s	92.58% (30)	95.56% (25)	85.34% (10)	87.72% (10)
DAFE	87.48% (10)	90.74% (10)	73.88% (5)	78.38% (5)
NWFE	92.28% (10)	96.09% (10)	81.40% (15)	84.46% (15)

method. Traditional supervised feature reduction methods exhibit lower accuracies. This can be seen in Table 5.1, which reports the best obtained results after testing different numbers of features (the optimal number of features is shown in the parentheses).
The second important issue addressed in this work is the possibility to estimate *a priori* the best possible number of features to retain when applying a dimensionality reduction algorithm. At a first sight, it can be easily understood the difficulty of the task. Different algorithms obtain better performances with different number of features. Moreover, in the case of the AVIRIS Kennedy Space Center data set, the same algorithm obtains the best classification accuracy with a number of features retained which varies according to the percentage of training samples considered. In general, an increase of the training set size allows to extract useful information from features which are not considered as the most informative, without incurring into the risk of caused by the curse of dimensionality, which could lead to the Hughes' phenomenon. Therefore, we can affirm that also if less informative features contain important information for the classification task, the limitation of samples available to train the classifier is a major limitation for exploiting such information.
The K-Means based unsupervised algorithm shows a more stable behavior in case of large number of components retained prior to classification, while the other unsupervised algorithms, once the best performance is reached, show a clear decrease in the results when increasing the number of features retained. In the case of supervised algorithms, the decrease of classification accuracy is faster, due to the appearance of problems connected to overfitting. In such a scenario, the possibility to automatically estimate the most suitable number of features to retain prior to classification seems to be an impossible task. Considering the three algorithms analyzed in this work, the approach which better estimates the dimensionality of the observed scene is the HySime, which in three of four cases gives a results close to the best classification accuracy obtained with the unsupervised methods, while it fails when applied to the ROSIS University data set. Quite opposite, the VD and PCADE methods (even after being tested with different parameter settings) could not provide dimensionality estimates able to maximize the overall classification accuracy for the different tested methods. Moreover, these methods suffer of the uncertainty due to the parameter which need to be set to estimate the components (the probability of false alarm for the VD and the percentage of variance spanned by the principal components).

5.4 Conclusions and future lines

In this chapter, we have investigated the issue of how many (and what type of) features can be used effectively for SVM-based classification of hyperspectral images. The study highlights several aspects, such as the difficulty in determining an adequate number of features to be retained, and the results provided by unsupervised and supervised feature extraction methods when very few training samples are available. Future developments of this work will include an investigation of additional techniques and hyperspectral image data sets.

Part III

Improving thematic map resolution

Chapter 6

Spectral Unmixing to improve the spatial resolution of thematic maps

Abstract

The second part of this thesis showed the importance of taking into account the sub-pixel information when classifying hyperspectral data. However, the information obtained when classifying an image containing mixtures of materials represents a thematic map which is intrinsically biased by the fact that regions of the image containing several land cover classes are represented with a single label. In this chapter, a new techniques taking into account the issue is proposed. By making a joint use of spectral unmixing and spatial regularization, classification maps with a finer spatial resolution are obtained.

Contents

6.1 Introduction

Target and structure detection and image classification are two important tasks of image processing, used in many application domains such as biomedical imaging and remote sensing [23,136,137]. In these fields, low image spatial resolution can highly affect the performance of the processing algorithms. Recent advances in sensor technology have made available images characterized by a very detailed spectral information on a wide spectral range, particularly suitable for these applications [138]. As shown in the previous chapters, hyperspectral images offers a huge quantity of spectral information from the visible to the infra-red region. For every recorded pixel, the rich spectral information provides a complete spectral description and a better characterization of the observed surface, thus resulting in a very powerful tool for materials discrimination. The advantageous characteristic of high spectral resolution data and the already planned civilian space missions which will make available in the next future a huge quantity of hyperspectral data (amongst the others, PRISMA, planned by the Italian Space Agency ASI in 2014, EnMap planned by German Space Agency DLR in 2014, Hyper-J and HyspIRI, planned by the Japan Space Agency and NASA, respectively, in the next future, besides the already on orbit sensors like the widely used Hyperion and AVIRIS [139], both of NASA) increase the attention given to this kind of data. However, a common drawback of hyperspectral sensors is the relatively low spatial resolution, which can vary from few to tens of meters, especially in the cases of high altitude sensors or instruments covering wide areas. There are many factors (such as imperfect imaging optics, atmospheric scattering, secondary illumination effects and sensor noise) that degrade the acquired image quality and make the development of new technology to improve the spatial resolution a very challenging task [5].

In the cases of structure detection and land cover classification, a low spatial resolution leads to the problem of mixed pixels, *e.g.*, pixels containing mixture of different materials [140]. The usual assumption that every pixel of the image can be associated with a unique class label is no longer verified, and mixed pixels cannot be correctly addressed by traditional classifiers [141]. Sometimes, spatial structures become hard to detect, with a substantive loss of information.
Because of this reason, the spectral unmixing problem has a high importance also in the analysis of hyperspectral images. The very high spectral resolution of this kind of data allows one a detailed characterization of the spectral signatures of the objects present in the investigated scene. This makes it possible to identify the abundances of constituents of a given material within the resolution cell. In these conditions, conventional crisp (hard) classification methods preclude a proper analysis of the image as it is not possible to model the sub-pixel abundances of each class in output from the classifier.

In this chapter, we address the problem of hard classification of pixels containing mixtures of pure materials. The question we try to answer is: is it possible to exploit the rich spectral information provided by hyperspectral data in order to obtain hard classification maps at a finer resolution with respect to the input data? A number of techniques were proposed in the literature to improve the spatial resolution of a hyperspectral image, but none of these approaches tried to exploit the big amount of spectral information in order to improve the resolution of a thematic map. The investigation of this issue is the objective of the chapter. The remainder of the chapter is as follows. Section 6.2 provides an overall presentation of different methods proposed in the literature to overcome problems related to low spatial resolution of hyperspectral images. In Section 6.3, the general framework of the proposed method is presented, and spectral unmixing is analyzed in detail. Section 6.4 is focused on the different techniques used to obtain a map resolution improvement. Section 6.5 is devoted to a toy experiment with synthetic data, where the effectiveness of the method is tested on a

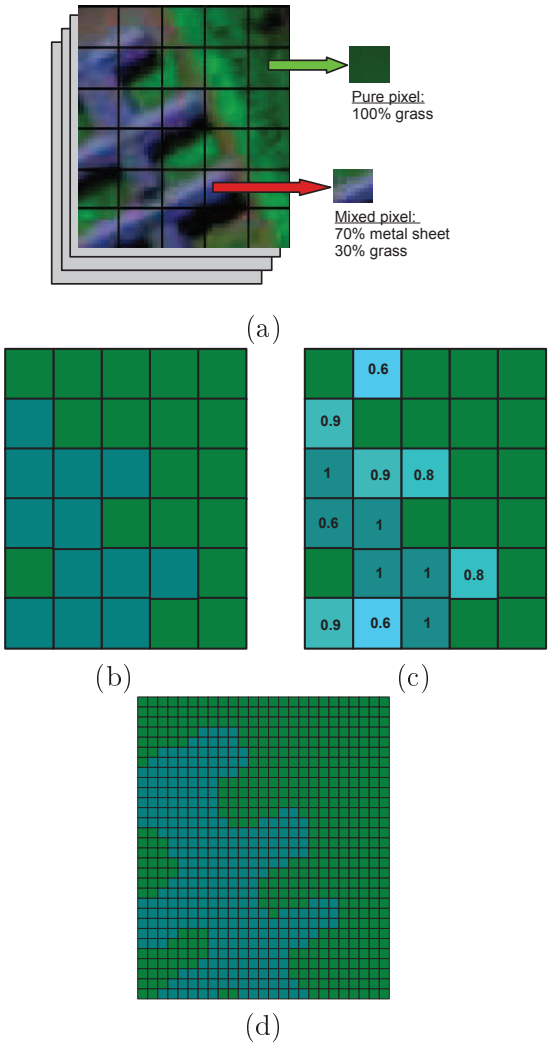


Figure 6.1: Illustration of the problem of mixed pixels in remote sensing images: (a) When the spatial resolution is not fine enough, several land cover classes can be found in the same pixel. (b) In this case, a hard classification process cannot give an accurate information about the pixel coverage, leading inevitably to a loss of information (in the example figure, the classes of grass and metal sheet are considered for classification). (c) Spectral unmixing can provide information at a sub-pixel level, computing the fractional abundances of a class within each pixel. However, the supplementary information obtained does not result in a resolution enhancement of the classification map. (d) The proposed method tries to handle the problem by jointly using the ideas of spectral unmixing and unsupervised classification, in order to obtain thematic maps at a finer scale, without needing any additional source of information.

scene where all the ground truth is known in detail. Experiments on two real hyperspectral data sets are discussed in Section 6.6, while the conclusions are finally drawn in Section 6.7.

6.2 Related works

Several techniques have been proposed during the last years to deal with the problem of mixed pixels and low spatial resolution of remote sensing images [94, 142–148]. These techniques can be divided in three main groups. The first group includes techniques which use high spatial resolution images jointly with the resolution images, in order to obtain a fused image with high spectral and spatial resolutions [142–144]. The resulting image can afterwards be used as input for classification. The main drawback of this approach, besides the need for an accurate coregistragion of the two images, is the need for ancillary data which is not easy to obtain.
The second group of techniques comprises super-resolution approaches independent from any high spatial resolution data. Tatem et al. propose an algorithm based on the Hopfield Neural Network [149] which does not need any secondary source of data to realize the super resolution mapping. This method has a good efficiency but suffers from high computational cost. Gu et al. tackle the problem of high computational cost through a fast learning-based algorithm to integrate the spatial and spectral information of hyperspectral images, back propagation neural network and some ground truth information which is unassociated to the considered test data [100]. The output of such techniques is a series of abundance maps at a higher resolution. For each class considered, an abundance map is created.
Finally, several methods which classify images assuming the possibility of mixed pixels were proposed in the last years. Examples of such techniques are soft classification algorithms [147], which provide a set of images (one per class) expressing the degree of membership to the class, and linear spectral mixture analysis (SMA) [94, 150, 151], which assumes every pixel to be the weighted sum of some constituent spectra, also called *endmembers*. Similar methods have been proposed for subpixel image labelling, based on the latest development of machine learning [152, 153]. These techniques can partially overcome the weakness of full pixel classification methods when analyzing mixed pixels, and they are suitable to be used for the analysis of these scenarios. However, the final output is a classification map (or a series of maps) representing the membership degree (or the abundance) of each pixel with respect to a class. When trying to obtain a crisp output, the additional information provided by fuzzyness is lost.
The work presented in this chapter tries to tackle the problem of mixed pixels from a different viewpoint, in order to consider as input a hyperspectral image (HSI) with a given spatial resolution and obtain a thematic map where the distribution of the classes is depicted in a classification map with finer spatial resolution. The main principle is to sufficiently mine the data advantages of HSI by spectral unmixing and superresolution mapping and to integrate the spectral and spatial information for resolution enhancement. One advantage of the proposed method is that no supplementary source associated with HSI is needed. The idea is to provide a unifying framework, able to address the problems highlighted here above, considering the rich information provided by hyperspectral data and exploiting the characteristics of both unsupervised classifiers and spectral unmixing to address the quantification of pure classes within mixed pixels, coupled with a spatial regularization which aims at correctly locate sub-pixels from a spatial viewpoint. According to the authors best knowledge, this is the first time that a similar unsupervised technique is proposed. Besides the novelty of the method, a careful investigation of the variables having an influence on the final classification accuracy is conducted, with particular attention to the well known problem of spectral variability. Preliminary results of this work were presented in [154], and encouraged to further

develop the method.

6.3 Spectral Unmixing

The general scheme of the proposed methodology is shown in Fig. 6.2 and Fig. 6.4. The first step is the determination of the classes within the image. Two approaches are proposed in this work:

6.3.1 Source Separation based technique

The first approach retrieves the constituent endmembers of the image by mean of a source separation technique. Let us denote a remotely sensed hyperspectral scene with n bands by $\mathbf{I}$, in which each pixel is represented by a vector $\mathbf{X} = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^n$, where $\mathbb{R}$ denotes the set of real numbers in which the pixel's spectral response x_k at sensor channels $k = 1, \ldots, n$ is included. Under the linear mixture model assumption, each pixel vector can be modeled using:

$$\mathbf{X} = \sum_{z=1}^{p} \Phi_z \cdot \mathbf{E}_z + \mathbf{n}, \tag{6.1}$$

where $\mathbf{E}_z$ denotes the spectral response of endmember z, Φ_z is a scalar value designating the fractional abundance of the endmember z at the pixel $\mathbf{X}$, p is the total number of endmembers, and $\mathbf{n}$ is a noise vector.
A number of techniques has been recently proposed to retrieve endmembers by mean of source separation [108].

In [155], the Vertex Component Analysis (VCA) is proposed as an effective method for extracting the endmembers which are linearly mixed. VCA makes use of the concept of orthogonal projection. Algorithms based on this concept, start by selecting the pixel vector with maximum length in the scene as the first endmember. Then, they look for the pixel vector with the maximum absolute projection in the space orthogonal to the space linearly spanned by the initial pixel and labels that pixel as the second endmember. A third endmember is found by applying an orthogonal subspace projector to the original image, where the signature that has the maximum orthogonal projection in the space orthogonal to the space linearly spanned by the first two endmembers. This procedure is repeated until the desired number of endmembers p is found.
VCA, as opposed to the methodology previously described, exploits the fact that the endmembers are the vertices of a simplex and that the affine transformation of a simplex is also a simplex. As a result, VCA models the data using a positive cone, whose projection onto a properly chosen hyperplane is another simplex whose vertices are the final endmembers. After projecting the data onto the selected hyperplane, the VCA projects all image pixels to a random direction and uses the pixel with the largest projection as the first endmember. The other endmembers are identified in sequence by iteratively projecting the data onto a direction orthogonal to the subspace spanned by the endmembers already determined. The new endmember is then selected as the pixel corresponding to the extreme projection, and the procedure is repeated until a set of p endmembers is found [155].
In our work, we have selected VCA due to its good performances and the very low computational burden.

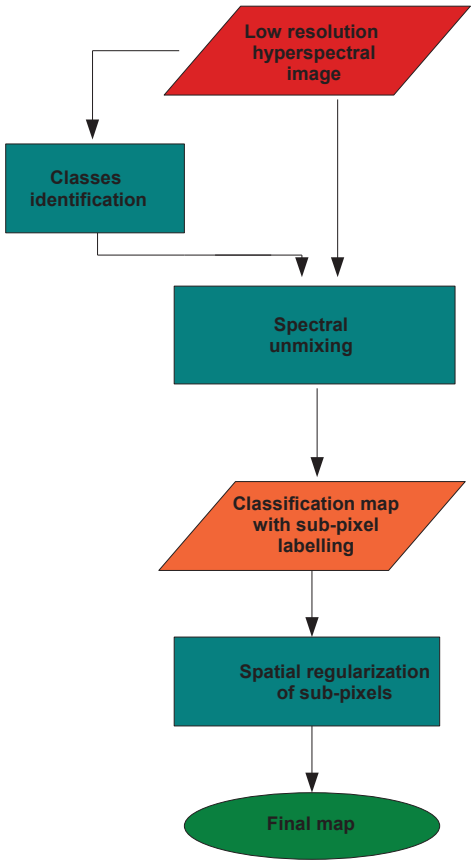


Figure 6.2: Block diagram of the proposed approach. In a first step, thematic classes are identified trough endmember extraction or unsupervised classification. Spectral unmixing is used to compute abundances of classes within each pixel. After splitting pixels in a number of sub-pixels and assigning them to a class according to the results of unmixing, a spatial regularization is performed to obtain the final map.

6.3.2 Clustering based technique

Endmember extraction techniques are an easy way to retrieve endmembers, especially in the case of images comprising mixed pixels. However, these techniques are in general sensible to 'outliers', *e.g.*, isolated pixels with anomalous values of reflectance, which are detected as extreme pixels and therefore endmembers. For such a reason, the second proposed technique is an extension of a traditional unsupervised classifier (the K-means classifier) as a method to detect classes composing the image, which is expected to be less sensitive to the issues related to the presence of outliers. Given a set of observations $x_1, x_2, \ldots, x_n$, where each observation is a d-dimensional real vector, K-means clustering aims to partition the n observations into p sets so as to minimize the within-cluster sum of squares:

$$min_S \sum_{i=1}^{p} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \mu_i\|^2 \tag{6.2}$$

where μ_i is the mean of points in the cluster S_i.

The K-means classifier is first applied to the image. At the end of the classification process, the centroids of the classes found by the algorithm are retained as constituent spectra of the image.

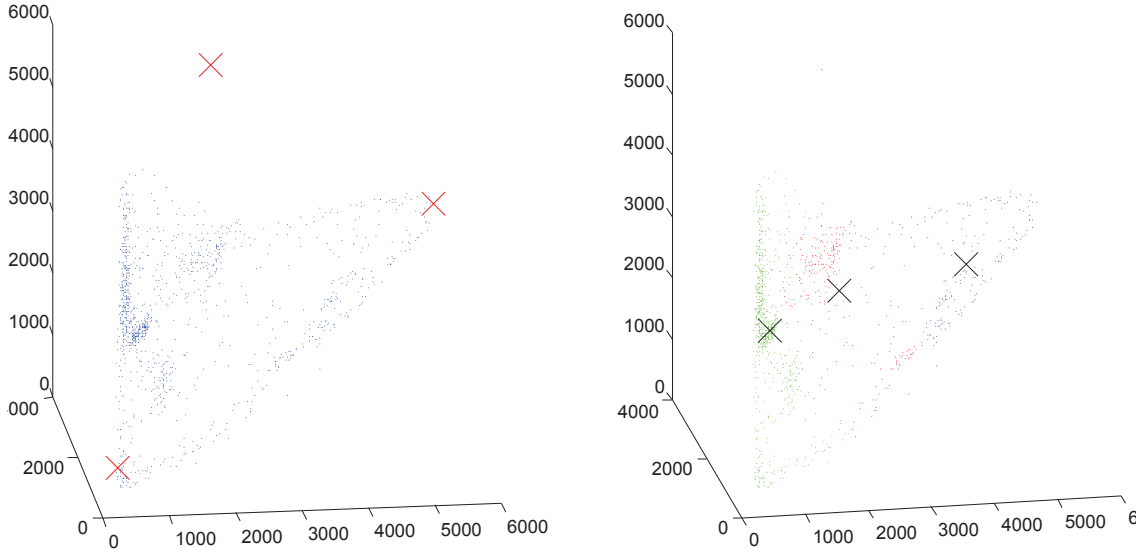


Figure 6.3: Examples of differences between source separation and unsupervised classification based approach for classes identification in a 3-dimensional data space with 3 classes (data set considered: ROSIS University, three principal components). Geometrical based source separation algorithms, such as VCA, generally identify as endmembers the pixels which are vertex of the simplex containing the data points (depicted as red 'x' in the left hand side figure). In opposite way, unsupervised classifiers separate the data into a specified number of clusters, where the centroid represents the central value of the class-specific data cloud (black 'x' in the right hand side figure). As it can be seen from the illustration, source separation algorithms are generally sensitive to outliers (*e.g.*, pixels with extreme value of reflectance), which are often identified as endmembers. On the other hand, in case of highly mixed scenes, the centroids depicted by unsupervised classifiers could correspond to mixed pixels, this representing a problem for the proposed algorithm.

Once the endmembers are extracted from the image, the abundance fractions of the elements within each pixel should be determined. Several algorithms have been developed to handle the linear mixing model according with the required physical constraint of abundance fractions, which are non negativity (all the abundances must be greater than or equal to zero) and full additivity (the sum of the endmember abundances within a pixel should be equal to one). Due to the efficiency from a computational point of view, a common choice is to use a fully constrained least squares (FCLS) algorithm, which satisfies both abundance constraints and is optimal in terms of least squares error [109]. In concrete processing, the main problem is that the FCLS does not have a closed-form mathematical solution due to the nonnegativity constraints; thus, a numerical solution is always required.

6.4 Improving Spatial Resolution

Spectral unmixing is useful to describe the scene at a sub-pixel level, but can only provide information about proportion of the endmembers within each pixel. Since the spatial location remains unknown, spectral unmixing does not perform any resolution enhancement. Here, we investigate two super-resolution mapping techniques, which take advantage of the information given spectral mixing analysis and use it to enhance the spatial resolution of thematic maps.
First, we set an abundance threshold to determine if a pixel can be considered as 'pure'. Since a single spectral signature can not represent extensively a class within the whole im-

age, the abundance determination is negatively affected by spectral variability. Therefore, all the pixels with *maximum* abundance greater than this threshold are considered as entirely belonging to the corresponding class. The other pixels are considered as mixed. Then each pixel is divided in a fixed number of sub-pixels, according to the desired resolution enhancement. Every sub-pixel is assigned to a class, in conformity with the fractional abundance computed in the first step. For each pixel, the number of sub-pixels n_i to assign to the class i is computed according to the equation:

$$n_i = \text{round}\Big(\frac{\text{abd}_i}{1/N^2}\Big), \tag{6.3}$$

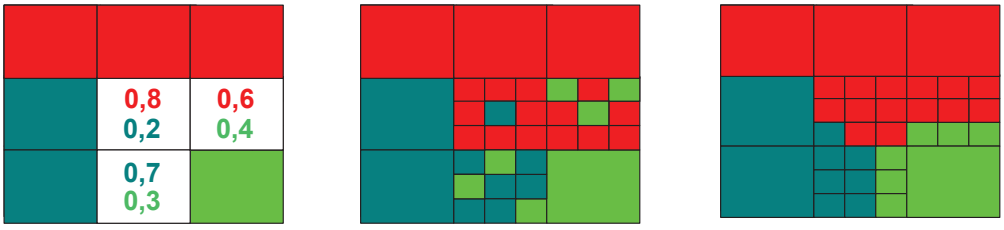
where abd_i is the fractional abundance of the class i within the considered pixel estimated with the FCLS, $N \times N$ is the total number of sub-pixels within each pixel, and round(x) returns the value of the closest integer to x.
As mentioned before, a well known problem of hyperspectral images source separation and spectral unmixing covering wide areas is that abundances determination is negatively related to the intra-class spectral-variability [156], and the assumption that a single endmember could extensively represent a class is generally far from reality. In other words, if we extract two spectra of the same material in two different corners of the image, they will probably present several differences, this having a negative effect if a single endmember is used for abundance estimation. In order to investigate the influence of spectral variability on the final results, we have tested two possible approaches. When using an endmember extraction algorithm to retrieve the endmembers, the spectral signatures retrieved are used as 'endmembers' in the whole image, since such techniques retrieve the purest pixel of the images, which could be used in the whole data set to determine abundances within mixtures. In the case of endmembers that are extracted using clustering techniques, it is computed a preliminary 'classification map', where only the purest pixels are labeled (a pixel is considered as pure if its maximum abundance is higher than a chosen threshold). Then, the abundances of unlabeled pixels are re-computed by considering as 'endmember candidates' only a number of pixels in the spatial neighborhood of the considered mixed pixel. The endmember candidates are therefore chosen among the pixels labeled as 'pure' in the first step, in order to use local endmembers and to handle the problem of spectral variability. This is done because endmember retrieved by clustering techniques are more likely to be mixed or slightly mixed pixels, which can not represent the correspondent class all in the whole image.

