INSTITUTO
SUPERIOR
TÉCNICO

UNIVERSIDADE TÉCNICA DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Discriminative Image Segmentation: Applications to Hyperspectral Data

Jun Li

Orientador: Doutor José Manuel Bioucas Dias

Co-orientador: Doutor Antonio J. Plaza

Thesis specifically prepared to obtain the PhD Degree in

Electrical and Computer Engineering

Draft

May 2011

Abstract

This thesis proposes several new techniques for hyperspectral image segmentation based on discriminative Bayesian approaches, where the posterior class distributions are modeled by the multinomial logistic regression (MLR) model and the spatial information is modeled by means of Markov random fields (MRFs). Our proposed framework introduces significant innovations with regards to previous approaches in the same field, many of which are mainly based on exploiting the spectral information alone. Another contribution of the thesis is that we enhance our proposed supervised techniques with semi-supervised learning capabilities, thus exploiting unlabeled samples by means of an active learning paradigm. Furthermore, the thesis introduces new active sampling strategies based on labeled query selection which are thoroughly discussed and compared with previous developments in the same field. Finally, we also develop subspace-based techniques that can better discriminate land-cover classes in the presence of heavily mixed pixels. The effectiveness of the proposed techniques is illustrated by comparing with state-of-the-art methods by using both simulated and real hyperspectral data sets.

Keywords:

Hyperspectral segmentation, active learning, semi-supervised learning, multinomial logistic regression, multi-level logistic prior, unlabeled samples, subspace learning, ill-posed problems.

Resumo

Esta tese introduz várias contribuições em segmentaçõo de imagem hiperespectral explorando abordagens discriminativas num quadro conceptual bayesiano. Para um dado pixel, a distribuição *a posteriori* das classes, dado o respectivo vector espectral, é modelizada pela regressão logística multinomial. A informação contextual espacial contida nas imagem hiperespectrais é modelizada por um campo estocástico de Markov, mais especificamente pelo modelo logístico multi-nível. Relativamente ao estado-da-arte em classificação hiperespectral, a presença de informação espacial é um elemento distintivo de todas contribuições. Foram consideradas métodos semi-supervisionados e supervisionados com aprendizagem activa baseada, respectivamente, em amostras sem e com etiquetas. Finalmente, desenvolveu-se uma nova técnica de segmentação baseada em subespaços concebida para lidar com os chamados "pixeis misturados", que frequentemente surgem em imagens hiperespectrais de média e baixa resolução. Referem-se duas componentes principais de todas as novas abordagens introduzidas: a) a eficiência dos algoritmos de aprendizagem propostos e b) a qualidade das aproximações, obtidas por em técnicas de cortes em grafos, para os problemas de optimização inteira associados segmentação de máxima probabilidade *a posteriori*. A eficiência e competitividade dos métodos propostos é documentada através de comparações exaustivas com o estado-da-arte usando imagens hiperespectrais reais e simuladas.

Palavras-Chave:

Segmentação hiperespectral, regressão logística multinomial, modelo logístico multi-nível, abordagen discriminativa, aprendizagem activa, aprendizagem supervisionada, aprendizagem náo-supervisionada, aprendizagem baseda em sub-espaços, optimização inteira, cortes em grafos.

Acknowledgments

First of all, I would like to thank my supervisors, Professor José Manuel Bioucas Dias and Professor Antonio J. Plaza. I thank Professor José Bioucas Dias for all of his help, support, guidance, patience and everything, who is not only a supervisor for research but also a father for life. I thank Professor Antonio J. Plaza for all his important contributions, comments, ideas, also his kindness and friendship.

I would like to thank Professor José Leitão for his help and support. I would like to thank Professor Jón Atli Benediktsson for our outstanding collaboration. I would like to thank all of the professors giving me invaluable knowledge in their courses: Professor Jorge Marques, Professor Mário Figueiredo, Professor João Xavier, and Professor Paulo Oliveira. I would like to thank my colleagues in TCRP group, IST: Daniel, Nuria, Manya, Andre, João, Gonçalo, Marta, Mariana, Rong, Liliana, Sónia, José Nascimento, Hugo Silva; and my colleagues in HyperComp, UEX: Javier, Gabriel, Sergio, Inmaculada, Abel, Alberto, Maciel, Prashanth and all of the other colleagues I do not mention here who definitely deserve my sincere acknowledgments.