6.4.1 Simulated Annealing

The first approach to locate the sub-pixels is based on a simulated annealing mapping function. The algorithm is used to create random permutation of these sub-pixels, in order to minimize a chosen cost function. Relying on the spatial correlation tendency of landcovers, we assume that each endmember within a pixel should be spatially close to the same endmembers in the surrounding pixels. We assume that, if the same material can be encountered in close pixels, it is more probable that form a single object rather than two isolated ones. Therefore, the cost function C to be minimized is chosen as the perimeter of the areas belonging to the same class:

$$C = \sum_{i=1}^{I} \sum_{j=1}^{N_i} P_j, \tag{6.4}$$

where I is the number of the classes, N_i is the number of connected components of the class i, and P_j is the perimeter of the connected component j, computed according to the



(a) (b) (c)

Figure 6.4: Basic steps of the proposed resolution improvement method: (a) Spectral unmixing provides information about abundance fraction within each pixel (b) Pixels are split into N sub-pixels, according to the desired zoom factor, assigned to an endmember and randomly positioned within the pixel (c) Sub-pixels are re-located by Simulated Annealing or Pixel Swapping. Simulated annealing performs random permutations of the sub-pixels position until *minimum* cost is reached (in the proposed example, the perimeter cost function of figure (c) is the best possible solution, and clearly lower than the case shown in figure (b)). Pixel Swapping changes sub-pixel positions as shown in Fig. 6.5.

8-connected border pixels model [157].
Simulated Annealing (SA) is a well established stochastic technique originally developed to model the natural process of crystallization [158]. This process is based on an analogy from thermodynamics where a system is slowly cooled in order to achieve its lowest energy state. More recently, SA has been proposed to solve global optimization problems [159], and it has been used in various fields.

The basic idea of the method is that, in order to avoid to be trapped in local minima, uphill movements, i.e. the points corresponding to worse objective function values could, sometimes, be accepted by the following iterative procedure. As with a greedy search, it accepts all changes that lead to improvements in the fitness of a solution. However, also changes which lead to worse solutions can be accepted. The probability of accepting a reversal is inversely proportional to the size of the reversal with the acceptance of smaller reversals, being more probable. This probability also decreases as the search continues, or as the system cools, allowing eventual convergence on a solution.

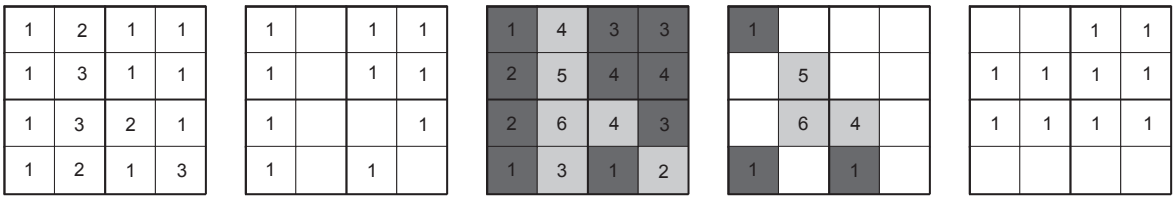
6.4.2 Pixel swapping

The second algorithm investigated is the so-called pixel-swapping algorithm, based on the concept of sub-pixels attractiveness introduced by Atkinson in [160, 161]. The original algorithm takes as input the sub-pixels hard land cover maps obtained after converting the fractional abundances of the classes and try to maximize the spatial correlation of same class sub-pixels while preserving the proportional composition within a "low-resolution" pixel. With respect to the original algorithm, three main differences are presented here:

- In [161], the fractional abundances are considered as a given information, while here we compute them in the first step of the proposed algorithm.

- The original algorithm was proposed for two classes problems. Here, we apply it to a multi-class classification problem.

The pixels swapping algorithm takes in consideration a weighted function representing the 'attractiveness' of a sub-pixel location. For each sub-pixel position i within the pixel, the attractiveness O_i is represented by:



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

Figure 6.5: Basic steps of the pixel swapping algorithm: (a) The hard sub-pixel values are random located within each pixel. (b) Example of sub-pixel distribution for class 1 (c) For each sub-pixel location, the attractiveness O_i is computed. In the example figure, sub-pixels allocated to class 1 are represented in dark gray, sub-pixels allocated to other classes are represented in light gray. (d) Within each pixel, the lowest attractiveness value of a sub-pixel allocated to the considered class and the highest attractiveness value of a sub-pixel allocated to another class are considered. If the first value is smaller than the second, the sub-pixels are swapped. (e) Example of sub-pixels relocation after the first iteration.

$$O_i = \sum_{j=1}^{n} \lambda_{ij} Z(X_j) \tag{6.5}$$

where n is the number of neighboring pixels considered, $Z(X_j)$ is the binary value of the class z in the j^{th} sub-pixel location X_j (1 if the sub-pixel belongs to the class z, 0 otherwise), and λ_{ij} is a weight computed as

$$\lambda_{ij} = \exp\left(\frac{-h_{ij}}{\alpha}\right) \tag{6.6}$$

where h_{ij} is the distance between the sub-pixel locations i and j, and α is a range parameter of the exponential model. Several weighting functions were explored in [162] as possible alternatives to the exponential function. The authors suggested that a simple Nearest Neighbor model could provide comparable accuracy of more complex spatial models, such as the exponential weighting function, with a much simpler model. Equation (6.5) becomes therefore the simple sum of the values in the nearest sub-pixels positions:

$$O_i = \sum_{j=1}^{n} Z(Xj) \tag{6.7}$$

The method can be described as follows: within a pixel, for each class, the attractiveness value O is computed for each sub-pixel position i and class z. If the least attractive value of a sub-pixel actually belonging to a class z is smaller than the most attractive value of a sub-pixel belonging to another class, the two sub-pixels are swapped. This procedure is repeated either for a previously fixed number of times or until a stopping criterion is reached.

6.5 Experiments on synthetic data

This section aims at giving an overall idea of the proposed method, when used for classifying synthetic data. Because of this reason, only the endmember extraction based method is considered, with Simulated Annealing sub-pixel mapping.

One of the main problem when dealing with spectral unmixing of real data is the difficulty to assess the results obtained, especially in the case of abundance fractions estimation. If

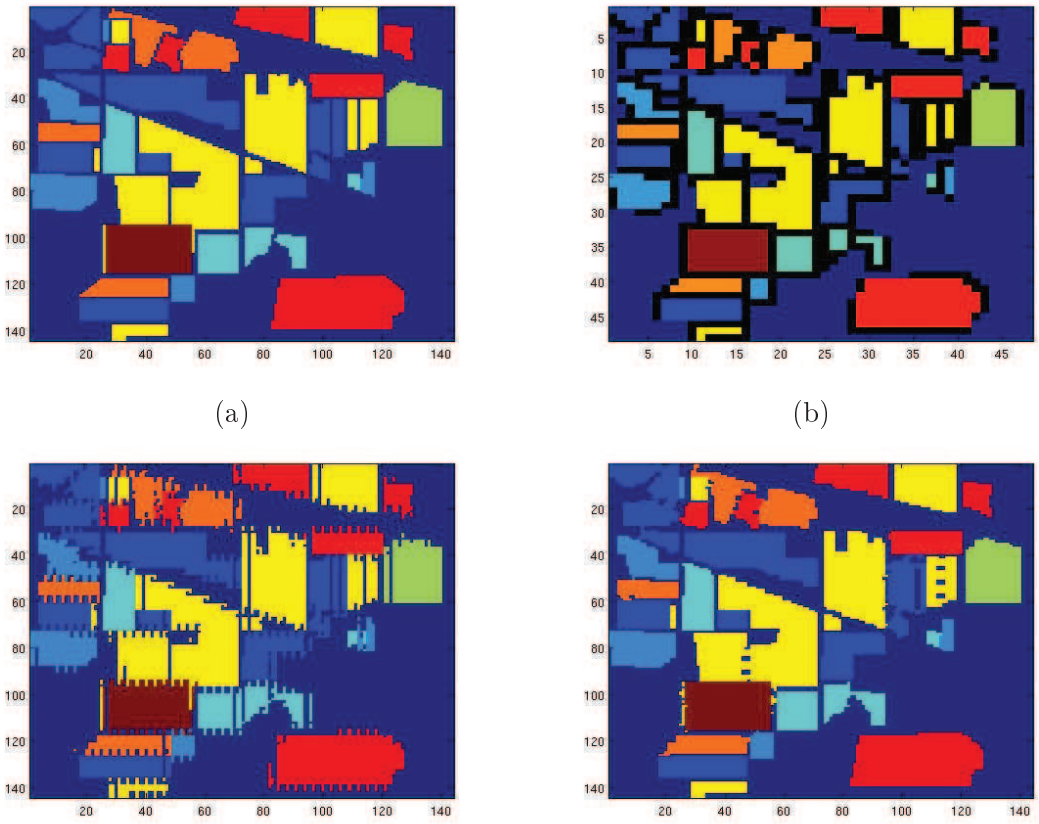


Figure 6.6: (a) Ground truth of the syntethic image created with reference spectra. (b) Ground truth after down-scaling of a factor 3. Mixed pixels are shown in black (c) Image obtained after source separation and random positioning of the elements (d) Final results after simulated annealing optimization

reference spectra can be easily obtained from a laboratory simulation or from *in situ* observations, to have a complete map of sub-pixels coverage and their fractional abundances is much more difficult. This is the main reason that lead us to first consider synthetic data, where all the spectra and the abundance fractions are known in a detailed way. In the following example, we have created an image using spectra obtained from a library, in order to better illustrate the working of the proposed method (for simplicity, we only show here the results obtained with the spectral unmixing based endmember extraction with Simulated Annealing sub-pixel location).

We have chosen 9 spectra from the USGS spectral library[1] [163], namely asphalt, brick, metal sheet, roofing felt, water, dry grass, oak leaves, lichen and green slime. This library provides a very detailed spectral description of the elements, with more than 2100 considered wavelengths. In order to consider as realistic as possible data, which could be comparable with the data provided by the last generation hyperspectral sensors, we have downsampled the original spectra of the library, so that each spectrum used in the experiment was composed by 216 values in the 0.3-2.5 μm range, which could roughly corresponds to the spectral range and the amount of information of a real data. Despite most of these spectra are not closely related, some of them have a very high spectral correlation making more difficult the source separation process.

In this simulation, we have tried to perform an accurate simulation of data containing spatial characteristic close to reality. In order to create a realistic fractional abundance map, we have considered the ground truth of a widely used image in remote sensing application, the AVIRIS Indian Pine image[2]. By considering the reference map of the AVIRIS data as ground truth of our simulated scene, we have substituted the original spectral values with those obtained after the convolution of the USGS library spectra. Therefore, every spectrum will have a spatial distribution of a land cover class of the AVIRIS image (or, in some cases, more than one, since the AVIRIS image contain 16 different classes, while in our experiment we have only nine). No spectral variability is considered in this first experiment.

After creating the synthetic image of 144×144 pixels which will be used later to assess the results of our method, we perform a down-scaling by substituting each 3×3 window of pixels with its average value, obtaining a new image composed by 48×48 pixels, with the same number of bands but a lower spatial resolution (of course, being a simulated imaged, the spatial resolution is relative; however, we obtain an image covering the same area where mixture of pixels are incorporated). This image will be the input data of the proposed approach, to improve the spatial resolution of its classification map. The ground truths of the original and filtered images are shown in figure 6.6, (a) and (b). In the image a majority of pure pixels is present, but it also contains a number of areas with mixtures of materials (totally the low resolution image contains 1722 pure and 582 mixed pixels). Due to the lack of spectra variability this is not a particularly challenging scenario, but still it can show the possibility offered by the proposed method.

Results are shown in Figures 6.6 (c)-(d), and in Table I. To have a quantitative assessment of the proposed method, we performed a comparison between the original high resolution image (figure 6.6a) and the image with enhanced resolution. The indicator that we have evaluated are the following: The correspondence of the retrieved spectra with the spectra used to build the image, the percentage of pixels which are not correctly retrieved after the spectral unmixing step, the percentage of pixels which are not correctly located after the Simulated Annealing step, and the percentage of mixed pixels (that are, pixels considered as pure in the high resoltion image and mixed in the low resolution one) which are not correctly located after applying SA.. More than 98% of the pixels are correctly located,

[1] http://speclab.cr.usgs.gov/spectral.lib06/ds231/datatable.html

[2] Available on-line: http://dynamo.ecn.purdue.edu/∼lbiehl/

Table 6.1: Spectral unmixing for unsupervised classification: Performance obtained in the different scenarios

Filtering window size	3x3	4x4		
SNR	Inf	50 dB	30 dB	25dB
Mean spectra correlation	1	1	1	0.98
Pixel unc. labeled after SS	0.15%	0.83%	1.80%	2.51%
Pixel unc. postioned after SA	1.89%	6.40%	7.31%	8.82%
Mixed pixel unc. positioned	7.50%	17.86%	19.47%	24.02%

which correspond to the 92.8% of the mixed pixels. All the pure pixels are correctly labeled. From this simple experiment, the effectiveness of the proposed method can be evaluated. We can preliminary conclude that when the classes are correctly retrieved with spectral unmixing, the proposed spatial regularization technique provides very good results.

6.5.1 Spectral variability

In order to search for more challenging scenarios, the image was down-scaled of a factor 4, and the influence of noise has been considered. The filtered image is therefore composed by 1296 pixels, 818 pure and 478 mixed. Zero mean Gaussian noise is added to the original signal, in order to obtain the desired SNR. Three different values of SNR have been tested: 50, 30 and 25 dB. When considering real hyperspectral images, the amount of noise is in general much lower than the one tested here; however, the Gaussian noise is useful to try to represent the spectral variability of real data, which is not considered in synthetic images. The presence of noise could affect the source separation step. If the endmembers are not correctly retrieved, the optimization step will start with wrong assumptions and inevitably lead to a bad result. As it can be seen from Table 7.4, the overall error increases with the noise, but still very good results are obtained. In the worst case, which is synthetic data with zoom factor of 4 and SNR of 25 dB, more than 91% of the image pixels are correctly labeled and positioned. Thus, the results obtained in the previous experiments are confirmed, stating the validity of the proposed spatial regularization and of the whole method proposed.

6.6 Experiments on real data

The experiments on real images were conducted by considering two different hyperspectral data. The first considered data set ROSIS data acquired over the University of Pavia, Italy, with 103 bands, ranging from 0.43 to 0.86 μm, with a 1.3 m spatial resolution. The very high value of the spatial resolution, which is not common in traditional hyperspectral satellite sensors, is due to fact that ROSIS is an airborne sensor. Here, we consider a small segment (120×90 pixels) of the image, which contains several land cover classes. Figure 6.7 (a) shows a gray scale image of the 30th band of the scene. The main element of interest is the metal sheet structure in the center of the image.
The second image analyzed in our experiments is an AISA Eagle dataset. It contains 252 bands ranging from 395 to 975 nm in the visible and NIR spectral range. The original spatial resolution of the image was 2 m measured on ground, but in order to be treatable and

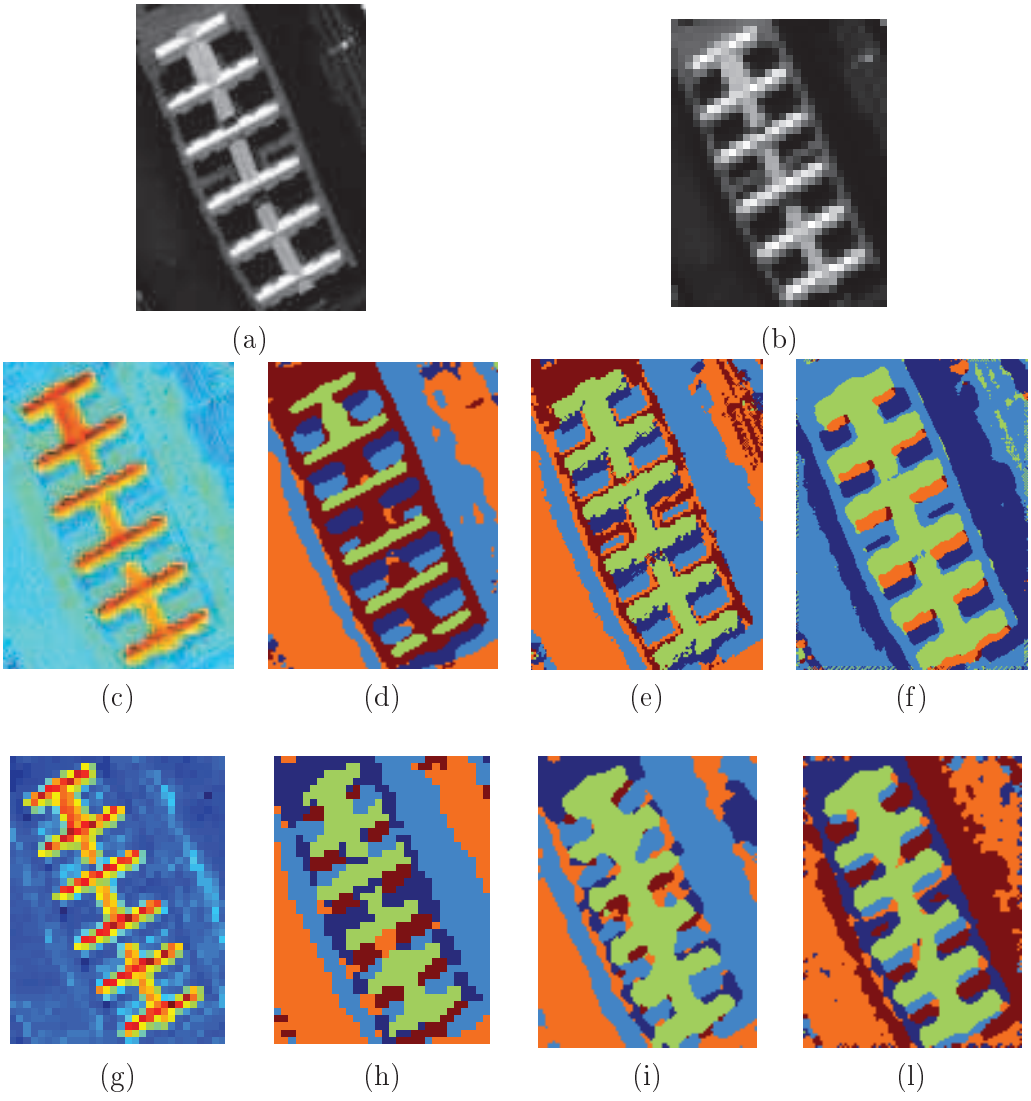


Figure 6.7: ROSIS University data set experiment: (a) Data set used in the experiment, band 30, original spatial resolution (b) Data set used in the experiment, band 30, spatial resolution degraded of a factor 3. (c) Fractional abundance map obtained with spectral unmixing (VCA + FCLS) (d) K-means classification map, after post-processing (e) Proposed method K-means+Spectral Unmixing classification map, after post-processing (f) Proposed method VCA+Spectral Unmixing classification map, after post-processing (g) Fractional abundance map obtained with spectral unmixing (VCA + FCLS) (h) K-means classification map, after post-processing (i) Proposed method K-means+Spectral Unmixing classification map, after post-processing (l) Proposed method VCA+Spectral Unmixing classification map, after post-processing.

still useful for the purposes of land cover interpretation it was downscaled to 6 m ground resolution while keeping the original spectral information as possible. The area is located in Hungary and contains arable lands near to the city of Heves. The area is mainly useful because of agricultural production. We considered a large subset of the image (400×500 pixels) containing six classes of interest.

The performances of the tested methods were evaluated in terms of overall accuracy (OA), that is the number of correctly classified test samples with respect to the total number oftest samples, average accuracy (AA), which represents the average of the classification accuracies for the individual classes, and the single classes accuracy. In the case of the ROSIS University experiment, where the classification of a single class is evaluated, the two accuracy indicators are equivalent.

6.6.1 ROSIS data set

The experiments carried out on the ROSIS data set are intended to evaluate the usefulness of the proposed method as a tool for structure detection. Two different tests were performed. The first one was on the original data, where all the pixels are considered as pure, in order to see the behavior of the proposed algorithms in such a situation. In the second experiment, the spatial resolution of the image was artificially degraded of a factor 3, so that the obtained images have a spatial resolution of 3.9 m, which is a realistic assumption in the case of airborne/satellite hyperspectral sensors. We introduce in this way a number of mixed pixels, useful to evaluate the performance of our method in such a situation. In order to have a comparison with a traditional unsupervised classification method, we have also classified both images with a K-means classifier. The number of classes to select was set to 5, after applying the Virtual Dimensionality method (setting the probability of false alarm to 0.001). Besides the number of classes, the only parameter which needs to be set in the proposed method is the threshold to determine whether a pixel can be considered as 'pure' after the first step. Instead of choosing an absolute value, we considered the difference between the two biggest abundances within a pixel, and set this value to 0.4. The decision to consider a relative value as threshold was taken by considering the characteristics of hyperspectral data, which are in general subject to high spectral variability. When performing spectral unmixing, endmembers which do not belong to a pixel could results in a small, but larger than zero abundance, mainly because of spectral variability or noise influence. With the proposed method, if a pixel contains two classes with abundances 0.65 and 0.35, it will be considered as mixed. However, if several classes are included in the pixel, the largest abundance being 0.65 and the others smaller than 0.2, the pixel will be considered as pure, since we assume

Table 6.2: Classification accuracies for the ROSIS data set experiment. In this experiment, only the producer accuracy for the main class of interest 'metal sheet' is considered. (KM = K-Means, SU = Spectral Unmixing, SA = Simulated Annealing, PS = Pixel Swapping, PP = Post Processing)

	ROSIS original		ROSIS low resolution	
	Before PP	After PP	Before PP	After PP
K-Means	50.86%	50.71%	93.75%	96.46
KM-SU-SA	95.89%	96.91%	97.10%	98.35
KM-SU-PS	95.24%	96.29%	96.91%	98.02
VCA-SU-SA	**96.95%**	**99.91%**	**97.12%**	**98.78%**
VCA-SU-PS	96.55%	99.76%	96.85%	98.43%

that low abundances are related to spectral variability and noise.

The performance of the methods was evaluated on the classification of the metal sheet structure present in the middle of the image. In order to have a quantitative comparison of the results, in the case of low resolution data, the classification map obtained with the traditional unsupervised classifier was evaluated by comparison with the low resolution ground truth available, while the proposed methods is evaluated by comparison with the high resolution ground truth data. However, we would like to highlight that all the methods take as input the low resolution data. A simple post-processing was applied to the classification map, in order to eliminate sparse pixels. For each pixel, a 3×3 window including its surrounding was used, and the value set to the most represented class within the window.
Both from Table II and Fig. 6.7 can be noticed the improvement provided by the proposed methods. Quite surprisingly, the K-means classifier provides better results in the case of low resolution data (also if the spatial accuracy of the method is clearly lower). The reason for this improvement is mainly due to two facts: 1) pixels labeled as "structure" in the low resolution data are composed by the average value of 9 pixels of the original image, this mitigating the problem of spectral variability 2) the number of pixels labeled as "structure" are much less than in the original case, since all the samples which were averaged with pixels belonging to other classes or unknown, were considered as mixed and therefore discarded from the ground truth.
It is highly remarkable that the proposed method obtains comparable results in the two cases, retrieving the metal structure as it is represented in the high resolution reference data. The qualitative improvement can be easily seen in Fig. 7.

6.6.2 AISA data set

The second experiment was carried out on the AISA data set, after reducing the spatial resolution of a factor 5. In spite of the high spatial resolution degradation, most of the pixels of the data set are to be considered as pure, since the image is mainly composed by large agricultural fields (high resolution ground truth represented in Fig. 6.8 (a)).
After the unsupervised classification, every cluster was assigned to the label of the class that was better represented, taking care that each cluster was assigned to only one class. The overall classification accuracy was then computed along with the accuracies of the single classes and the average class accuracies. As in the previous experiment, the K-means output map was compared with the low resolution ground truth obtained after filtering (not considering pixels which become mixed), while the proposed method was compared with the high resolution ground truth. We want to stress that this type of comparison is highly unfavorable to our method, which is expected to correctly classify pixels which are mixed in the input image, and to correctly locate the obtained sub-pixels in order to have the same spatial distribution of the original image, while for the K-means classifier these pixels are not considered in the ground truth. The post-processing considered in the first experiment was applied also in this case to eliminate isolated pixels in the classification maps.
The quantitative results are shown in Table III. Both proposed methods show better performances in terms of overall classification accuracy. The use of spectral unmixing with global endmembers results in a high percentage of mixed pixel, as it can be noticed from the improvement obtained with the classification post-processing. Instead, the unsupervised classifier with local endmembers shows slightly better results in terms of average class accuracy after the post processing step. The results of the experiment suggest that once the additional information about sub-pixel class abundances is retrieved by mean of spectral unmixing, the classification errors due to spectral variability can be easily corrected with a simple major voting post-processing.

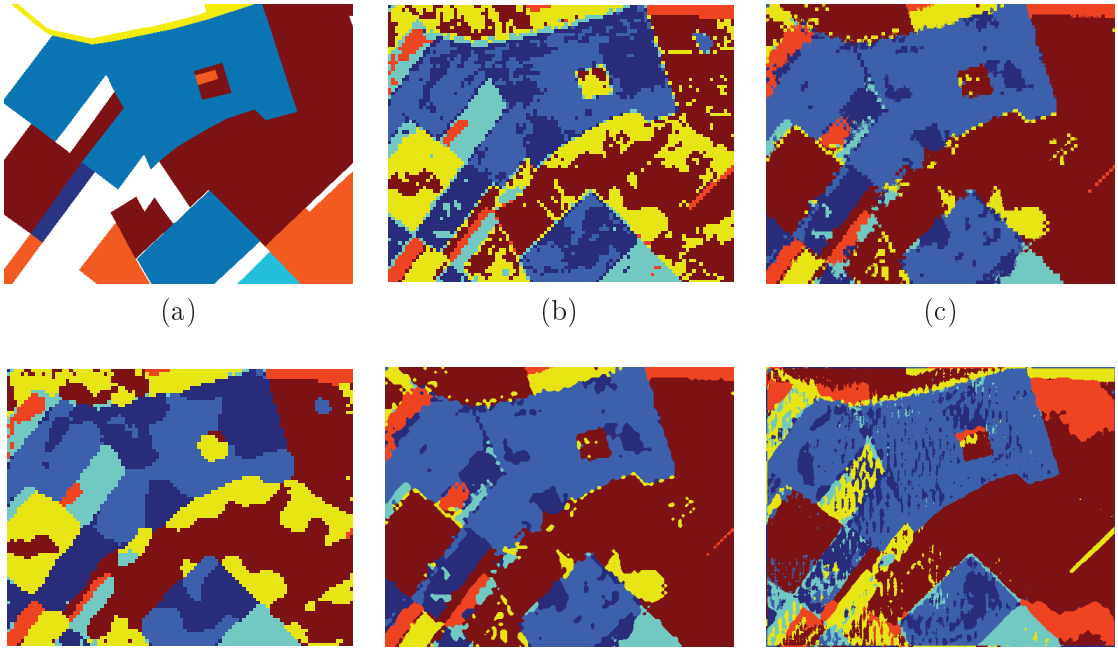


Figure 6.8: ROSIS data set experiment: (a) Original ground truth (b) K-means classification map, before post-processing (c) Proposed method K-means+Spectral Unmixing classification map, before post-processing (d) K-means classification map, after post-processing (e) Proposed method K-means+Spectral Unmixing classification map, after post-processing (f) Proposed method VCA+Spectral Unmixing classification map, after post-processing

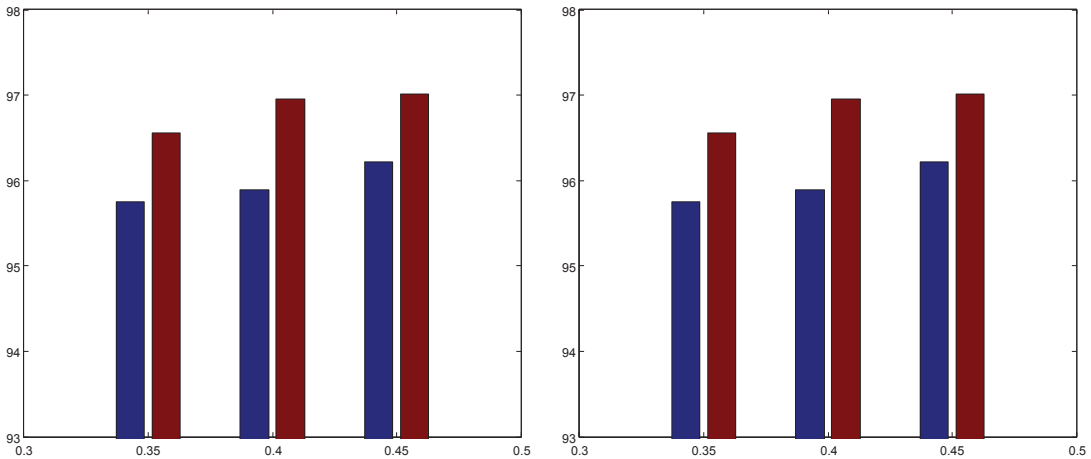


Figure 6.9: Variation of the overall classification accuracy versus the threshold parameter to determine the purity of a pixel: (a) AISA data set (b) ROSIS low resolution data set. Red bars: Simulated Annealing. Blue bars: Pixel Swapping.

Table 6.3: Classification accuracies for the AISA data set experiment. (KM = K-Means, SU = Spectral Unmixing, SA = Simulated Annealing, PS = Pixel Swapping, PP = Post Processing).

	AISA Data set - Before PP				
	K-Means	KM-SU-SA	KM-SU-PS	VCA-SU-SA	VCA-SU-PS
Overall Acc.	51.61%	**75.72%**	75.65%	59.69%	59.54%
Average Acc.	61.37%	**64.20%**	64.15%	56.83%	56.41%
Class 1	**93.75%**	59.72%	59.21%	58.26%	58.11%
Class 2	56.41%	87.34%	87.32%	67.87%	**67.88%**
Class 3	95.83%	**99.19%**	99.11%	30.69%	30.60%
Class 4	59.15%	46.90%	46.91%	**73.50**	73.21%
Class 5	6.67%	7.08%	7.00%	**51.19%**	51.13%
Class 6	56.44%	**85.03%**	84.97%	58.96%	58.87%
	AISA Data set - After PP				
	K-Means	KM-SU-SA	KM-SU-PS	VCA-SU-SA	VCA-SU-PS
Overall Acc.	52.75%	**76.24%**	76.11%	70.57%	70.51%
Average Acc.	65.60%	64.37%	64.21%	**65.73%**	65.70%
Class 1	**100%**	60.21%	60.22%	75.63%	75.61%
Class 2	55.75%	**88.07%**	88.01%	77.41%	77.40%
Class 3	**100%**	99.38%	99.38%	28.18%	28.12%
Class 4	71.83%	45.84%	45.64%	**94.75%**	**94.75%**
Class 5	6.67%	7.16%	7.18%	70.05%	**70.06%**
Class 6	59.09%	**85.58%**	85.50%	59.77%	59.60%

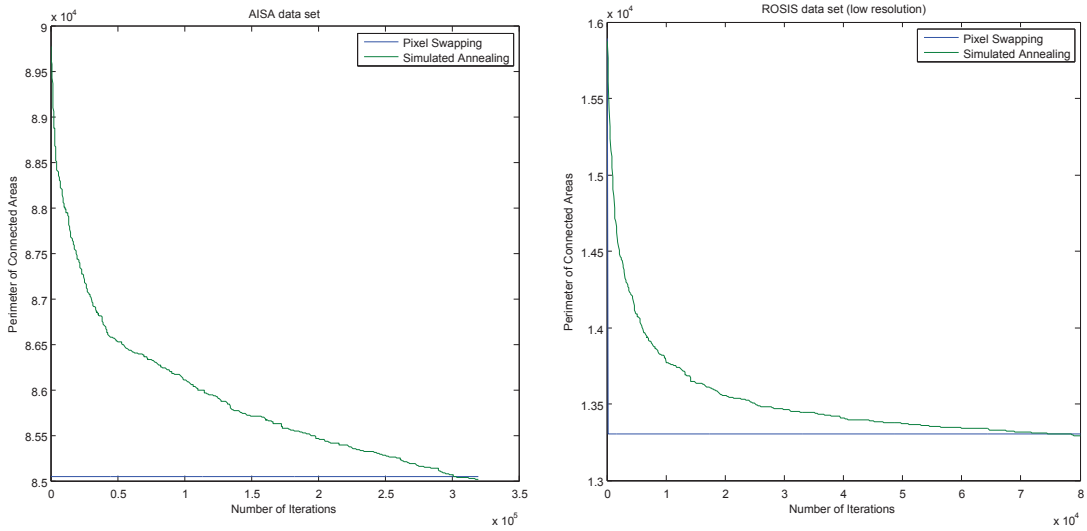
6.6.3 Discussion

The results obtained in the previous experiments prove that the proposed method has a very high potentiality for the unsupervised classification of hyperspectral images with low spatial resolution. In the case of highly mixed scene (as in the first considered data set, where there is a large number of mixed pixels with respect to the size of the data set), endmember extraction based methods provide better results. On the contrary, when the number of pure pixels is much higher than the number of mixed ones, the characteristics of clustering based methods have proven to be more suitable, as in the experiments conducted on the AISA data. The two approaches tested for sub-pixel locations have shown to be equivalent in terms of classification accuracy, with slightly better results provided by Simulated Annealing. In the proposed method, the only parameter having an influence on the overall classification accuracy obtained is the threshold to determine if a pixel can be considered as 'pure'. How the classification accuracy changes by changing the value of the parameter can be seen in Fig. 7.8. It can be noticed that the proposed method shows similar classification accuracies for the three tested values. The parameters, although very important for the classification, show a range of values for which the algorithms provide good results in terms of accuracy. As could be expected, a high value of the comparative threshold to determine if a pixel can be considered as 'pure' provides slightly higher accuracies, since only the most reliable pixels are labeled for the preliminary classification. By setting a low value of the threshold parameter, the preliminary classification map will tend to be like a common hard classification map obtained with a traditional classifier.
Regarding the computational burden, the two techniques considered to determine land cover classes, are equivalent in terms of processing time. The main difference between the methods tested in our experiments resides in the super-resolution algorithms used to locate sub-pixels.

Figure 6.10 shows the value of the perimeter of connected areas (which is the cost function considered in our experiments for Simulated Annealing) versus the number of iterations. It can be noticed that the Pixel Swapping method reaches the stability condition after few iterations (usually in the order of tens iterations), while Simulated Annealing, due to the random permutations performed, needs a much higher number of iterations to reach the same results. The computational burden required by Simulated Annealing is much higher in the case of AISA data set, due to the larger number of pixels and the higher resolution enhancement factor.

6.7 Conclusions

The unsupervised classification of hyperspectral images in presence of mixed pixels was addressed. Two methods for structure detection and improvement of the spatial resolution of classification maps were proposed. The method exploits the advantages of source separation, unsupervised classifiers and spectral unmixing algorithms, in order to determine the fractional abundances of the classes at a sub-pixel scale. A spatial regularization by Simulated Annealing is finally performed to spatially locate the land cover classes within each pixel. Experiments were carried out on a synthetic and two real data sets. The experimental results show that the proposed method clearly outperforms classical unsupervised classification techniques when areas with mixtures of materials are located in the scene, providing excellent results both from a visually and quantitative point of view. Further research will be devoted to the investigation of advanced methods to better discriminate pure and mixed pixels, and of the possibility of alternative techniques of spatial regularization.



(a) (b)

Figure 6.10: Perimeter of connected areas versus number of iteration for the two proposed super-resolution algorithms: (a) AISA data set (b) ROSIS low resolution data set.

Chapter 7

Spectral unmixing to obtain finer map resolution

Abstract

The integration of the super-resolution method into the framework of supervised classification is discussed in this chapter. When training set is available, a probabilistic SVM can be used to perform a preliminary classification. The results are the starting point to integrate the proposed method to improve the spatial resolution of the classification map. Comparison with a traditional SVM shows the effectiveness of the proposed method.

Contents

7.1 Introduction

The spatial resolution of hyperspectral images has a large influence on the classification accuracy which can be achieved. Although an increase of the spatial resolution does not lead automatically to better performances, the low spatial resolution makes appear problems connected to the presence of mixtures of materials [164]. In particular, the main assumption on which conventional supervised classification techniques for hyperspectral imagery are based is that the classes to be separated are discrete and mutually exclusive, i.e., it is assumed that each pixel vector contains information of a single spectral class. Often, however, this is not a realistic assumption. In particular, due to the tradeoff which is observed between spectral and spatial resolution, most of the pixels collected by hyperspectral imaging instruments contain the resultant mixed spectra from the reflected surface radiation of various constituent materials at a sub-pixel level. The presence of mixed pixels is due to several reasons [5]. First, the spatial resolution of the sensor is generally not high enough to separate different pure signature classes at a macroscopic level, and the resulting spectral measurement can be a composite of individual pure spectra (often called endmembers in hyperspectral analysis terminology) which correspond to materials that jointly occupy a single pixel. Second, mixed pixels also result when distinct materials are combined into a microscopic (intimate) mixture, independently of the spatial resolution of the sensor.

As pointed out in the previous chapters, one of the main limitation of remote sensing is handling of the intrinsic scale of variation of land cover, often being finer than the scale of sampling imposed by the satellite sensor [161]. The problem is particularly significant in the case of multispectral imagery, due to the tradeoff which generally exists between spatial and spectral resolution. A number of so-called full pixel techniques, based on the assumption that each pixel corresponds to the spectral signature of one predominant land cover type, have been proposed during the last decades for the classification of hyperspectral images [138,165]. These techniques are not suitable for the analysis of mixed pixels and will inevitably lead to a high error rate when used for scenarios with a high number of sites with mixtures of land cover classes. The issue of mixed pixels has been considered in several works. A widely investigated approach is the use of soft classification techniques [147]. These classifiers do not assign a pixel to only one class, but they produce a set of images (one per class) that express for each pixel the degree of membership in the class in question [147]. However, the membership degree does not necessarily reflect the fractional abundance of a class within a mixed pixel, and the probability of a pixel to belong to one class does not necessarily correspond to the fractional part of the pixel covered by the considered class.
Linear spectral mixture analysis (SMA) [94] is a soft classification technique explicitly designed to address this problem. Following the spectral mixing model, the spectral signature of a mixed pixel is assumed to be the weighted sum of some constituent spectra, also called endmembers. Spectral unmixing is the procedure by which the measured spectrum of a pixel is decomposed into a collection of endmembers, and a set of corresponding fractions, or abundances, that indicate the proportion of each endmember within the pixel. A number of techniques, exploiting both statistical and geometrical properties of the data, was proposed over the last few years [108,150]. These techniques can partially overcome the weakness of full pixel methods when analyzing mixed pixels. However, when used to obtain crisp classification maps, the endmembers selection and the abundances determination are negatively affected by spectral variability [166], and common hard classification methods are more suitable in such a case [167].
The remainder of the chapter is organized as follows. Section 7.2 presents the idea of the proposed approach. Section 7.3 presents in greater details the proposed approach. Section 7.4 shows the experiment on a synthetic data set, while Section 7.5 illustrates the experimental

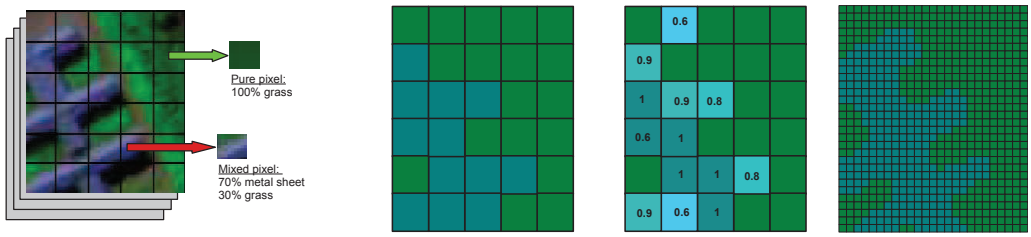


Figure 7.1: Illustration of the problem of mixed pixels in remote sensing images: (a) When the spatial resolution is not fine enough, several land cover classes lie in the same pixel. (b) In this case, a hard classification process cannot give an accurate information about the pixel coverage, leading inevitably to a loss of information (in the example figure, the classes of grass and metal sheet are considered for classification). (c) Spectral unmixing can provide information at a sub-pixel level, computing the fractional abundances of a class within each pixel. However, the supplementary information obtained does not result in a resolution enhancement of the classification map. (d) The proposed method tries to handle the problem by jointly using the ideas of spectral unmixing and unsupervised classification, in order to obtain thematic maps at a finer scale, without needing any additional source of information.

results on real hyperspectral data. Section 7.6 finally draws the conclusions.

7.2 Supervised super-resolution

In this chapter, a new supervised technique, which takes advantage of both probabilistic classification and spectral unmixing mapping techniques, is proposed in order to handle the issue of mixed pixels. The concept of sub-pixel mixing is also considered, in the attempt of obtaining land cover maps with an improved spatial resolution. The idea of subpixel mapping was first presented by Atkinson in [168]. He proposed to use the output of a soft classification technique in order to obtain a super-resolution mapping, trying to maximize the spatial correlation of the land cover classes to determine sub-pixels spatial locations. Since then, a number of techniques focused on better estimating sub-pixel fractional abundances determination and obtaining land cover maps with higher spatial resolution have been proposed [169]. In this chapter, we propose the use of Simulated Annealing (SA) for this purpose, due to its simplicity and ease of use. This method has shown good results in a number of optimization and real problems, and its wide range of parameters grants a high flexibility with respect to the analysed problem. In multi-hyperspectral remote sensing, it has successfully been used for classification [148, 170], and abundances estimation [171]. The method proposed in this chapter is in three steps. In a first step, a coarse classification is performed, based on the probabilistic output of an SVM. Every pixel can be assigned to a class, if the probability value obtained in the classification process is greater than a chosen threshold, or be unclassified. Pixels with a low probabilistic output are either mixed pixels or pixels hard to classify due to spectral variability, and their classification is addressed in a second step. In the second step, spectral unmixing is performed on the unclassified pixels by considering the preliminary results of the coarse classification step and by applying a Fully Constrained Least Squares (FCLS) method to every unlabeled pixel, in order to obtain the abundances fractions of each land cover type. Finally, in a third step, spatial regularization by SA is performed to obtain the resolution improvement. Experiments are carried out on synthetic and real hyperspectral data sets. The results are excellent both numerically and visually and show that the proposed method clearly outperforms traditional hard classification methods when the data contain mixed pixels.