I would like to thank all of my friends Cynthia, Sabrinna, Li Yigang, Liu Wei, Peter, Lumi, Nuno, Baiqiao, Huidong, Mingcai, Hanbing, Lidong, Xiao Liping, Yang Jie, Xueqian, Xuemei, He Hui, Jiang Youcheng, Li Kun, Li Lireng, Wang Fangxin, Mao Jing, Ma Zhen, Wang Gongliang, Yu Ning, Fengliang, Liang Dawei, Zhang Shixia ...

I would like to thank my dear friends Hugo Gamboa, Ana Rita and their lovely daughters: Alice, Diana and Carolina. I have much more than thanks to say, my friends.

Special words go to my dear Portuguese Mother and Grandmother. Kiss and miss, my dears.

Last but not least, I would like to thank my parents, my sisters and brothers, Dong and my baby, Miguel.

To Miguel Zeng

Table of Contents

Chapter 1

Introduction

1.1 Context

The work presented in this thesis was supported by the European Community's Marie Curie Research Training Networks Program under contract MREST-CT-2005-021175 (European Doctoral Program in Signal Processing, SIGNAL), by a Instituto de Telecomunicações (IT) PhD grant and by the Spanish Ministry of Science and Innovation (HYPERCOMP/EODIX project, reference AYA2008-05965-C04-02). The SIGNAL project has been awarded funding (about 3.000.000 Euros) for 16 PhD grants + 9 short stays) from the EU Human Resources and Mobility program. These Early Stage Research Training Host Fellowships are the most competitive EU Marie Curie Actions. SIGNAL is a consortium of four universities:

- Signal and Systems laboratory (I3S), University of Nice, France.
- KOM department, University of Aalborg, Denmark.
- Technical Institute (IST), University of Lisbon, Portugal.
- Signal and Systems division (ESAT), University of Leuven, Belgium.

SIGNAL was aimed at providing a unified training in signal processing, focusing on the fundamental research aspects of signal processing, offering early stage researchers (ESRs) an in-depth knowledge of the field, not restricted to a particular sub-domain of applications. Moreover, due to the strong links of the participants in industrial projects and in various types of applications, the researchers had the opportunity to apply their results in the real world.

The author of this thesis, Ms. Jun Li, joined SIGNAL as an *ESR* in September 2007, when she started her research activity at Instituto Superior Técnico (IST), Lisbon, Portugal, under the joint supervision of Prof. José M. Bioucas Dias and Prof. Antonio Plaza from University of Extremadura (UEX), Cáceres, Spain. She registered as a PhD student at IST in October 2008. Her contract with SIGNAL project ended in April 2010. Then she was supported by an

Table 1.1: Overview of some present and future remote sensing missions including hyperspectral sensors.

	Hyperion*	**Prisma†**	**EnMAP‡**	**HyspIRI§**
Country of origin	USA	Italy	Germany	USA
Spatial Resolution	30 meters	5-30 meters	30 meters	60 meters
Revisit Time	16 days	3/7 days	4 days	18 days
Spectral Range	400-2500 nanometers	400-2500 nanometers	420-2450 nanometers	380-2500 nanometers
Spectral Resolution	10 nanometers	10 nanometers	6.5-10 nanometers	10 nanometers
Swath width	7.7 kilometers	30 kilometers	30 kilometers	120 kilometers
Earth coverage	Partial	Full	Full	Full
Launch	2000	2010	2012	2018
Lifetime	10 years	$\approx$ 6 years	$\approx$ 6 years	$\approx$ 6 years

*http://eo1.gsfc.nasa.gov †http://www.asi.it/en/flash_en/observing/prisma ‡http://www.enmap.org §http://hyspiri.jpl.nasa.gov

IT PhD grant for 6 months from May to October 2010. Then she was appointed as a researcher with the Hyperspectral Computing Laboratory (HyperComp) research group coordinated by Prof. Antonio J. Plaza at the Department of Technology of Computers and Communications, University of Extremadura, Cáceres, Spain.