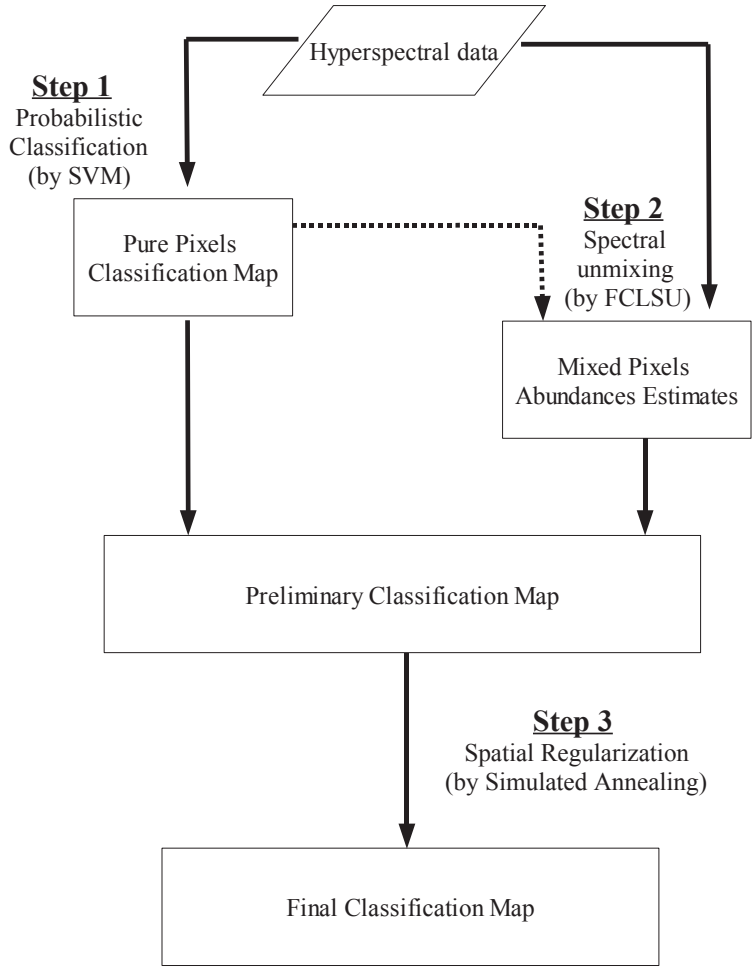


Figure 7.2: Flow chart scheme of the proposed approach to obtain thematic maps at a finer spatial scale in a supervised classification framework.

7.3 Methodology

The flow chart scheme of the proposed approach is presented in Fig. 7.2. The hyperspectral data are used as input for the hard classification method, in order to obtain a preliminary classification of all the pixels considered as "pure". The results of this step are the input (along with the original hyperspectral image) for the spectral unmixing, so that an appropriate set of endmembers can be found and the negative impact of spectral variability on the classification map minimized. In the last step, the results obtained are processed with a Simulated Annealing algorithm. Based on the assumption of spatial correlation of the land cover classes, SA is used to optimize a function where spatial proximity of pixels belonging to the same land cover class are preferred to the opposite case.

7.3.1 Pixel-wise classification

The first step of the proposed method consists in performing a pixelwise classification of the hyperspectral image, in order to obtain, for every pixel, a probability value for it to belong to one of the land cover classes. The pixels with a probability higher than a chosen treshold are considered as pixels where a single class is represented, and thus assigned to the considered class. These pixels are going to provide a preliminary classification map, where only the pixels containing a predominant land cover class are labeled. All the other pixels

Figure 7.3: Basic steps of the proposed method: (a) A probabilistic classification map is computed for each class. (b) The pixels with highest probability greater than a chosen treshold are considered as pure and classified (in the figure, we set the treshold to 70%). The other pixels are considered as mixed (MIX in the figure). (c) For each mixed pixel, a set of possible endmembers is selected, considering the results of the preliminary classification. The other pixels, pure or mixed, are just ignored. (d) Spectral unmixing provides information about the abundance fraction of a class within each pixel. (e) Pixels are split into n sub-pixels, according to the desired zoom factor, assigned to an endmember and randomly positioned within the pixel. The number of sub-pixels assigned to each class reflects the fractional value estimated in the previous step. (f) Simulated annealing performs random permutations of the sub-pixels position until *minimum* cost is reached.

are not labeled, and their classification will be addressed in a second step.
As long as it can provide a probabilistic output, every classifier can be used for the data analysis. In this work, we propose to use a probabilistic Support Vector Machine (SVM) classifier [172], due to the good performances shown in the classification of hyperspectral data, also in case of limited training sets [15, 173, 174]. In the following, we give a short description of the principles on which SVM is based. Due to the lack of space, we refer the reader to [172] and [173] for further details on the theory of SVM and its application for hyperspectral images classification.

7.3.2 Probabilistic Support Vector Machine

The SVM is surely one of the most commonly used kernel learning algorithm. It performs robust non-linear classification of samples using the kernel trick [175]. The idea is to find a separating hyperplane in some feature space induced by the kernel function while all the computations are done in the original space [172]. Given a training set $\mathcal{S} = \{(\mathbf{x}^1, y_1), \ldots, (\mathbf{x}^\ell, y_\ell)\} \in \mathbb{R}^n \times \{-1; 1\}$ where vector $\mathbf{x}$ contains the band intensities and y the class label, the SVM computes a decision function f(x) such that sign(f(x)) can be used to predict the label of any test sample x. The decision function can be written as

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i \widetilde{k}(\mathbf{x}^i, \mathbf{x}^j) + b \tag{7.1}$$

and it is found by solving the convex optimization problem:

$$\begin{aligned} \max_{\alpha} \; g(\alpha) \quad &= \quad \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \widetilde{k}(\mathbf{x}^i, \mathbf{x}^j) \\ \text{subject to} \quad &\quad 0 \leq \alpha_i \leq C \text{ and } \textstyle\sum_{i=1}^{\ell} \alpha_i y_i = 0 \end{aligned} \tag{7.2}$$

where α are the Lagrange coefficients, $\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \delta_{ij}/C$, k the kernel function, C a constant that is used to penalize the training errors, δ_{ij} a function such that $\delta = 1$ if $i{=}j$, $\delta = 0$ otherwise. To be an acceptable kernel, k should be a positive semi-definite function [175].

In [176] Platt proposes approximating the posterior class probabilities $P(y = 1|x)$ by a sigmoid function:

$$P(y = 1|x) \approx P_{A,B}(f) = \frac{1}{1 + \exp(A\hat{f} + B)} \tag{7.3}$$

where $\hat{f}$ is an estimation of the decision function f(x) computed by the SVM, A and B two parameters that need to be optimized. The best parameter setting $(A^\star, B^\star)$ is determined by solving the following regularized maximum likelihood problem (with N_+ of the y_i's positive, and N_- negative training samples):

$$\min_{z=A,B} \quad F(z) = \sum_{i=1}^{l}(t_i log(p_i) + (1-t_i)(1-log(p_i)), \tag{7.4}$$

where

$$p_i = P_{A,B}(f_i), \quad t_i = \begin{cases} \frac{N_+ +1}{N_+ +2} & \text{if} \quad y_i = +1 \\ \frac{1}{N_- +2} & \text{if} \quad y_i = -1 \end{cases}, \quad i = 1, \ldots, l.$$

A detailed description of the method can be found in [176]. In this work, we have used an improved implementation of the above algorithm [177], which is included in the LIBSVM library [79].

In this first classification step, we consider two outputs:

1) A complete probability map, containing the probability estimates for each pixel to belong to the assigned class.

2) A coarse classification map of the pixels considered as 'not mixed', containing class labels for the samples with a probability belonging to the class higher than a chosen threshold.

It is not a straightforward task to choose the threshold to determine if a pixel should be considered as pure or mixed. When labelling the pure pixels, we are interested in correctly classifying most of the pure pixels, because of two already mentioned reasons: i) When dealing with pure pixels, the full pixel methods work better than spectral unmixing, ii) this preliminary classification will be the input of the second step, and a large number of correctly classified pure pixel helps to provide suitable endmember candidates for the mixed pixels. For this same reason, the misclassification of pure pixels could lead to critical issues and cause a large error in the spectral unmixing step, thus a tradeoff is observed. The experiments carried out to investigate this issue have shown that in general a high threshold (close to 80%) allows to obtain a higher classification accuracy, since only the pixels which are reasonably sure to belong to a class are labeled. However, the choice of this parameter is not crucial for the classification accuracy of the proposed method, as it will be shown in the next sections.

7.3.3 Spectral Unmixing

After obtaining a coarse classification map, where some pixels considered as "pure" (due to the high probability to belong to the assigned class) were classified, the labelling of the other pixels is addressed in the second step.
Spectral mixture analysis (SMA) techniques have overcome some of the weaknesses of full pixel approaches by using linear statistical modeling and signal processing techniques [94,108]. They are inherently either nonlinear techniques or linear techniques. Nonlinear mixed pixel analysis estimate multiple scattering effects that may arise when the different materials form intimate association at microscopic level [94]. Although they can be useful for some types of analysis, in the majority of applications a linear mixing model can be considered without significant loss of information [94]. The key task in linear SMA is to find an appropriate set of pure spectral constituents -called "endmembers" in hyperspectral analysis terminology-,

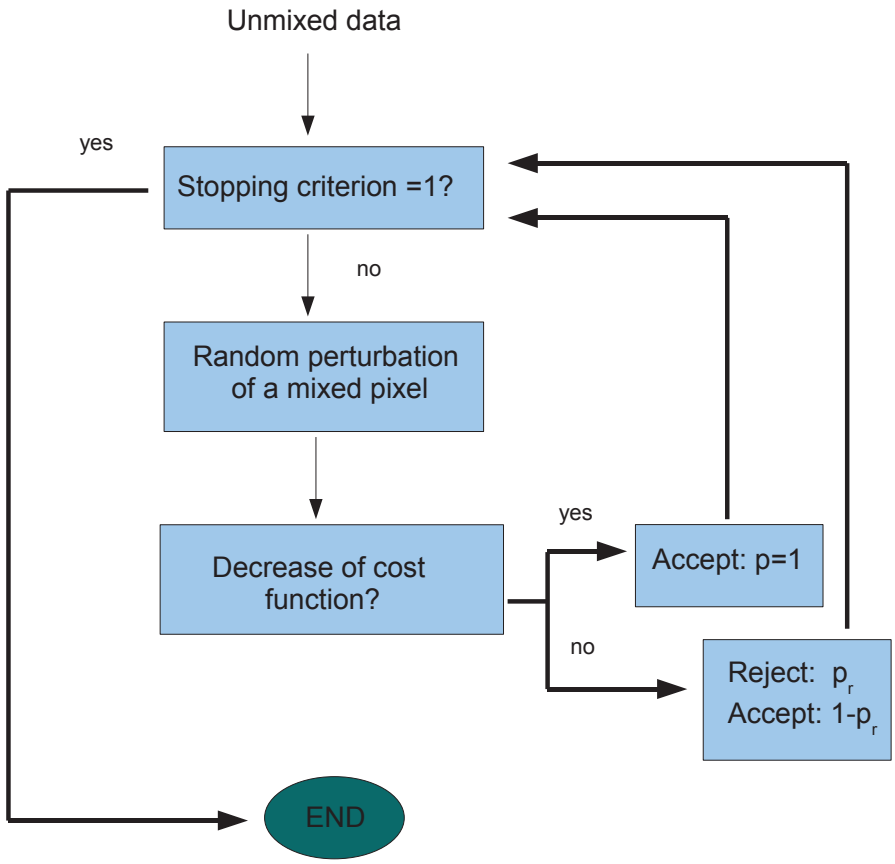


Figure 7.4: Flow chart scheme of the proposed spatial regularization approach based on Simulated Annealing.

which are then used to estimate the fractional abundances of each mixed pixel from its spectrum and the endmember spectra by using a linear mixture model.
In the Linear Mixture Model (LMM), the spectrum of a mixed pixel is represented as a linear combination of component spectra (endmembers). The weight of each endmember spectrum (abundance) is proportional to the fraction of the pixel area covered by the endmember. If there are M spectral bands, the spectrum of the pixel and the spectra of the endmembers can be represented by M-dimensional vectors. Therefore, the general equation for LMM is described as a linear regression form

$$\mathbf{z} = \sum_{\mathbf{i=1}}^{\mathbf{L}} \mathbf{a_i s_i} + \mathbf{e} = \mathbf{As} + \mathbf{e} \tag{7.5}$$

where $\mathbf{z}$ is an $M \times 1$ column pixel vector which describes the spectrum of the mixed pixel, $\mathbf{s} = [\mathbf{s_1 s_2} \ldots \mathbf{s_L}]$ is an $M \times L$ endmember matrix of material signature, $\mathbf{s_i}$ $(i = 1, 2, \ldots, L)$ are the M-dimensional spectra of the endmembers, $\mathbf{a}$ is an $L \times 1$ column vector and is composed of abundance coefficients $\mathbf{a_i}$ $(i = 1, 2, \ldots, L)$, $\mathbf{e}$ is an M-dimensional error vector accounting for lack-fit and noise effects, and L is the number of the endmembers. Due to physical reasons, (7.5) has to respect the following constraints of non-negativity (abundance fractions within a pixel cannot be negative) and sum to one (the sum of all the abundances

fraction within a pixel must have 1 as a result):

$$a_i \geq 0 \tag{7.6}$$

$$\sum_{i=1}^{L} a_i = 1. \tag{7.7}$$

In recent years, several algorithms have been developed for automatic or semi-automatic extraction of spectral endmembers directly from the image data and to determine their fractional abundances within each pixel [150]. Assuming that ground truth is available, we do not need to determine the endmembers composing the data, but simply the abundance of each land cover type within the pixels. In this case, a major issue is how to handle the spectral variability which affects the data. As shown in [156], soft classification of hyperspectral images covering wide areas is negatively related to the intra-class spectral-variability, and the assumption that a single endmember could extensively represent a class is generally far from reality. The choice of appropriate endmembers is very important in order to correctly estimate the fractional abundances. If the endmembers do not represent the land cover classes well, the estimates of the sub-pixel coverage can be highly biased and lead to misclassification errors.
In order to overcome this problem, we propose an adaptive approach to select the best endmember candidates for each pixel. This approach is based on two main assumptions:

1. The spatial correlation of the classes, i.e., for each pixel, it is probable that the best endmember candidates lie in the spatial proximity of the considered pixel.

2. The probabilistic output provided by the SVM, i.e., if a candidate is not spatially close to the selected pixel, but the probabilistic value of the class to which it belongs is high, it is presumably a good candidate.

For each mixed pixel which has to be classified, we consider a set of 10 different spectra, that represent the endmember candidates. The reason for this choice is the above mentioned spectral variability: It is hard to represent a class with a single endmembers. The following experiment will show that this number can be varied within a wide range of values without affecting the performances of the proposed method. These candidates are chosen from the labeled samples of the training data and the results of the preliminary classification of step one, considered as a set of pure pixels correctly classified. If one of the land cover classes has a high probabilistic output (we consider a probabilistic output as high if its difference from the threshold chosen at step 1 is smaller than 5%), at least five spectra of this class are considered, otherwise all the 10 candidates are selected from the spectral signatures spatially closest to the considered pixel, after the coarse classification step. Once the spectral signatures representative of each class are extracted from the image, the abundance fraction of the elements within each pixel should be determined. Several algorithms have been developed for the linear mixing model according to the required constraints of abundances fractions. The fully constrained least squared unmixing algorithm is a widely adopted practical solution to avoid the appearance of theoretical problems, such as negative fractional abundances or abundances that sum up to more than one. Due to its computational efficiency, we have chosen this algorithm, which satisfies both abundance constraints and is optimal in terms of least squares error [109]. After applying FCLS, we obtain the fractional abundances of each endmember. Due to the fact that in many cases several endmembers represent the same class, by summing the fractional abundances of all the endmembers belonging to the same land cover class, we obtain the cover percentage of a class within a mixed pixel. It should

be noted that not all the pixels to be classified in this second step are mixed, but there are many 'pure' pixels not labeled because of the low probability output provided by the SVM. However, the proposed method allows to label them as 'pure' pixels also in this second step.

7.3.4 Improving Spatial Resolution

Spectral unmixing is useful to describe the scene at a sub-pixel level, but can only provide information about proportions of the endmembers within each pixel. Since the spatial location remains unknown, spectral unmixing does not perform any resolution enhancement. In this chapter, we propose a sub-pixel mapping technique, which takes advantage of the information given by the spectral mixing analysis and uses it to enhance the spatial resolution of thematic maps. Our proposed approach is as follows: In a first step, each pixel is divided in a fixed number of sub-pixels, according to the desired resolution enhancement. Every sub-pixel is assigned to an endmember, in conformity with its fractional abundance within the pixel. For example, if we want to have a zoom factor of N, we have to divide each pixel into N×N sub-pixels. For each pixel, the number of subpixels n to assign to the class i is computed according to the equation:

$$n_i = \text{round}\Big(\frac{\text{abd}_i}{1/N^2}\Big), \tag{7.8}$$

where abd_i is the fractional abundance of the class i within the considered pixel estimated with the FCLS and round(x) returns the value of the closest integer to x.

A Simulated Annealing (SA) mapping function is then used, to create random permutation of these sub-pixels, in order to minimize a chosen cost function. Relying on the spatial correlation tendency of landcovers, we assume that each endmember within a pixel should be spatially close to the same endmembers in the surrounding pixels. Therefore, the cost function C to be minimized is chosen as the perimeter of the areas belonging to the same class:

$$C = \sum_{i=1}^{I} \sum_{j=1}^{N_i} P_j, \tag{7.9}$$

where I is the number of the classes, N_i is the number of connected components of the class i, and P_j is the perimeter of the connected component j, computed according to the 8-connected border pixels model [157].

SA is a well established stochastic technique originally developed to model the natural process of crystalization [158]. This process is based on an analogy from thermodynamics where a system is slowly cooled in order to reach its lowest energy state. More recently, SA has been proposed to solve global optimization problems [159], and it has been used in various fields. The basic idea of the method is that, in order to avoid to be trapped in local minima, uphill movements, i.e., points corresponding to worse values of the objective function could, sometimes, be accepted for the following iteration. As with a greedy search, it accepts all the changes that improve the solution. Changes degrading the solution can be accepted, but with a probability that is inversely proportional to the size of the degradation (small degradations are accepted with a higher probability). This probability also decreases as the search continues, or as the system cools down, allowing eventual convergence to the optimal solution.

An example of how SA spatial regularization works can be seen by looking at Fig.7.3, where 7.3 (e) represents the initial sub-pixel distribution and 7.3 (f) the optimal one. First, a mixed pixed is selected, according to the information provided by the spectral unmixing step. Then, a random permutation of the subpixels within the chosen pixel is performed by SA. If this

Table 7.1: Simulated data set. Description of the pureness of each pixel. abd represents the *maximum* value of abundance within a pixel

Pureness	Number of Pixels
abd>95%	7105
95%>abd>85%	828
85%>abd>75%	484
75%>abd>65%	576
65%>abd>55%	384
abd<55%	323

permutation leads to a decrease of the cost function (which is in our case the perimeter of connected components, i.e., components belonging to the same class), the change is accepted. Otherwise, as decribed above, the change will be probably rejected. The algorithm stops when *minimum* cost is reached, that is when a previously fixed number of consecutively rejected changes is reached (in our case, we set the number of consecutive operations to 100.000, since it is large enough to avoid sub-optimal solutions and Matlab takes only a few seconds to perfom this computation).

7.4 Experiments on simulated data

The first experiment was carried out on a synthetic data set. The advantage of using a synthetic data set is the perfect knowledge that we have about the analysed image. This experiment has two main aims: the first one is to verify the assumption that the pixels classified by the SVM with a high probability value effectively correspond to 'pure' pixels. The second one is to validate the proposed method with a data set known in details, in order to evaluate the performances under different conditions of mixtures. The creation of a synthetic data set for classification purposes is not a trivial task, especially in this case where the spatial information is very important to fully exploit the potentiality of the proposed method. In order to build a data set as realistic as possible, we have considered a thematic map of an AVIRIS image, taken over the area of San Diego, with 9 land cover classes. We have created a hyperspectral data cube by substituting every class with a spectrum taken from the USGS spectral library, available on-line [163]. The classes were chosen mainly from the vegetation library, in order to make more difficult the discrmination. The chosen classe were: shadow, asphalt, green grass, dry grass, maple, pine lodgepole, pine white, pinon and rub. The original image created was 400×400 pixels. Gaussian noise was added in order to reach a SNR of 30 dB. To have the possibility to analyse data sets where the ground truth cover is known in details, and to evaluate the obtained results from a quantitative point of view, we decided to use the original ground truth data only to compare the obtained results, and to decrease the spatial resolution of the image by applying an 4×4 low pass filter, so that we obtained an image of the same area with a resolution degraded of a factor 4. This way, we have the possibility to test the proposed method on a data set know in details, where there are pure, close to pure, and mixed pixels. The complete description of the data set is given in Table 7.1. In order to compare the proposed approach with a common hard classification method, the same data were also classified with an SVM wih Gaussian Kernel, One vs One multiclass strategy and 10 fold cross-validation. Among the several multi-class strategies available for the SVM, we have chosen the One vs One because of its good performances in

Table 7.2: Results of the experiment on the synthetic data set. abd represent the *maximum* value of abundance within a pixel. The third column describes the average of the maximum probability value within a group of pixels with the same purity degree. The last column represents the number of pixels with a maximum probability value greater than 0.7, according to their purity degree.

	SVM	Proposed Method	Average Probabilistic SVM Output	Max(prob)>0.7
OA	88.21%	90.84%	-	-
κ	87.34%	89.52%	-	-
AA	88.32%	90.25%	-	-
abd>95%	99.50%	99.50%	0.74	5317
95%>abd>85%	92.25%	91.44%	0.68	328
85%>abd>75%	79.57%	77.57%	0.54	27
75%>abd>65%	54.12%	61.17%	0.39	5
65%>abd>55%	34.82%	50.33%	0.30	0
abd<55%	15.36%	40.22%	0.26	1

terms of robustness and computational burden [178]. The comparison of the low resolution map obtained with SVM to the high resolution ground truth was not possible due to different number of pixels of the two images. However, we know that every pixel of the low resolution image corresponds to $n \times n$ pixels of the high resolution image. By comparing a pixel of the low resolution classification map with the $n \times n$ corresponding in the high resolution ground truth map, we can compute per-pixel classification accuracy. By doing this, we have to keep in mind that in case of a mixed pixel the hard classification method will inevitably lead to an error, because it will assign the corresponding high resolution $n \times n$ pixels to just one class, considered as predominant within the mixed pixel. However, this is exactly the issue that the proposed method is expected to address.