1.2 Thesis overview

This thesis addresses the problem of remotely sensed hyperspectral image segmentation. Remotely sensed hyperspectral imaging instruments are capable of collecting hundreds of images, corresponding to different wavelength channels, for the same area on the surface of the Earth. The concept of hyperspectral imaging was first introduced at NASA's Jet Propulsion Laboratory [59], where a system called Airborne Imaging Spectrometer (AIS) was built to demonstrate this technology. Today, NASA is continuously gathering high-dimensional image data with instruments such as Jet Propulsion Laboratory's Airborne Visible-Infrared Imaging Spectrometer (AVIRIS). This advanced sensor for Earth observation records the visible and near-infrared spectrum of the reflected light using more than 200 spectral bands, thus producing a stack of images in which each pixel (vector) is represented by a spectral signal that uniquely characterizes the underlying objects (see Figure 1.1). Nowadays, the concept of hyperspectral imaging is extended to describe systems with hundreds to thousands of spectral channels, with many new instruments currently in development for spaceborne operation. Table 1.1 presents a summary of several hyperspectral sensor systems (of satellite type) which are currently in operation or under development.

The number and variety of processing tasks in hyperspectral remote sensing is enormous [107]. However, the majority of algorithms can be organized according to the following specific tasks [125]:

- *Dimensionality reduction* consists of reducing the dimensionality of the input hyperspectral

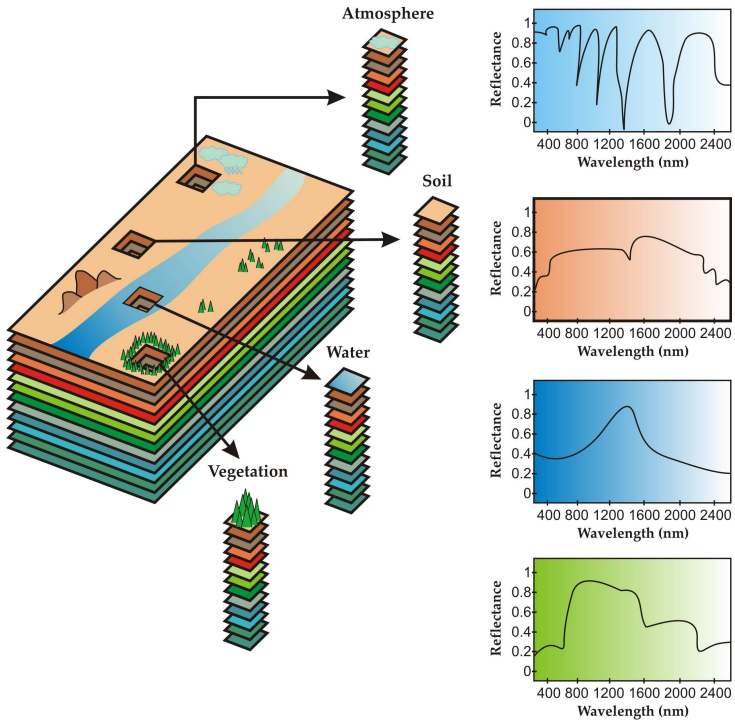


Figure 1.1: Hyperspectral data cube.

scene in order to facilitate subsequent processing tasks.

- *Target and anomaly detection* consist of searching the pixels of a hyperspectral data cube for "rare" (either known or unknown) spectral signatures.

- *Change detection* consists of finding the "significant" (i.e., important to the user) changes between two hyperspectral scenes of the same geographic region.

- *Classification/segmentation* consist of assigning a label to each pixel/region in order to generate a thematic land-cover map.

- *Spectral unmixing* consists of estimating the fraction of the pixel area covered by each material present in the scene.

In this thesis, we particularly focus on the problem of supervised and semi-supervised hyperspectral image segmentation (i.e., how to partition an image into spatially consistent regions associated to different land-cover classes starting from some –limited– reference information available *a priori*). The high dimensionality of hyperspectral data in the spectral domain poses critical problems for supervised algorithms [19, 107], most notably, in order for supervised classifiers to perform properly there is a need for large training sets in order to avoid the well-known Hughes effect [68, 84]. However, training samples are limited, expensive and very difficult to obtain in real remote sensing scenarios.