The performances of the two methods were compared in terms of overall accuracy (OA), that is the number of correctly classified test samples with respect to the total number of test samples, average accuracy (AA), which represents the average of the classification accuracies for the individual classes, and the Kappa coefficient of agreement (κ), that is a parameter that estimates the correct percentage classification without the amount that could be expected due to chance alone [80]. In addition to this, we have computed the number of mixed pixels correctly classified, in order to show the performances of the two methods when dealing with mixtures of classes, and the number of pixels correctly labeled after the spectral unmixing step but incorrectly positioned after Simulated Annealing.

For each class, 2% of the labeled samples were selected for training the algorithm. The threshold between pure and mixed pixels has been set to 0.7. In this approach, we have decided to use a fix threshold rather than a relative one as in the previous chapter, because of two main reasons: the higher number of classes and the increased robustness of the method, which can rely on the information given by the training samples. Because of these reason, the choice of an absolute threshold was considered. The results of the experiment are shown in Table 7.2. The first three columns represent the results obtained on the entire data set in terms of OA, κ and AA, and show that the proposed method provide an improvement in the overall accuracy classification. The second part of the Table II show a comparison of the performance of the two methods over groups of pixels with different degrees of purity, varying from pure pixels (where the predominant class has an abundance larger than 95%) to highly

mixed pixels (where the predominant class has an abundance smaller than 55%). It can be noticed that while in case of pure and close to pure pixels the results of the two methods are quite similar, the proposed approach provided a dramatic increase of classification accuracy for mixed and highly mixed pixels, where a traditional classifier completely fails while the proposed method improves the accuracy of up to 35 percentage points. In order to evaluate the correctness of the assumption that pure pixels are classified with a high probability value, we have computed the mean probability of the *maximum* value of each group of pixels and the number of pixels of each group which are classified with a probability value higher than 70% (two leftmost columns of Table 7.2). It can be easily noticed that the larger is the *maximum* value of abundance within the pixel, the higher is the probabilistic output provided by the SVM. The 80% of pure pixels were classified with a probability larger than 0.7, while in case of mixed and highly mixed pixels this quantity drops to 0.4%.

7.5 Experiments on real data

The experiments on real data were carried out considering three different data sets from two hyperspectral images. The first two data sets are from an Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) image taken over NW Indiana's Indian Pine test site in June 1992. This image has been widely used in the remote sensing community for both classification and spectral unmixing purposes, and thus represents an interesting benchmark for the proposed method. According to [179], we considered for the first experiment a part of the scene, consisting of pixels [31-116] $\times$ [27-94] for a size of 86 $\times$ 68, which contains four labeled classes (the background pixels were not considered for classification purposes). We will refer to this data set as the "Subset scene". The second experiment was carried out on the whole AVIRIS data set. Sixteen land cover classes were considered. The original image is composed by 145 $\times$ 145 pixels. The calibrated data are available online[1] with detailed ground-truth information.

Finally, 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 10 years, in 1970, 1980-81, 1991 and in 2000. The volcano is located on the South-Western 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 AVIRIS sensor operates in the visible, near- and mid- infrared portions of the electromagnetic spectrum, its sensitivity range spanning wavelengths from 0.4 μm to 2.4 μm. As on the previous case, the sensor system has 224 data channels, utilizing four spectrometers, whereas each spectral band is approximately 10nm in width. During the image acquisition, spectrometer 4 was not working properly. This particular spectrometer operates in the wavelength range from 1.84 μm to 2.4 μm (64 bands). These 64 bands were deleted from the imagery along with the first channels for all the other spectrometers, and the remaining 157 data channels were left. A subset of 180 $\times$ 180 pixels has been used for this experiment. In order to address the issue of the random choice of the training samples, for each data set we have repeated the experiment with ten different training sets.
As in the previous experiment, due to the difficulty to have a perfect knowledge of the fractional abundances of each land cover type, we decided to use the original ground truth data only to compare the obtained results, and to decrease the spatial resolution of the image by applying an $n \times n$ low pass filter, where n varies according to the considered data set.

[1] http://dynamo.ecn.purdue.edu/~biehl/

Table 7.3: Information about the training and the testing set of the three considered data sets. It has to be noticed that the training set is selected from the low resolution image used as input of the method, while the test set is selected from the high resolution reference data used for comparison.

	AVIRIS Indian Pine Subset			AVIRIS Indian Pine Complete			AVIRIS Hekla		
No.	Name	Train	Test	Name	Train	Test	Name	Train	Test
1	Corn-no till	20	1434	Alfa Alfa	4	54	Andesite lava 1970	24	672
2	Grass-Trees	20	747	Corn-no till	44	1434	And. lava 1980 I	126	3350
3	Soybean-no till	20	727	Corn-min till	25	834	And. lava 1980 II	523	11916
4	Soybean-min till	20	1926	Corn	7	234	And. lava 1991 I	220	4709
5	-	-	-	Grass-Pasture	15	497	And. lava 1991 II	279	6918
6	-	-	-	Grass-Trees	23	747	Lava tephra covered	103	2310
7	-	-	-	Grass-Mowed	4	26	Rhyolite	6	181
8	-	-	-	Hay-windrowed	17	489	Scoria	51	1286
9	-	-	-	Oats	4	20	Firn-glacier ice	42	1058
10	-	-	-	Soybean-no till	32	968	-	-	-
11	-	-	-	Soybean-min till	83	2468	-	-	-
12	-	-	-	Soybean-clean t	19	614	-	-	-
13	-	-	-	Wheat	5	212	-	-	-
14	-	-	-	Woods	42	1294	-	-	-
15	-	-	-	Bldg-Trees-Drive	12	380	-	-	-
16	-	-	-	Stone-Steel Tower	4	95	-	-	-

This way, we know exactly the quantity of each class within a pixel, and we can use the low resolution image obtained after filtering as input for the proposed method. The information about the classes, the training and the test sets can be found in Table 2.1.

7.5.1 AVIRIS subset

The first experiment was carried out on the AVIRIS subset image. The goal of this experiment is to illustrate the effectiveness of the method when used for the analysis of a simple hyperspectral data set. In this subimage, composed by 86×68 pixels, there are four classes with uneven number of labeled samples, namely, "Corn-notill", "Grass/Trees", "Soybeans-no till", and "Soybeans-min". The complete description of the training and test sets can be found in Table 2.1. A 3×3 low-pass filter was applied to the original image, so that a new image with lower spatial resolution was obtained. The new image was composed by 28×23 pixels, and it was used as input for the proposed method. The low resolution image obtained after filtering and the ground truth can be seen in Fig. 7.5 (a-b). Twenty pixels per class, considered as "pure" in the low resolution image, were randomly chosen and used for training the SVM classifier.
The result of the classification with the SVM is presented in Fig. 7.5 (c). As it can be seen from the classification map, the two main problems are represented by the mixed pixels, which make hard to distinguish the border between different land cover areas, and the high spectral variability, which results in a noisy classification map. The proposed method provides an overall improvement for both issues (classification map shown in Fig. 7.5 (f)). It can be seen at the top of the image that the border of the Corn-no till field (represented in light blue in the map) is estimated with improved accuracy when compared to the traditional

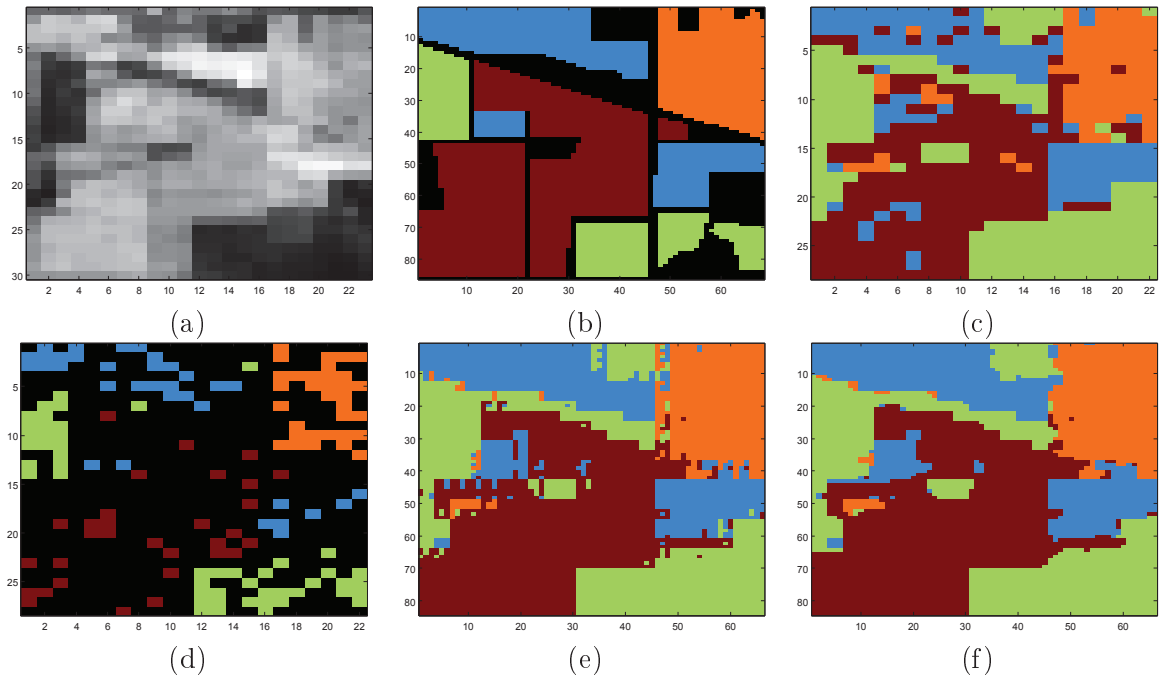


Figure 7.5: AVIRIS subset data: (a) Low resolution image (band 30) obtained after applying a 3×3 filter to the high resolution image. (b) Ground truth of the high resolution image. Unknown pixels are represented in black. (c) Classification map obtained with traditional SVM. (d) Results of the preliminary classification. (e) Classification map obtained before applying the spatial regularization. (f) Final classification map obtained with the proposed method.

SVM classification, thus assessing the effectiveness of the proposed approach also to provide classification maps with a better spatial resolution. The problem of spectral variability is also solved, since the classification map is much less noisy and, therefore, a lower number of sparse pixels is observed.
The quantitative results obtained with the two methods are presented in Table 7.4, along with the results of the other data sets. The proposed method provides an improvement of the overall accuracy which is greater than 10%. As it will be in the other experiments, the improvement of the classification accuracy of the mixed pixel with respect the SVM is larger than for the whole data set. This demonstrates that the proposed method is effective in improving the results of data sets with mixtures of land cover classes. To assess the effectiveness of SA to locate sub-pixels in the classification map, we have also computed the number of sub-pixels correctly classified after the spectral unmixing step, but incorrectly located after the spatial regularization. In this case, we can see that the error due to bad positioning of sub-pixels is extremely low.

7.5.2 AVIRIS complete

The second experiment was carried out on the whole AVIRIS data set. Sixteen land cover classes were considered for classification. The original image is composed by 145 $\times$ 145 pixels, and it was used as reference data. After applying a 2$\times$2 low pass filter, an image composed by 72$\times$72 pixels was obtained. The land cover ground truth can be seen in Fig. 7.6 (a). For training set, we have randomly selected, for each class, 15% of all the samples which were considered as "pure" in the low resolution image (that would correspond to about 10% of pixels of each class in the high resolution image). To have the possibility to compare the results of the proposed method with the available ground truth, we chose a zoom factor equal to 2, lower than in the previous case. However, the higher number of classes and their spectral similarity make this data set more challenging than the first one.
Figure 7.6 (b) and (e) shows the classification maps obtained with a conventional SVM and the proposed method. Also in this case, an improvement can be clearly seen in the classification maps, resulting in a less noisy map and an improved detection of the borders of spatial structures (in this case, agricultural fields). To have a quantitative comparison of the results obtained with the two methods, the overall accuracy of pixels correctly classified has been compared. The mean overall accuracy obtained in the five experiments with the SVM is 72.31%. As in the previous case, the low value of accuracy is due to two main factors, which are the impossibility of a common hard classification technique to distinguish different land cover classes at a sub-pixel level, and the difficulty to handle the high spectral variability. The proposed method obtained an average overall accuracy of 91.10%, showing the capability of the proposed approach to better deal with the aforementioned two main issues. By comparing Figure 7.6 (d) and (e), it can be noticed the effectiveness of the proposed spatial regularization with Simulated Annealing.

7.5.3 Hekla data set

For the last experiment, we consider a subset of the Hekla data set, located in the top-left corner of the scene. This subset is composed by 180 $\times$ 180 pixels, and it contains nine classes of interest. Also in this case a 2$\times$2 low pass filter was applied to the original image, leading to a low resolution image of 90 $\times$ 90 pixels. Due to the insufficient availability of ground truth to quantify the results provided by the proposed method, we have considered as ground truth the classification map obtained by a spectral-spatial method, proposed in [46], where the overall accuracy computed on the reference test set was close to 100%. Thus, also if we have to keep in mind that the results are estimated by comparison with a classification map

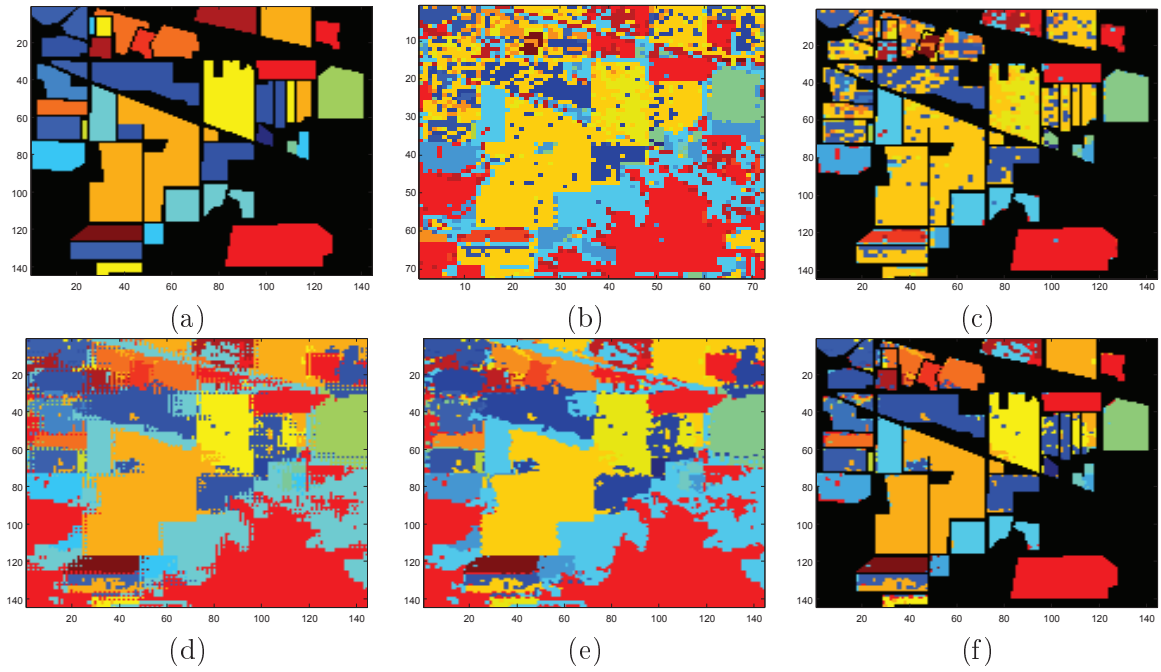


Figure 7.6: AVIRIS whole data set. (a) Ground truth data (b) Classification map obtained with an one versus one SVM (training set 1) (c) Classification of the ground truth pixels (d) Classification map obtained before applying the spatial regularization. (e) Final classification map obtained with the proposed approach. (f) Classification of the ground truth pixels.

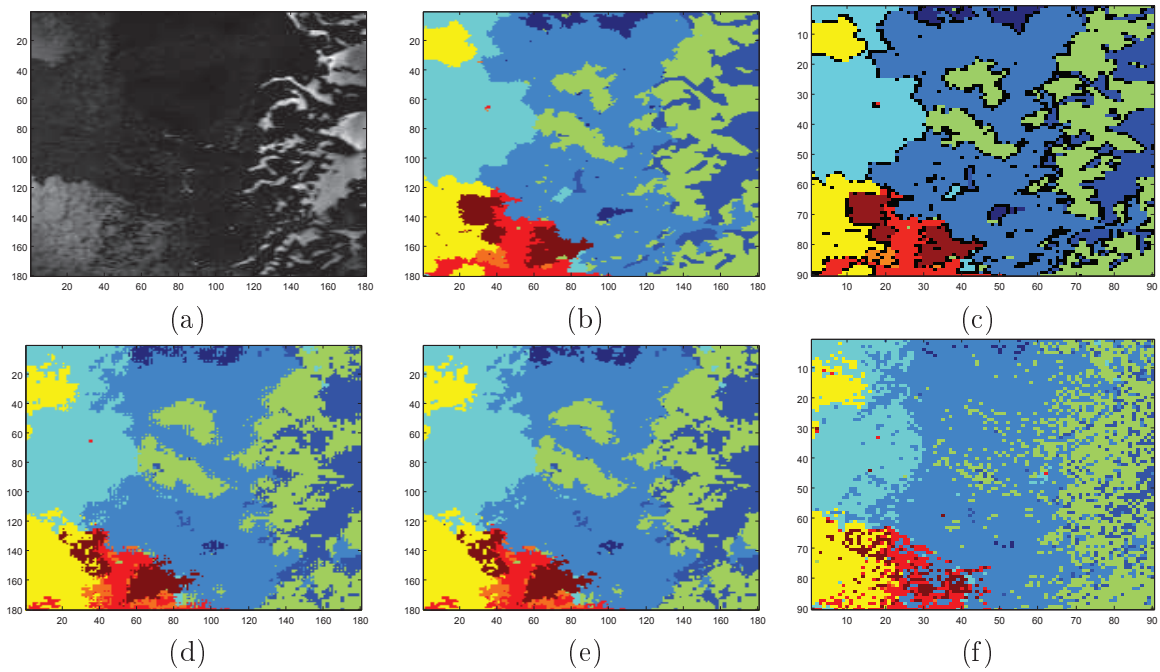


Figure 7.7: AVIRIS Hekla data set. (a) AVIRIS Hekla, band 80 (b) Classification map obtained in [46]; (c) Low resolution ground truth. In black are represented the mixed pixels (d) Classification map obtained with the proposed approach after spectral unmixing (training set 1, treshold 0.7) (e) Final classification map obtained with the proposed approach (f) Classification map obtained with SVM

Table 7.4: Comparison of classification accuracy obtained with the proposed method and SVM in the three analysed data sets. "Mixed Pixels" refers to the percentage of correctly classified mixed pixels (i.e., pixels considered as pure in the reference image but mixed in the low resolution one). "Spatial Error" refers to the percentage of sub-pixels correctly retrieved after spectral unmixing but uncorrectly located after spatial regularization.

	AVIRIS Indian Pine subset		AVIRIS Indian Pine complete		AVIRIS Hekla	
Approach	SVM	Proposed Method	SVM	Proposed Method	SVM	Proposed Method
OA	78.22 ± 0.94%	**90.65 ± 2.41%**	72.31 ± 1.64%	**91.10 ± 1.42%**	69.19 ± 2.10%	**81.71 ± 2.34%**
κ	68.14 ± 1.63%	**84.38 ± 3.76%**	67.53 ± 1.78%	**88.84 ± 1.65%**	63.96 ± 1.91%	**76.23 ± 2.36%**
AA	81.47 ± 1.59%	**91.36 ± 1.21%**	64.34 ± 1.19%	**90.73 ± 1.73%**	62.83 ± 2.71%	**74.72 ± 3.50%**
Mixed Pixels	73.85%	88.13%	50.21%	72.77%	48.10%	67.65%
Spatial Error	-	0.62%	-	1.38%	-	2.92%
Class 1	74.49%	87.21%	36.30%	88.15%	50.30%	52.98%
Class 2	99.24%	93.59%	61.83%	87.07%	89.22%	62.18%
Class 3	76.39%	93.67%	40.65%	79.21%	74.99%	85.84%
Class 4	75.74%	90.98%	26.24%	84.70%	90.27%	90.61%
Class 5	-	-	82.45%	84.55%	42.89%	82.87%
Class 6	-	-	90.63%	95.53%	78.96%	84.55%
Class 7	-	-	76.92%	99.23%	50.82%	60.83%
Class 8	-	-	93.54%	98.94%	34.37%	68.35%
Class 9	-	-	80.00%	94.00%	53.68%	84.31%
Class 10	-	-	51.51%	84.55%	-	-
Class 11	-	-	85.41%	96.56%	-	-
Class 12	-	-	36.03%	85.18%	-	-
Class 13	-	-	66.79%	99.06%	-	-
Class 14	-	-	92.92%	99.30%	-	-
Class 15	-	-	44.47%	84.32%	-	-
Class 16	-	-	63.79%	91.37%	-	-

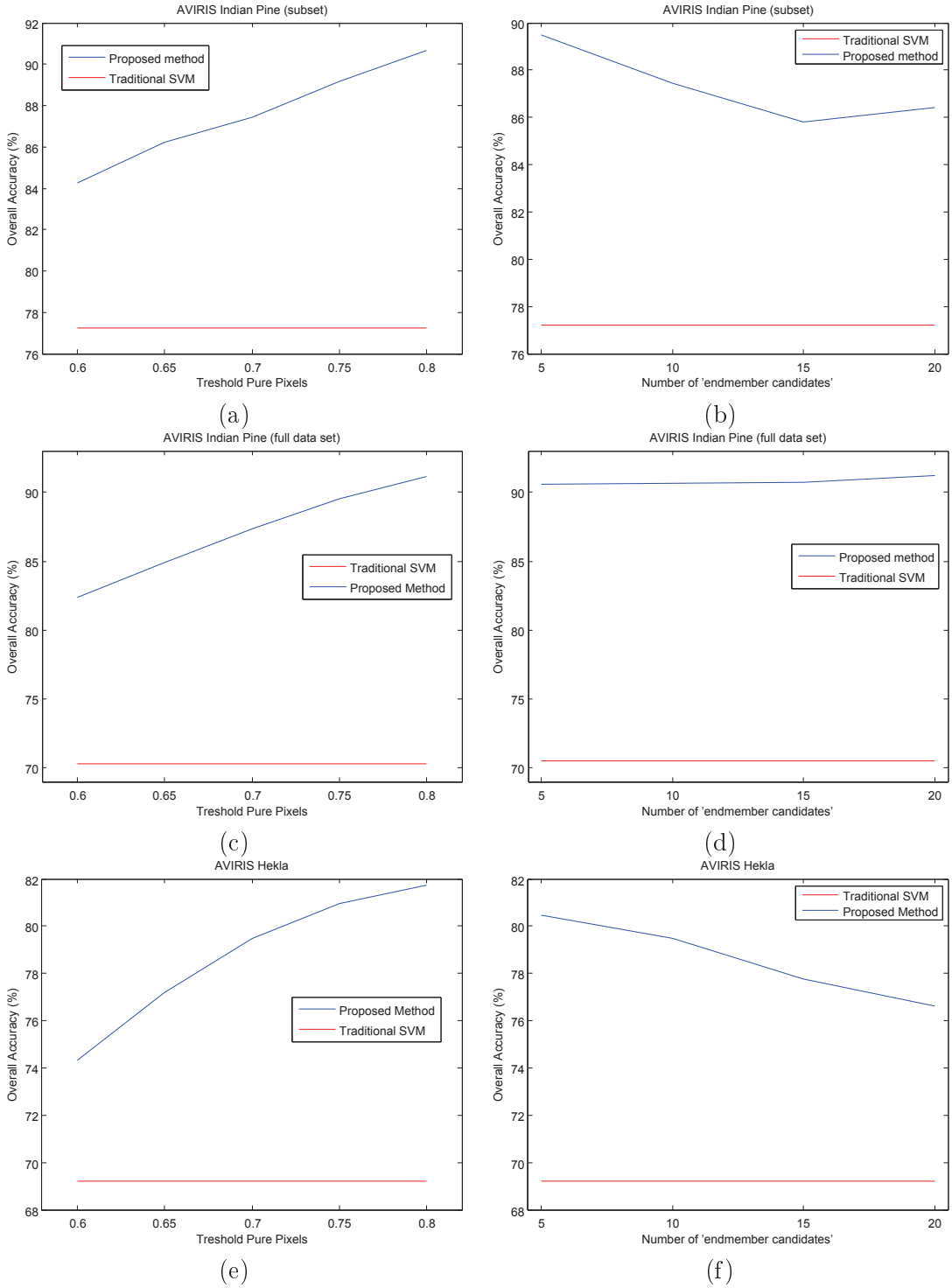


Figure 7.8: Variation of the overall classification accuracy versus the value of the parameter treshold to determine if a pixel can be considered as 'pure' for (a) AVIRIS Indian Pine subset (b) AVIRIS Indian Pine complete (c) AVIRIS Hekla data sets

Table 7.5: Computational burden for the three data sets.

	AVIRIS Indian Pine subset		AVIRIS Indian Pine complete		AVIRIS Hekla	
Approach	SVM	Proposed	SVM	Proposed	SVM	Proposed
Classification	30 s	43 s	14 min	15 min	35 min	37 min
Spectral Unmixing	-	15 s	-	1.5 min	-	2 min
Simulated Annealing	-	50 s	-	3 min	-	4 min
Total elapsed time	30 s	88 s	18 min	23 min	35 min	43 min

and not with a selected land cover ground truth map, this classification map seems to be a reliable source of knowledge about the land class coverage of the area. The original image, the classification map obtained in [46] and the obtained classification maps are shown in Fig. 7.7. 15% of the labeled pixels of the low resolution data were randomly selected from each class and used to train the classifier, and the experiment repeated ten times with different training sets.
The quantitative results in this experiment confirm those obtained in the previous ones: the proposed method provides not only a better classification map from a qualitative point of view, but also a large improvement of the overall accuracy of correctly classified pixels. Due to the irregular spatial structures in which the land cover classes are grouped, the spatial regularization method proposed in this chapter was expected to be less effective than in the previous cases. The quantitative results agree with this supposition, also if the overall accuracy is penalized by only 2 percentage points. More advanced techniques could be investigated in our future works, in order to have an improvement of the classification accuracy of the mixed pixels.

7.5.4 Discussion about the choice of parameters and computational burden

In the proposed method, the parameters having an influence on the overall classification accuracy obtained (apart from the parameters of the SVM, which are automatically selected through cross-validation) are the threshold to distinguish between pure and mixed pixels and the number of 'endmember candidates' to consider for the spectral unmixing, in the second step. How the classification accuracy changes by changing the value of the parameters can be seen in Fig. 7.8. It can be noticed that the proposed method outperforms the traditional SVM in terms of accuracy over the whole range tested, being the choice of the parameters not crucial for the classification. As could be expected, a high value of the threshold to determine if a pixel can be considered as 'pure' provides a higher accuracy, since only the most reliable pixels are labeled for the preliminary classification. By setting a low value of the threshold parameter, the preliminary classification map will tend to be like the hard classification map obtained with a traditional SVM, thus decreasing the interest of the proposed method.
The number of 'endmember candidates' considered in the spectral unmixing step points out the importance of the spatial information. The best results are in general obtained by considering a low number of candidates (which are the spatially closest to the considered pixel). When setting a larger value of the parameter, endmember candidates spatially far from the analyzed pixel can be selected, introducing useless information and thus leading to a slight decrease in the classification accuracy.
The computational burden of the proposed method can be seen in Table 7.5. The training of the SVM, which quadratically depends on the size of the training set, is the most computationally expensive step of the proposed approach. The spectral unmixing step depends on the number of the mixed pixel to unmix, while the Simulated Annealing regularization

depends on the number of mixed pixels and on the zoom factor desired. When requiring a larger zoom, the number of possible sub-pixel combinations grows exponentially, thus requiring a heavier computational burden to reach the optimal configuration. In case of desired zoom factor equal or higher than 4, the computational burden of the spatial regularization is expected the be the most important.

7.6 Conclusions

Classification of hyperspectral images in presence of mixed pixels was addressed in this chapter. A new method for the improvement of the spatial resolution of the classification maps was proposed. The method exploits the advantages of both soft classification techniques and spectral unmixing algorithms, in order to determine the fractional abundances of the classes at a sub-pixel scale. After the fractional abundances have been determined, spatial regularization by Simulated Annealing is finally performed to spatially locate the land cover classes within each pixel. Experiments were carried out on three different data sets and show that the proposed method clearly outperforms classical classification techniques when areas with mixtures of materials are located in the scene, providing excellent results both from a visually and quantitative point of view. Further research will be devoted to the investigation of advanced methods to better discriminate pure and mixed pixels, and of the possibility of alternative techniques of spatial regularization.

Conclusions

The supervised classification of hyperspectral remote sensing data was addressed in this thesis. The aim was to propose innovative technologies able to deal with the main issues related to this kind of data, and to incorporate different sources of information. Two main strategies were analyzed: the proposition of advanced techniques exploiting the spectral and spatial information and the developments of approaches exploiting sub-pixel information estimated from the data. For the first strategy, Independent Component Analysis was proposed as a valid alternative for dimensionality reduction. Starting from this point, two important approaches were proposed:

1. A statistical classifier exploiting the spectral information of hyperspectral data. Independent Component Discriminant Analysis exploits the ability of ICA of projecting the data into an independent space. By minimizing all the statistical dependencies up to the fourth order, the proposed approach ICDA gives the possibility to estimate the multivariate densities as the product of univariate densities. The method has proven to be very effective both in terms of classification accuracy and computational burden when compared to the Support Vector Machines.
   The tests conducted with small amount of training samples showed the comparative effectiveness of the method also in such a case, when the training set is well representative of the problem at hand. The method performed reasonably well also in case of disjoint training and test sets, but in such a case the advantages provided over SVM are less evident.

2. The second approach incorporated the contextual information provided by Morphological Attribute Filters in the classification framework. Based on the good results obtained by the application of Morphological Profiles to remote sensing problems, several attribute filters were tested, extracting information about the area, the diagonal of the bounding box, the moment of inertia and the standard deviation of the objects. The experiments have shown the importance of using the complementarity of the information. While single object could be well represented by their peculiar features, a complex data scene needs differentiated information in order to be exhaustively represented. Indeed, the best results were always obtained by considering the multi-filters approaches.
   It has to be noticed that when all the filters were considered, the problems related to the curse of dimensionality did not appear in spite of the unfavorable ration between the number of training samples and the number of dimensions considered.

The second strategy tested in this work was the incorporation of sub-pixel information within the classification framework. As a general conclusion, the joint use of full pixel and mixed pixel techniques is desirable to improve the classification accuracy of hyperspectral images, both from a quantitative and a qualitative point of view. More specifically, the approaches developed in this work pointed out two main results:

1. Sub-pixel information can be usefully incorporate into a dimensionality reduction method in order to mitigate problems related to the high dimensionality of the data but also to the presence of mixed pixels within an image. Incorporating information about the (possibly) mixed nature of the training samples during the learning stage resulted highly beneficial for the classification accuracy, with the potential advantage of improved interpretability of features due the physical nature of the extracted abundance maps.
   The most suitable number of features to be retained is still an open question. The methods proposed in the literature for estimating the dimensionality of a data set often provide discordant results, and these results are not necessarily related to the most suitable number of features to be retained for classification. Although final classification

results are highly dependent on the number of features retained, the chains tested in this work suggest higher accuracies with respect to traditional methods, such as PCA and ICA, and comparable or higher accuracies related to MNF.

2. The joint use of hard classification methods and spectral unmixing techniques can be useful to obtain thematic maps at higher resolution than the original input data. This possibility offers a new perspective for the analysis of data with moderate-low spatial resolution. The proposed approach is based on the creation of abundance maps through spectral unmixing, and the transformation of these maps into a classification map where each original pixel is split into a number of sub-pixels labelled according the unmixing results. Finally, a spatial regularization step is performed to correctly locate these sub-pixels from a spatial point of view. The possibility to incorporate this approach into a supervised classification framework was also discussed, showing very promising results for the analysis of data sets containing mixed pixels.

There are many possible perspectives for pursuing this work. First of all, related to spectral-spatial information of hyperspectral. With the improvement of computer technologies and computing power, the use of kernel versions of traditional feature reduction techniques (for example, Kernel ICA), could be analyzed. These approaches are very promising for extracting a high amount of useful information from hyperspectral data, rather than concentrating all the information into few first components. However, they require a big computational burden, and their use can severely affect the computational efficiency of the classification approaches. However, the development of powerful machines and the use of parallel computing framework can be very important to mitigate these shortcomings.

New techniques for integrating hard classification methods with mixed pixel techniques could be developed as an advancement of this thesis. Experiments carried out in this work showed the importance of considering sub-pixel information for obtaining accurate results. However, the research is far from being exhausted. Here, the sub-pixel spatial information was only considered as the land cover label of the neighboring pixels. In future works, it could be interesting to consider the characteristics of the objects to which a pixel belongs, in order to detect important information such as the distance from the border (pixels close to the border of an object will be more likely to be mixed).

As a final conclusion of this thesis, the importance of integrating different sources of information should be highlighted. In particular, considering the characteristic of new generation hyperspectral sensors which will be launched during the next years (such as EnMap and PRISMA), the incorporation of sub-pixel information opens new perspective for image analysis. Here, a pioneering attempt of developing techniques based on this rational was investigated. In spite of the fact that results leave a lot of space for improvements, the very interesting results obtained suggest that many possibilities exist for continuing on this direction.

Appendix A

Data sets

This chapter is dedicated to the description of the data used in this thesis. When a data set is used for some experiment, a short introduction is provided in the manuscript. However, in order to avoid repetitions, a complete description of all the data sets considered in this work is only presented here. The spectral coverage and spatial resolution of each image, along with the ground truth associated with the data, are detailed in the following.

A.1 ROSIS data sets

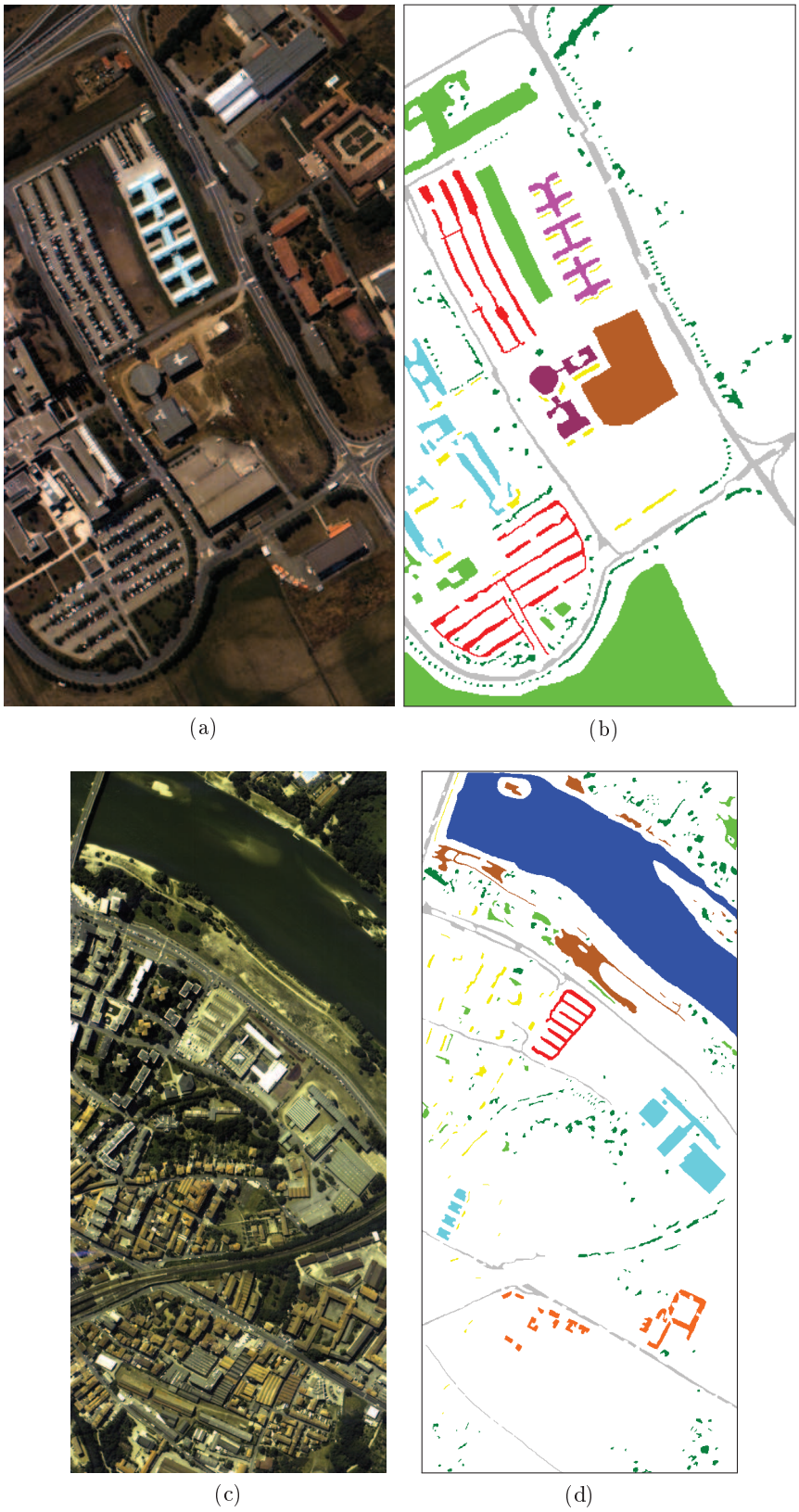
Airborne data from the ROSIS-03 (Reflective Optics System Imaging Spectrometer) optical sensor are used for the experiments. The flight over the city of Pavia, Italy, was operated by the Deutschen Zentrum fur Luft- und Raumfahrt (DLR, the German Aerospace Agency) in the framework of the HySens project, managed and sponsored by the European Union. According to specifications the number of bands of the ROSIS-03 sensor is 115 with a spectral coverage ranging from 0.43 to 0.86 μm. The spatial resolution is 1.3 m per pixel. Two data set were available: the University area and the Pavia Center.

A.1.1 University Area

The original data set is 610 by 340 pixels. Some channels (12) have been removed due to noise. The remaining 103 spectral dimensions are processed. Nine classes of interest are considered, namely: trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow and soil. The image was acquired around the Engineering School at the University of Pavia. False color image is presented in Figure A.1.(a) and the available testing set in Figure A.1 (b). Testing and training set are detailed in Table A.1.

A.1.2 Center Area

The second ROSIS data set is the center of Pavia. The Pavia center image was originally 1096 by 1096 pixels. A 381 pixel wide black in the left part of image was removed, resulting in a "two part" image. This "two part" image is 1096 by 715 pixels. Some channels (13) have been removed due to noise. The remaining 102 spectral dimensions are processed. Nine classes of interest are considered, namely: water, tree, meadow, brick, soil, asphalt, bitumen, tile and shadow. False color image is presented in Figure A.1 (c) and the available testing set in Figure A.1 (d). Testing and training set are detailed in Table A.1.



(a) (b)

(c) (d)

Figure A.1: ROSIS data sets: (a) University data set, false color image (b) University test set (c) Center data set, false color image (d) Center test set

	University			Center		
	Class	Samples		Class	Samples	
No.	Name	Train	Test	Name	Train	Test
1	Asphalt	548	6641	Water	824	65971
2	Meadow	540	18649	Tree	820	7598
3	Gravel	392	2099	Meadow	824	3090
4	Tree	524	3064	Brick	808	2685
5	Metal Sheet	265	1345	Bare Soil	820	6584
6	Bare Soil	532	5029	Asphalt	816	9248
7	Bitumen	375	1330	Bitumen	808	7287
8	Brick	514	3682	Tile	1260	42826
9	Shadow	231	947	Shadow	476	2863
	Total	3921	42776	Total	7456	148152

Table A.1: Information about training and test samples of the ROSIS (University and Center) data sets

A.2 AVIRIS data sets

Four hyperspectral images provided by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) were used in this thesis. The AVIRIS sensor operates in the visible, near- and mid- infrared portions of the electromagnetic spectrum, its sensitivity range spanning wavelengths from 0.4 μm to 2.4 μm. The sensor system has 224 data channels, utilizing four spectrometers, whereas each spectral band is approximately 10nm in width. The spatial resolution is 20 m by pixel.

A.2.1 Indian Pine

The first AVIRIS image was taken over NW Indiana's Indian Pine test site in June 1992. This image has been widely used in the remote sensing community for both classification and spectral unmixing purposes, and thus represents an interesting benchmark for the proposed method. The original image is composed by 145 $\times$ 145 pixels. It contains sixteen classes of interest. Some works presented in the literature, as [179], have considered as data set only a small part of the image , consisting of pixels [31-116] $\times$ [27-94] for a size of 86 $\times$ 68, which contains four labeled classes (the background pixels were not considered for classification purposes). We will refer to this data set as the "Subset scene". The calibrated data are available online [1] with detailed ground-truth information.

A.2.2 Hekla

The second AVIRIS image represents 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 10 years, in 1970, 1980-81, 1991 and in 2000. The volcano is located on

[1] http://dynamo.ecn.purdue.edu/~biehl/

the South-Western 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. During the image acquisition, spectrometer 4 was not working properly. This particular spectrometer operates in the wavelength range from 1.84 μm to 2.4 μm (64 bands). These 64 bands were deleted from the imagery along with the first channels for all the other spectrometers, and the remaining 157 data channels were left. The original data sets contains 12 land cover classes. In this thesis, both the original image, and a subset of 180 $\times$ 180 pixels have been used.

A.2.3 Kennedy Space Center

The third AVIRIS data set was collected over the Kennedy Space Center, Florida, in March 1996. The portion of this scene used in our experiments has dimensions of 292 $\times$ 383 pixels. After removing water absorption and low-SNR bands, 176 bands were used for the analysis. Twelve ground-truth classes were available, where the number of pixels in the smallest class is 134 while the number of pixels in the largest class is 761.

A.2.4 Salinas

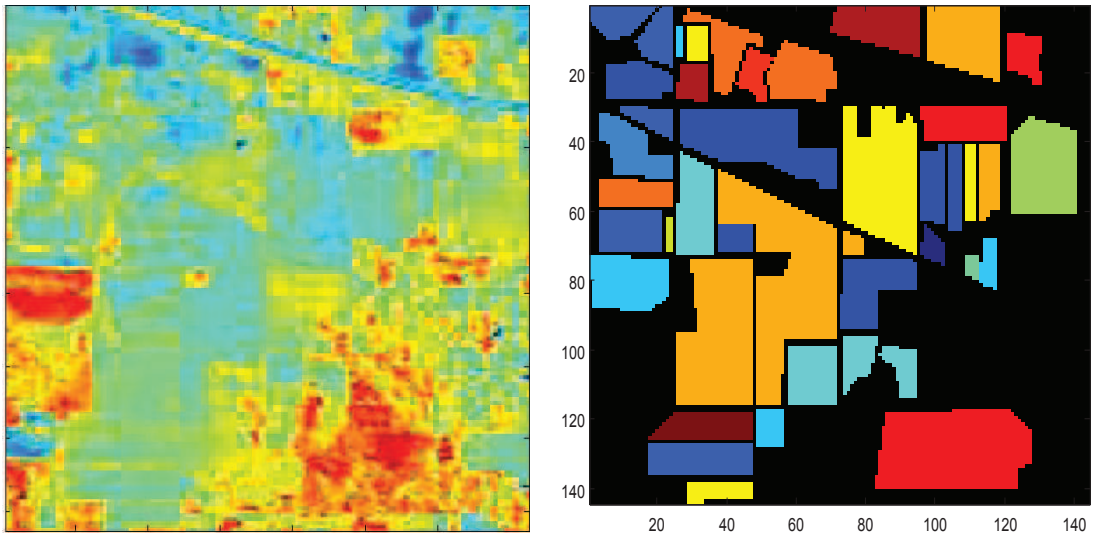
The fourth AVIRIS data set used in experiments was collected over the Valley of Salinas, Southern California, in 1998. It contains 217 $\times$ 512 pixels and 224 spectral bands from 0.4 to 2.5 μm, with nominal spectral resolution of 10 nm. It was taken at low altitude with a pixel size of 3.7 m. The data include vegetables, bare soils, and vineyard fields. The upper leftmost part of Fig. A.4 shows the entire scene (with overlaid ground-truth areas). The upper rightmost part of Fig. A.4 shows the available ground-truth regions for the scene, and the bottom part of Fig. A.4 shows some photographs taken in the field for the different agricultural fields at the time of data collection.

A.3 HYDICE data set

Airborne data from the HYDICE sensor (Hyperspectral Digital Imagery Collection Experiment) was used for the experiments. The HYDICE was used to collect data from flightline over the Washington DC Mall. Hyperspectral HYDICE data originally contained two hundred and ten 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 class were defined, namely: roof, road, grass, tree, trail, water and shadow. False color images is presented in FigureA.5(a) and the available testing set in Figure A.5 (b). Testing and training set are detailed in Table A.2.

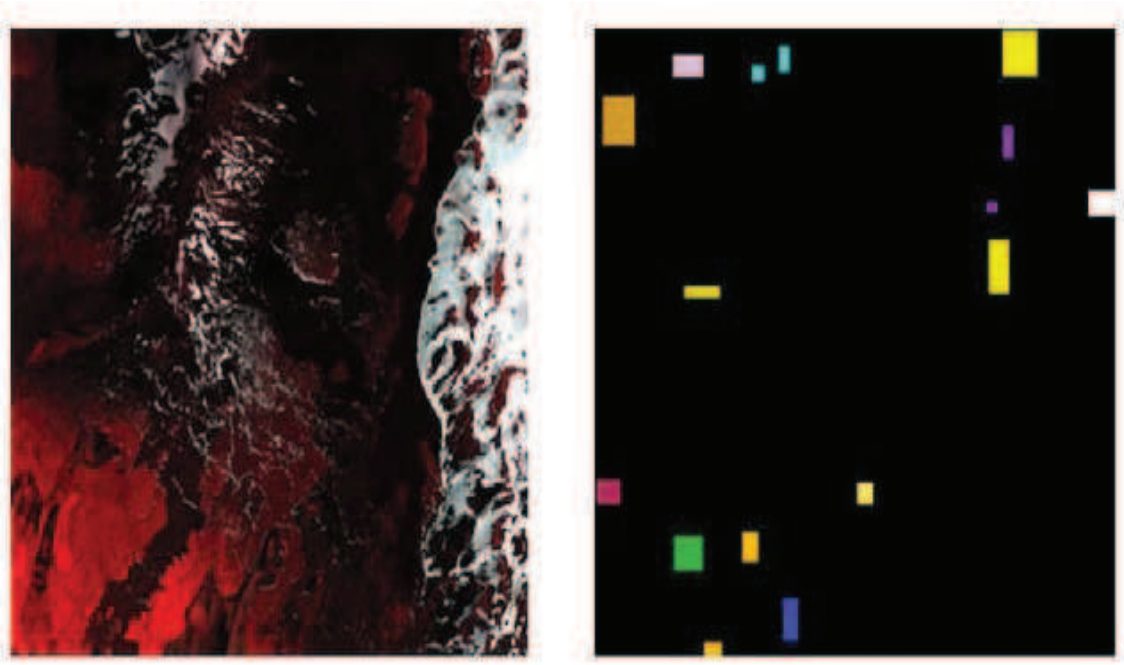
A.4 AISA data set

The last image analyzed in our experiments is an AISA Eagle dataset. It contains 252 bands ranging from 395 to 975 nm in the visible and NIR spectral range. The original spatial resolution of the image was 2 m measured on ground, but in order to be treatable and still useful for the purposes of land cover interpretation it was downscaled to 6 m ground resolution while keeping the original spectral information as possible. The area is located in Hungary and contains arable lands near to the city of Heves. The area is mainly useful because of agricultural production. In our experiments, we considered a large subset of the image (400$\times$500 pixels) containing six classes of interest.



(a) (b)

Figure A.2: AVIRIS Indian Pines data set: (a) False color image; (b) Ground truth data.



(a) (b)

Figure A.3: AVIRIS Hekla data set: (a) False color image; (b) Ground truth data.

Table A.2: AVIRIS and HYDICE data sets ground truth information

	AVIRIS Indian Pine		AVIRIS Hekla		HYDICE	
No.	Name	Samples	Name	Test	Name	Test
1	Alfalfa	54	Andesite lava 1970	1023	Roof	3794
2	Corn-no till	1434	And. lava 1980 I	684	Road	376
3	Corn-min till	834	And. lava 1980 II	458	Trail	135
4	Corn	234	And. lava 1991 I	550	Grass	1888
5	Grass-Pasture	497	And. lava 1991 II	1496	Tree	365
6	Grass-Trees	747	And. lava moss cover	700	Water	1184
7	Grass-Mowed	26	Hyaloclastite formation	342	Shadow	57
8	Hay-windrowed	489	Lava tephra covered	404		
9	Oats	20	Rhyolite	708		
10	Soybean-no till	968	Scoria	713		
11	Soybean-min till	2468	Firn-glacier ice	410		
12	Soybean-clean t	614	Snow	318		
13	Wheat	212	-	-		
14	Woods	1294	-	-		
15	Bldg-Trees-Drive	380	-			
16	Stone-Steel Tower	95	-	-		

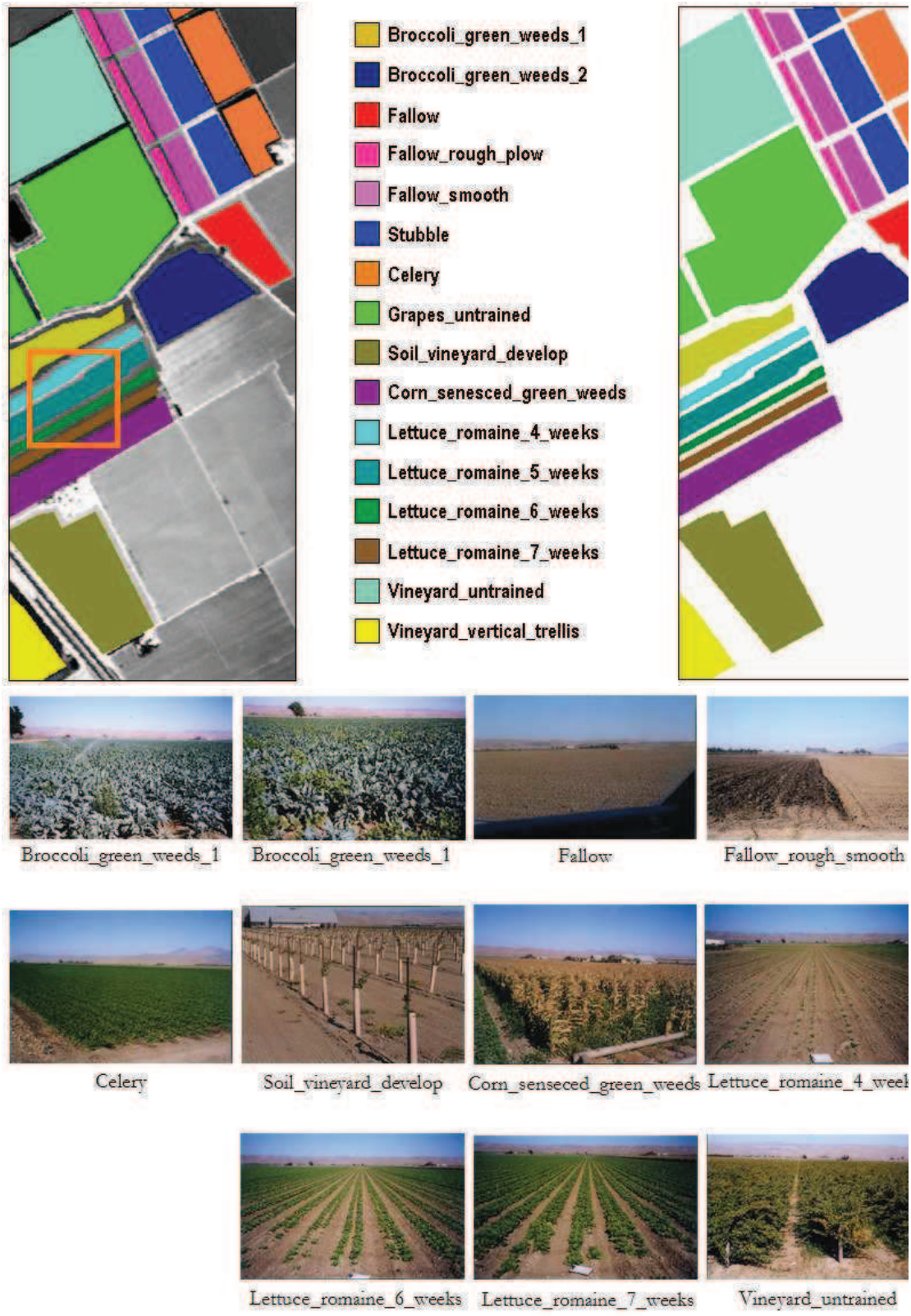
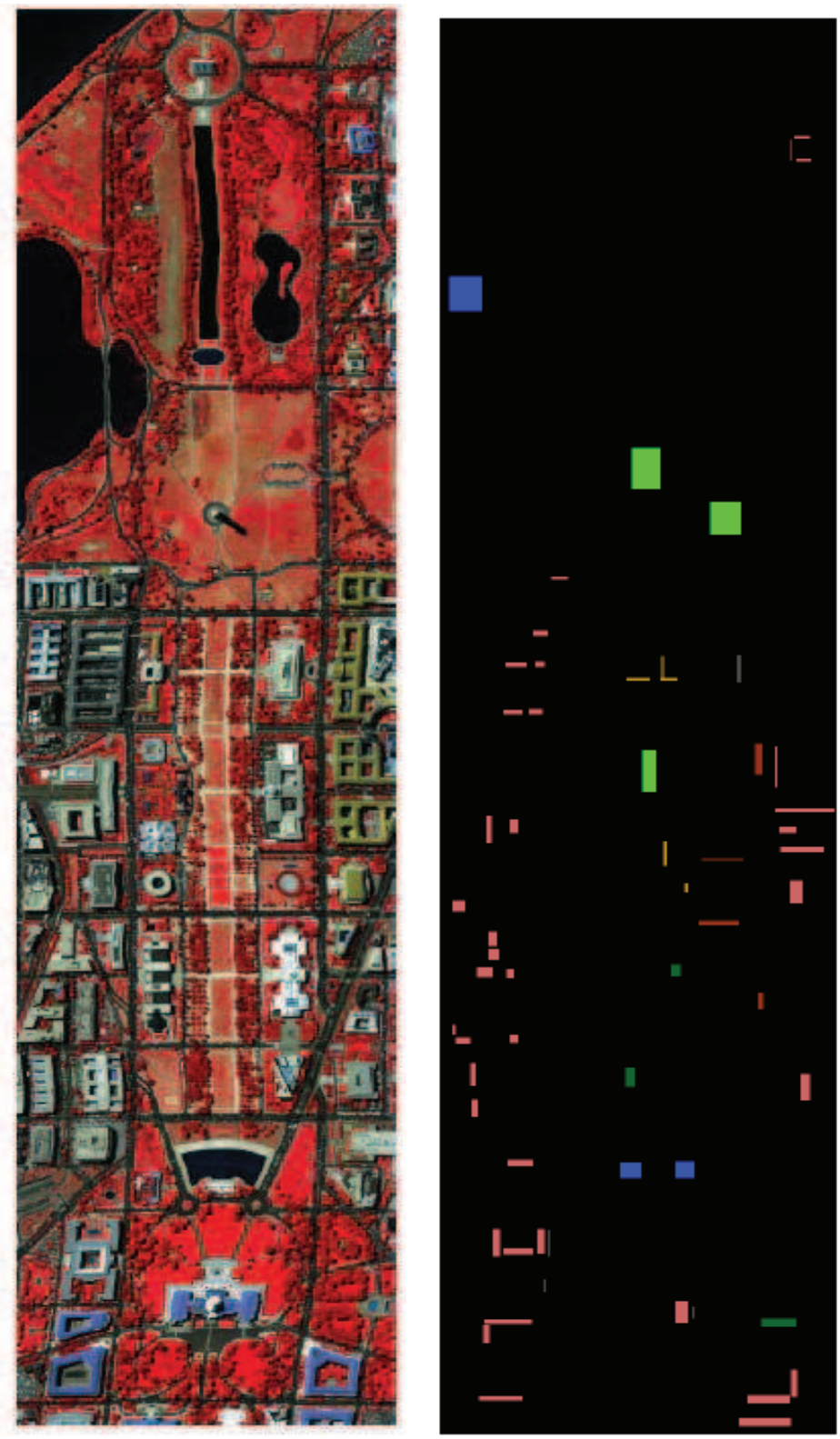


Figure A.4: AVIRIS Salinas data set, ground truth and pictures of the different classes.



(a) (b)

Figure A.5: HYDICE Washington DC Mall data set: (a) Original data, false color image; (b) Ground truth.

Appendix B

Accuracy assesment

The estimation of the classification accuracy is based on the confusion matrix. From that matrix it is possible to evaluate the exactitude of a given classification map by comparison to the reference map. Several estimates, from global estimation to specific estimation are extracted from the confusion matrix. They are detailed in the following as well as the confusion matrix.

Definition 1 *In the field of artificial intelligence, a confusion matrix is a visualization tool typically used in supervised learning . Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see where the system is confusing (i.e., commonly mis-labelling one class as another).*

An example of confusion matrix is given Table B.1 for a 3-classes problem. C_i represents the class i and C_{ij} is the number of pixels assign to the class j by the classifier which are referenced as class i.

Definition 2 *The overall accuracy (OA) is the percentage of correctly classified pixels:*

$$OA = \frac{\sum_{i}^{N_c} C_{ii}}{\sum_{ij}^{N_c} C_{ij}} \quad x \quad 100 \tag{B.1}$$

Definition 3 *The Class Accuracy (or producer's accuracy) (CA) is a measure how well a certain area has been classified. It includes the error of omission which refers to the proportion of observed features on the ground that are not classified in the map. The more errors of omission exist, the lower the producer accuracy. It is derived by the number of correct pixels in one class divided by the total number of pixels as derived from reference data:*

$$CA_i = \frac{C_{ii}}{\sum_{j}^{N_c} C_{ij}} \quad x \quad 100 \tag{B.2}$$

Definition 4 *The average accuracy (AA) is the mean of class accuracy for all the classes*

$$AA = \frac{C_{ii}}{\sum_{j}^{N_c} CA_i} \quad \text{x} \quad 100 \tag{B.3}$$

An OA or an AA is closed to 100% (0%) means that the classification accuracy is almost perfect (wrong). When a referenced set is unbalanced, the OA may not be representative of the true performance of the classifier. For instance, if a class has very few number of

Table B.1: Confusion Matrix, N is the number of referenced pixel and N_c is the number of classes.

Percentage	Classification data				
Reference data	C_1	C_2	C_3	Row total	Producers'accuracy
C_1	C_{11}	C_{12}	C_{13}	$\sum_i^{N_c} C_{1i}$	$\frac{C_{11}}{\sum_i^{N_c} C_{1i}}$
C_2	C_{21}	C_{22}	C_{23}	$\sum_i^{N_c} C_{2i}$	$\frac{C_{22}}{\sum_i^{N_c} C_{2i}}$
C_3	C_{31}	C_{32}	C_{33}	$\sum_i^{N_c} C_{3i}$	$\frac{C_{33}}{\sum_i^{N_c} C_{3i}}$
Column total	$\sum_i^{N_c} C_{i1}$	$\sum_i^{N_c} C_{i2}$	$\sum_i^{N_c} C_{i3}$	N	
User's accuracy	$\frac{C_{11}}{\sum_i^{N_c} C_{i1}}$	$\frac{C_{11}}{\sum_i^{N_c} C_{2i}}$	$\frac{C_{11}}{\sum_i^{N_c} C_{3i}}$		

referenced pixels, its influence will be very low in the computation of the OA, while it will be more influent in the AA since the mean is done the number of classes rather than the whole number of pixels. Strong difference between OA and AA may indicate that a specific class is wrongly classified with a high proportion.

Definition 5 *The Kappa Coefficient (k) is a statistical measure of agreement. It is the percentage agreement corrected by the level of agreement that could be expected due to chance alone. It is generally thought to be a more robust measure than simple percent agreement calculation since k takes into account the agreement occurring by chance.*

$$k = \frac{P_o - P_e}{1 - P_o} \tag{B.4}$$

where

$$\begin{aligned} P_o &= OA \\ P_e &= \frac{1}{N^2} \sum_i^{N_c} C_{i\cdot} C_{\cdot i} \\ C_{i\cdot} &= \sum_j^{N_c} C_{ij} \\ C_{\cdot i} &= \sum_j^{N_c} C_{ji} \end{aligned} \tag{B.5}$$

Bibliography

[1] A. F. H. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *Science*, vol. 228, pp. 1147–1153, 1985.

[2] D. A. Landgrebe, *Signal theory methods in multispectral remote sensing.* Hoboken, NJ: John Wiley and Sons, 2003.

[3] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis: An Introduction.* Springer, 2006.

[4] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, pp. 55–63, January 1968.

[5] A. Cracknell, "Synergy in remote sensing: What's in a pixel?" *Int. J. Remote Sens.*, vol. 19, pp. 2025–2047, 1998.

[6] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287 – 314, 1994.

[7] C. Lee and D. A. Landgrebe, "Decision boundary feature extraction for neural networks," *IEEE Trans. Neural Networks*, vol. 8, pp. 75–83, 1997.

[8] ——, "Feature extraction based on decision boundaries," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 388–400, Apr 1993.

[9] I. Bell and G. Baranoski, "Reducing the dimensionality of plant spectral databases," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 3, pp. 570 – 576, march 2004.

[10] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, pp. 480–491, 2005.

[11] X. Jia and J. Richards, "Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 37, no. 1, pp. 538 –542, jan 1999.

[12] V. Tsagaris, V. Anastassopoulos, and G. Lampropoulos, "Fusion of hyperspectral data using segmented pct for color representation and classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 10, pp. 2365 – 2375, oct. 2005.

[13] S. Z. Li, *Markov Random Field Modeling in Image Analysis second edition.* Springer, 2001.

[14] J. A. Richards, "Analysis of remotely sensed data: the formative decades and the future," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, pp. 422–432, 2005.

[15] M. Fauvel, J. Chanussot, J. Benediktsson, and J. 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.

[16] B. Luo and J. Chanussot, "Hyperspectral image classification based on spectral and geometrical features," *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2009.

[17] ——, "Unsupervised classification of hyperspectral images by using linear unmixing algorithm," *Proceedings of the IEEE International Conference on Image Processing*, pp. 2877–2880, 2009.

[18] D. Landgrebe, "Hyperspectral image data analysis," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 17 –28, jan 2002.

[19] S. Kumar, J. Ghosh, and M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 7, pp. 1368 –1379, jul 2001.

[20] L. Bruzzone and L. Carlin, "A multilevel context-based system for classification of very high spatial resolution images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 9, pp. 2587 –2600, sept. 2006.

[21] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *Int. J. of Rem. Sens.*, vol. 31, pp. 5975–5991, 2010.

[22] M. G. Kendall, *A course in the geometry of n-dimensions*, N. York, Ed. Dover Publication, 1961.

[23] K. D., D. Hatzinakos, and H. Leung, "Robust classification of blurred imagery," *IEEE Trans. Image Proc.*, vol. 9, no. 2, pp. 243–255, Feb. 2000.

[24] S. Cai, Q. Du, and R. Moorhead, "Hyperspectral imagery visualization using double layers," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 10, pp. 3028 –3036, oct. 2007.

[25] N. Kosaka, K. Uto, and Y. Kosugi, "Ica-aided mixed-pixel analysis of hyperspectral data in agricultural land," *Geoscience and Remote Sensing Letters, IEEE*, vol. 2, no. 2, pp. 220 – 224, april 2005.

[26] J. Wang and C.-I. Chang, "Applications of independent component analysis in endmember extraction and abundance quantification for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 9, pp. 2601–2616, 2006.

[27] J. Wang and C. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Trans. Geosc. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, Jun 2006.

[28] J. A. Palmason, J. A. Benediktsson, J. R. Sveinsson, and J. Chanussot, "Classification of hyperspectral data from urban areas using morphological preprocessing and independent component analysis," in *Proc. IEEE IGARSS '05*, vol. 1, Jul. 25–29, 2005, pp. 176–179.

[29] K. Fukunaga, *Introduction to statistical pattern recognition*, S. Diego, Ed. CA: Academic Press, 1990.

[30] L. Jimenez and D. A. Landgrebe, "Supervised classification in high dimensional space: geometrical, statistical and asymptotical properties of multivariate data," *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 28, pp. 39–54, February 1993.

[31] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* New York: Wiley, 2001.

[32] J. Nascimento and J. B. Dias, "Does independent component analysis play a role in unmixing hyperspectral data?" *IEEE Trans. on Geosci. Remote Sens.*, vol. 43, no. 1, pp. 175–187, Jan. 2005.

[33] P. Birjandi and M. Dactu, "Multiscale and dimensionality behavior of ica components for satellite image indexing," *Geosc. Remote Sens. Letters*, vol. 7, no. 1, pp. 103–107, Jan 2010.

[34] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* Wiley, New York, 2001.

[35] T. Lee, M. Girolami, A. Bell, and T. Sejnowski, "An unifying information-theoretic framework for independent component analysis," *Comp. Math. with Appl.*, vol. 31, no. 11, pp. 1–21, Mar 2000.

[36] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation. Independent Component Analysis and Applications.* Academic Press, 2010.

[37] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neur. Netw*, vol. 10, no. 3, pp. 626–634, 1999.

[38] A. Bell and T. Sejnowski, *Adv. in Neur. Inf. Process. Sys. 7.* Cambridge Press, MIT, 1995, vol. The MIT Press, Cambridge, ch. A non linear information maximization algorithm that performs blind separation, pp. 467–474.

[39] J. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals," *IEEE Proceedings*, vol. 140, no. 6, pp. 363–370, 1993.

[40] S. Moussaoui, H. Hauksdottir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Doute, and J. Benediktsson, "On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation," *Neurocomputing*, vol. 71, pp. 2194–2208, 2008.

[41] A. Villa, J. Chanussot, C. Jutten, J. Benediktsson, and S. Moussaoui, "On the use of ica for hyperspectral image analysis," in *IEEE International Geoscience and Remote Sensing Symposium*, 2009.

[42] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 3rd edition, Ed. Mc-Graw Hill, 1991.

[43] M. Girolami, *Self-organising Neural Networks - Independent Component Analysis and Blind Source Separation.* Springer Verlag, 1999.

[44] C. Nikias and A. Petropulu, *Higher-order Spectral Analysis - A Nonlinear Signal Processing Framework.* Prentice Hall, 1993.

[45] M. Rosenblatt, *Stationary sequences and Random Fields.* Birkhauser, 1985.

[46] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson, "Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers," *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 5, pp. 1267–1279, 2010.

[47] M. Dalla Mura, J. A. Benediktsson, F. Bovolo, and L. Bruzzone, "An unsupervised technique based on morphological filters for change detection in very high resolution images," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 433–437, Jul. 2008.

[48] J. Serra, *Image Analysis and Mathematical Morphology*. U.K. Academic, 1982.

[49] ——, *Image Analysis and Mathematical Morphology, Volume 2: Theoretical Advances*. U.K. Academic, 1988.

[50] P. Soille, *Morphological image analysis: principles and applications*. Berlin: Springer-Verlag, 2003.

[51] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. on Geosci. and Rem. Sens.*, vol. 48, pp. 3747–3762, 2010.

[52] E. J. Breen and R. Jones, "Attribute openings, thinnings, and granulometries," *Comput. Vis. Image Underst.*, vol. 64, no. 3, pp. 377–389, 1996. [Online]. Available: http://portal.acm.org/citation.cfm?id=246245

[53] P. H. R. Duda and D. Stork, *Pattern Classification*, second ed, Ed. John Wiley & Sons, 2001.

[54] J. Besag, "Toward bayesian image analysis," *J. Appl. Stat.*, vol. 16, no. 3, pp. 395–407, 1989.

[55] J. Friedman, "Exploratory projection pursuit," *J. Amer. Statist. Assoc.*, vol. 82, pp. 249–266, 1987.

[56] D. Scott, *Multivariate density estimation: theory, practice and visualization*, N. York, Ed. Wiley,, 1992.

[57] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 763–767, 1996.

[58] S. Tadjudin and D. A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans. Geosci. Remote. Sensing*, vol. 37, pp. 2113–2118, Jul 1999.

[59] 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. Sensing*, vol. 39, pp. 2664–2679, Dec 2001.

[60] L. O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and feature reduction via projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, pp. 2653–2667, Nov. 1999.

[61] B. Kuo and D. Landgrebe, "A robust classification procedure based on mixture classifiers and nonparametric weighted feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, pp. 2486–2494, 2002.

[62] S. Haykin, *Neural Networks: A Comprehensive Foundation.* Englewood Cliffs, NJ: Prentice-Hall, 1999.

[63] J. Benediktsson, J. Sveinsson, and K. Arnason, "Classification and feature extraction of aviris data," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, pp. 1194–1205, Sep 1995.

[64] A. Filippi and J. 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, 2007.

[65] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, pp. 1351–1362, 2005.

[66] C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

[67] S. Mika, G. Rätsch, B. Schölkopf, A. Smola, J. Weston, and K. Müller, *Advances in Neural Information Processing Systems.* Cambridge, MA, 1999, ch. Invariant feature extraction and classification in kernel spaces.

[68] A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik, "Support vector clustering," *Mach. Learn. Res.*, vol. 2, pp. 125–137, 2001.

[69] G. Rätsch, B. Schölkopf, A. Smola, S. Mika, T. Onoda, and K.-R.Müller, *Advances in Large Margin Classifiers.* Cambridge, MA: MIT Press, 1999, ch. Robust ensemble learning, pp. 207–219.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] U. Amato, A. Antoniadis, and G. Gregoire, "Independent component discriminant analysis," *Intern. Math. Journal*, vol. 3, no. 7, pp. 735–753, 2003.

[76] B. Efrom and C. Morris, "Families of minmax estimators of the mean of a multivariate normal distribution," *Annals of Statistics*, no. 4, pp. 11–21, 1976.

[77] E. Fix and J. Hodges, "Discriminatory analysis - nonparametric discrimination: Consistency properties," *Int. Stat. Rev.*, vol. 57, pp. 238–247, 1989.

[78] A. Webb, *Statistical Pattern Recognitionsecond ed.* John Wiley & Sons, 2002.

[79] C. Chang and C. Lin, "Libsvm: A library for support vector machines," 2001, [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[80] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 46, pp. 20–37, 1960.

[81] 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.

[82] J. Kittler and J. Foglein, "Contextual classification of multispectral pixel data," *Image and Vision Computing*, vol. 2, no. 1, pp. 13 – 29, 1984. [Online]. Available: http://www.sciencedirect.com/science/article/B6V09-4998R32-3/2/42035282d318b3283e8cdbe333448d72

[83] A. Frery, A. Correia, and C. da Freitas, "Classifying multifrequency fully polarimetric imagery with multiple sources of statistical evidence and contextual information," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 10, pp. 3098 –3109, oct. 2007.

[84] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using svms and morphological profiles," *IEEE Trans. on Geosci. and Rem. Sens.*, vol. 46, no. 11, pp. 3804–3814, 2008.

[85] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral & spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. on Geosci. and Rem. Sens.*, vol. 47, no. 8, pp. 2973–2987, 2009.

[86] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, pp. 93–97, 2006.

[87] P. Zhong and R. Wang, "Learning conditional random fields for classification of hyperspectral images," *Image Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 1890 –1907, july 2010.

[88] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. on Geosci. and Rem. Sens.*, vol. 43, no. 3, pp. 480–491, 2005.

[89] M. Dalla Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using morphological attribute filters and independent component analysis," in *Proc. IEEE Whispers*, Reykjavik, Iceland, 14–16 June 2010.

[90] P. Salembier, A. Oliveras, and L. Garrido, "Antiextensive connected operators for image and sequence processing," *IEEE Trans. on Im. Proc.*, vol. 7, no. 4, pp. 555–570, 1998.

[91] M. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. on Inf. Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[92] A. Villa, M. Fauvel, J. Chanussot, P. Gamba, and J. A. Benediktsson, "Gradient optimization for multiple kernel's parameters in support vector machines classification," in *Proc. IEEE IGARSS '08*, vol. 4, Jul. 7–11, 2008, pp. IV–224–IV–227.

[93] A. Plaza, J. A. Benediktsson, J. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, J. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, pp. 110–122, 2009.

[94] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Proc. Mag.*, vol. 19, pp. 44–57, 2002.

[95] J. B. Adams, M. O. Smith, and P. E. Johnson, "Spectral mixture modeling: a new analysis of rock and soil types at the viking lander 1 site," *Journal of Geophysical Research*, vol. 91, pp. 8098–8112, 1986.

[96] K. Segl, L. Guanter, H. Kaufmann, J. Schubert, S. Kaiser, B. Sang, and S. Hofer, "Simulation of spatial sensor characteristics in the context of the enmap hyperspectral mission," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 7, pp. 3046 –3054, july 2010.

[97] C. Ekaterina and C. Veronique, "Spectral un-mixing of natural surfaces scenarios," in *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International*, july 2010, pp. 4491 –4494.

[98] G. Storvik, R. Fjortoft, and A. Solberg, "A bayesian approach to classification of multiresolution remote sensing data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 3, pp. 539 – 547, march 2005.

[99] L. Gomez-Chova, G. Camps-Valls, J. Calpe-Maravilla, L. Guanter, and J. Moreno, "Cloud-screening algorithm for envisat/meris multispectral images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 12, pp. 4105 –4118, dec. 2007.

[100] Y. Gu, Y. Zhang, and J. Zhang, "Integration of spatial-spectral information for resolution enhancement in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1347–1358, Mai 2008.

[101] J. Verrelst, J. Clevers, and M. Schaepman, "Merging the minnaert- k parameter with spectral unmixing to map forest heterogeneity with chris/proba data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 11, pp. 4014 –4022, nov. 2010.

[102] F. Garcia-Vilchez, J. Munoz-Mari, M. Zortea, I. Blanes, V. Gonzalez-Ruiz, G. Camps-Valls, A. Plaza, and J. Serra-Sagrista, "On the impact of lossy compression on hyperspectral image classification and unmixing," *Geoscience and Remote Sensing Letters, IEEE*, vol. 8, no. 2, pp. 253 –257, march 2011.

[103] D. Heinz and C.-I. Chang, "Fully constrained least squares linear mixture analysis for material quantification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, pp. 529–545, 2001.

[104] D. A. Roberts, G. T. Batista, J. L. G. Pereira, E. K. Waller, and B. W. Nelson, *Remote Sensing Change Detection: Environmental Monitoring Applications and Methods.* Ann Arbor, MI: Ann Arbor Press, 1998, ch. Change identification using multitemporal spectral mixture analysis: Applications in eastern Amazonia, pp. 137–161.

[105] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, pp. 542–552, 1994.

[106] C.-I. Chang, S.-S. Chiang, J. Smith, and I. Ginsberg, "Linear spectral random mixture analysis for hyperspectral imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 40, no. 2, pp. 375 –392, feb 2002.

[107] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, pp. 650–663, 2004.

[108] A. Plaza, G. Martin, J. Plaza, M. Zortea, and S. Sanchez, *in Optical Remote Sensing - Advances in Signal Processing and Exploitation Techniques.* Springer, 2010, ch. "Recent Developments in Spectral Unmixing and Endmember Extraction".

[109] D. Heinz and C. Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 3, pp. 529–545, Mar. 2001.

[110] K. H. Haskell and R. J. Hanson, "An algorithm for linear least squares problems with equality and nononegativity constraints generalized," *Math. Prog.*, vol. 21, pp. 98–118, 1981.

[111] J. J. Settle and N. A. Drake, "Linear mixing and estimation of ground cover proportions," *International Journal of Remote Sensing*, vol. 14, pp. 1159–1177, 1993.

[112] J. J. Settle and N. A. Campbell, "On the errors of two estimators of sub-pixel fractional cover when mixing is linear," *IEEE Trans Geosci. Remote Sensing*, vol. 36, pp. 163–170, Jan 1998.

[113] A. A. Nielsen, "Linear mixture models and partial unmixing in multiand hyper-spectral image data," in *Proc. 1st EARSel Workshop on Imaging Spectroscopy*, D. S. M. Schaepman and E. K. Itten, Eds., Zurich, Switzerland, 1998, pp. 165–172.

[114] W. H. Farrand and J. C. Harsanyi, "Mapping the distribution of mine tailings in the couer d'alene river valley, idaho, through the use of a constrained energy minimization technique," *Remote Sens. Environ.*, vol. 59, pp. 64–76, 1997.

[115] J. J. Settle, "On constrained energy minimization and the partial unmixing of multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, pp. 718–721, 2002.

[116] J. M. Bioucas-Dias and J. M. P. Nascimento, "Hyperspectral subspace identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 8, pp. 2435–2445, 2008.

[117] C.-I. Chang and Q. Du, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 608–619, 2004.

[118] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, pp. 650–663, 2004.

[119] J. C. Harsanyi and C.-I. Chang, "Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 4, pp. 779–785, 1994.

[120] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex Component Analysis: A Fast Algorithm to Unmix Hyperspectral Data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, 2005.

[121] M. E. Winter, "N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data," *Proc. SPIE Image Spectrometry V*, vol. 3753, pp. 266–277, 2003.

[122] M. Zortea and A. Plaza, "Spatial preprocessing for endmember extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, pp. 2679–2693, 2009.

[123] Q. Du, H. Ren, and C.-I. Chang, "A comparative study for orthogonal subspace projection and constrained energy minimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, pp. 1525–1529, 2003.

[124] G. Foody., "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogrammetric engineering and remote sensing*, vol. 70, no. 5, pp. 627–633, 2004.

[125] A. Plaza, P. Martinez, J. Plaza, and R. Perez, "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 466–479, 2005.

[126] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing.* John Wiley & Sons: New York, 2003.

[127] I. Dopido, M. Zortea, A. Villa, A. Plaza, and P. Gamba, "Unmixing prior to supervised classification of remotely sensed hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 8, pp. 760–764, Jul. 2011.

[128] M. Graham and D. Miller, "Unsupervised learning of parsimonious mixtures on large spaces with integrated feature and component selection," *Signal Processing, IEEE Transactions on*, vol. 54, no. 4, pp. 1289 – 1303, april 2006.

[129] P. Bajorski, "Statistical inference in pca for hyperspectral images," *Selected Topics in Signal Processing, IEEE Journal of*, vol. PP, no. 99, p. 1, 2011.

[130] ——, "Second moment linear dimensionality as an alternative to virtual dimensionality," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 2, pp. 672 –678, feb. 2011.

[131] L. O. Jimenez-Rodriguez, E. Arzuaga-Cruz, and M. Velez-Reyes, "Unsupervised linear feature-extraction methods and their effects in the classification of high-dimensional data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 2, pp. 469 –483, 2007.

[132] 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 *et al.*, "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sensing of Environment*, vol. 65, no. 3, pp. 227–248, 1998.

[133] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, pp. 65–74, 1988.

[134] J. W. Boardman, F. A. Kruse, and R. O. Green, "Mapping target signatures via partial unmixing of AVIRIS data," *Proc. JPL Airborne Earth Sci. Workshop*, pp. 23–26, 1995.

[135] J. C. Harsanyi and C.-I. Chang, "Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, pp. 779–785, 1994.

[136] B. Chalmond, B. Francesconi, and S. Herbin, "Using hidden scale for salient object detection," *IEEE Trans. Image Proc.*, vol. 15, no. 9, pp. 2644–2656, Sep. 2006.

[137] B. Guo, S. Gunn, R. Damper, and J. Nelson, "Customizing kernel functions for svm-based hyperspectral image classification," *IEEE Trans. Image Proc.*, vol. 17, no. 4, pp. 622–629, Apr. 2008.

[138] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing.* John Wiley and Sons, 2003.

[139] G. Vane, R. Green, T. Chrien, H. Enmark, E. Hansen, and W. Porter, "The airborne visible/infrared imaging spectrometer (aviris)," *Remote Sens. Environ.*, vol. 44, pp. 127–143, 1993.

[140] L. Miao, H. Qi, and H. Szu, "A maximum entropy approach to unsupervised mixed-pixel decomposition," *IEEE Trans. Image Proc.*, vol. 16, pp. 1008–1021, 2007.

[141] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "A new approach to mixed pixel classification of hyperspectral imagery based on extended morphological profiles," *Pattern Recognition*, vol. 37, pp. 1097–1116, 2004.

[142] M. T. Eismann and R. C. Hardie, "Hyperspectral resolution enhancement using high-resolution multispectral imagery with arbitrary response functions," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 455–465, Mar. 2005.

[143] R. Hardie, M. Eismann, and G. Wilson, "Map estimation for hyperspectral image resolution enhancement using an auxiliary sensor," *IEEE Trans. Image Proc.*, vol. 13, no. 9, pp. 1174–1184, Sept 2004.

[144] T. Akgun, Y. Altunbasak, and R. Mersereau, "Super-resolution reconstruction of hyperspectral images," *IEEE Trans. Image Proc.*, vol. 14, no. 11, pp. 1860–1875, Nov. 2005.

[145] G. Foody, "Sharpening fuzzy classification output to refine the representation of sub-pixel land cover distribution," *Int. J. Remote Sens.*, vol. 19, pp. 2593–2599, 1998.

[146] M. Nguyen, P. Atkinson, and H. Lewis, "Superresolution mapping using a hopfield neural network with fused images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 3, pp. 736–749, 2006.

[147] M. Nachtegael, D. Weken, E. van der Kerre, and W. Philips, Eds., *Soft Computing in Image Processing.* Springer, 2007.

[148] A. Robin, S. L. Hegarat-Mascle, and L. Moisan, "Unsupervised subpixelic classification using coarse-resolution time series and structural information," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1359–1374, May 2008.

[149] A. J. Tatem, H. G. Lewis, P. M. Atkinson, and M. S. Nixon, "Superresolution target identification from remotely sensed images using a hopfield neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 4, pp. 781–796, Apr 2001.

[150] A. Plaza, P. Martinez, R. Perez, , and J. Plaza, "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 650–653, Mar. 2004.

[151] C. Chang, C. Wu, and C. Tsai, "Random n-finder (n-findr) endmember extraction algorithms for hyperspectral imagery," accepted for publication 2011.

[152] A. Baraldi, E. Binaghi, P. Blonda, P. Brivio, and A. Rampini, "Comparison of the multilayer perceptron with neuro-fuzzy techniques in the estimation of cover class mixture in remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 5, pp. 994–1005, May 2001.

[153] F. Bovolo, L. Bruzzone, and L. Carlin, "A novel technique for subpixel image classification based on support vector machine," *IEEE Trans. on Image Process.*, vol. 19, no. 11, pp. 2983–2999, Nov 2010.

[154] A. Villa, J. Chanussot, J. Benediktsson, M. Ulfarsson, and C. Jutten, "Super-resolution: an efficient method to improve spatial resolution of hyperspectral images," in *Proceedings IGARSS 2010*, 2010.

[155] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, 2005.

[156] G. Foody and H. Doan, "Variability in soft classification prediction and its implications for sub-pixel scale change detection and super-resolution mapping," *Photogramm. Eng. Remote Sens.*, vol. 73, pp. 923–933, 2007.

[157] Z. Kulpa, "More about areas and perimeters of quantized objects," *Computer Vision, Graphics, and Image Processing*, vol. 22, no. 2, pp. 268–276, May 1983.

[158] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, pp. 1087–1092, 1958.

[159] S. Kirkpatrick, C. Gelatt, and M. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.

[160] P. Atkinson, *Innovations in GIS*. New York Taylor and Francis, 1997, vol. 4, ch. Mapping sub-pixel boundaries from remotely sensed images, pp. 166–180.

[161] P. M. Atkinson, "Super-resolution target mapping from soft-classified remote sensed imagery," *Photogramm. Eng. Remote Sens.*, vol. 71, no. 7, pp. 839–846, 2005.

[162] Y. Makido and A. Shortridge, "Weighting function alternatives for a sub-pixel allocation model," *Photogramm. Eng. Remote Sens.*, vol. 73, no. 11, pp. 1233–1240, 2007.

[163] R. N. Clark, G. A. Swayze, R. Wise, K. E. Livo, T. M. Hoefen, R. F. Kokaly, and S. J. Sutley, "Usgs digital spectral library splib06a," U.S. Geological Survey, Data Series 231, 2007, available on-line at: http://speclab.cr.usgs.gov/spectral.lib06/ds231/datatable.html.

[164] P.-F. Hsieh, L. Lee, and N.-Y. Chen, "Effect of spatial resolution on classification errors of pure and mixed pixels in remote sensing," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 12, pp. 2657 –2663, dec 2001.

[165] G. Camps-Valls and L. B. (eds.), *Kernel methods for Remote Sensing Data Analysis.* John Wiley and Sons, 2009.

[166] J. Li and L. M. Bruce, "Wavelet-based feature extraction for improved endmember abundance estimation in linear unmixing of hyperspectral signals," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 644–649, Mar. 2004.

[167] L. Wang and X. Jia, "Integration of soft and hard classifications using extended support vector machines," *IEEE Geosci. Remote Sens. Letters*, vol. 6, no. 3, pp. 543–547, Jul. 2009.

[168] P. M. Atkinson, "Super-resolution target mapping from soft-classified remote sensed imagery," in *Online Proceedings of the 6th International Conference on GeoComputation*, 2001.

[169] K. Mertens, "Sub-pixel mapping: A comparison of techniques," Ph.D. dissertation, Ghent University, 2008.

[170] A. Bardossy and L. Samaniego, "Fuzzy-rule based classification of remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 2, no. 2, pp. 362–274, 2002.

[171] P. Debba, E. Carranza, F. van der Meer, and A. Stein, "Abundances estimation of spectrally similar minerals by using derivative spectra in simulated annealing," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 12, pp. 3469–3658, Dec. 2006.

[172] V. Vapnik, *Statistical Learning Theory.* New York: Wiley, 1998.

[173] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote-sensing images with support vector machines„" *IEEE Trans. Geosci. Remote Sens.*, vol. 42, pp. 1778–1790, 2004.

[174] G. Licciardi, F. Pacifici, D. Tuia, S. Prasad, T. West, F. Giacco, C. Thiel, J. Inglada, E. Christophe, J. Chanussot, and P. Gamba, "Decision fusion for the classification of hyperspectral data: Outcome of the 2008 grs-s data fusion contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, pp. 3587–3865, 2009.

[175] N. Aronszajn, "Theory of reprodusing kernel," Harvard University, Division of engineering sciences, Tech. Rep., 1950.

[176] J. Platt, *Advances in Large Margin Classifiers.* Cambridge, MA: MIT Press, 2000, ch. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods.

[177] H. Lin, C. Lin, and R. C. Weng, "A note on platt's probabilistic outputs for support vector machines," Dept. Comput. Sci., Nat. Taiwan Univ., Taipei, Taiwan, Tech. Rep., 2003.

[178] C. Hsu and C. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[179] J. A. Gualtieri and R. F. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *Advances in Computer-Assisted Recognition*, ser. Proceedings of SPIE, R. J. Mericsko, Ed., vol. 3584, 1999, pp. 221–232.