Further, in this problem it is very important to take advantage of the fact that, in addition to the very rich spectral information available in the hyperspectral data, hyperspectral images exhibit (as many other classes of images) some kind of piecewise statistical continuity among neighboring pixels. As a result, hyperspectral image segmentation should exploit such spatial information in conjunction with spectral information in order to partition an image into a set of homogeneous regions (in statistical sense). In this regard, hyperspectral image segmentation provides an extension of multi-class image classification, where the spatial interdependencies among class labels are enforced by a suitable model. Without loss of generality, in this thesis we will use the term *classification* when the learning process only considers the spectral information. Similarly, we use the term *segmentation* when the spatial contextual information in the original scene is used.

In the thesis, we particularly focus on the problem of supervised and semi-supervised hyperspectral image segmentation. We introduce several new Bayesian approaches for hyperspectral image segmentation which include spatial-contextual information in the analysis. Another important contribution of the thesis is the inclusion of unlabeled samples (which are easy to obtain in practice) by means of active learning paradigms. Finally, we also develop innovative strategies to cope with one of the most important problems in hyperspectral image analysis: the presence of mixed pixels (with possibly many participating constituents at a sub-pixel level) due to limited spatial resolution, mixing phenomena happening at different scales, etc. For instance, the pixel vector labeled as "vegetation" in Figure 1.1 may actually comprise a mixture of vegetation and soil, or different types of soil and vegetation canopies. To address this issue we resort to subspace-based techniques that can better discriminate land-cover classes in the presence of heavily mixed pixels.

Combined, these topics intend to address cutting-edge problems in hyperspectral image analysis and interpretation. To introduce these topics, which will be presented in detail in the remaining chapters of the present document, we have organize the rest of this introductory chapter as follows. In Section 1.2.1, we focus on the classification of hyperspectral images, describing the advantages of discriminative versus generative models in our context, and present the state-of-the-art in discriminative hyperspectral image classification. Then, we focus on two widely used discriminative methods: the multinomial logistic regression (MLR) and the support vector machine (SVM). In Section 1.2.2 we discuss the importance of integrating spatial and spectral information in hyperspectral image segmentation. We also describe related works in this area. Furthermore, in Section 1.2.3 we specifically address the problems that supervised classifiers can found when limited training sets are available. Then, we present available solutions

in the literature to cope with this problem by adopting semi-supervised learning and active learning strategies. Finally, in Section 1.2.4 we present and categorize our main contributions in this thesis. We particularly address the strategies that we have adopted in order to overcome the aforementioned problems.

1.2.1 Hyperspectral image classification

The problem of hyperspectral image classification has been tackled in the past using several different approaches. For instance, several machine learning and image processing techniques have been applied to extract relevant information from hyperspectral data during the last decade [107]. In the context of supervised classification, a relevant challenge is the fact that we need to deal with very high-dimensional data volumes (with limited training samples available *a priori*). In other words, due to the small number of training samples and the high number of features available in remote sensing applications, reliable estimation of statistical class parameters is a very challenging goal [85]. As a result, with a limited training set, classification accuracy tends to decrease as the number of features increases. This is known as the Hughes effect [68]. High-dimensional spaces have been demonstrated to be mostly empty [72], thus making density estimation even more difficult. One possible approach to handle the high-dimensional nature of hyperspectral data sets is to consider the geometrical properties rather than the statistical properties of the classes, which leads to the use of kernel methods which have been shown to be a very effective tool for hyperspectral image interpretation [28]. Another widely used solution is to resort to Bayesian techniques [70], possibly combined with spatial-contextual information.

In this context, we can define the classification problems in mathematical terms as follows. Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be the observed data (as collected by an imaging spectrometer) and n is the number of pixels, and $\mathbf{y} = (y_1, \ldots, y_n)$ be the label configuration (available *a priori*). The posterior density $p(\mathbf{y}|\mathbf{x})$ is the central element of the risk-based inference, and we associate the term *classification* with $\widehat{y}_i \Leftarrow \arg\max p(y_i|\mathbf{x}_i)$ (only spectral information is considered) and the term *segmentation* with $\widehat{\mathbf{y}} \Leftarrow \arg\max p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$ (both spectral and spatial information are considered). In general, there are two rather different points of view in modeling this density $p(\mathbf{y}|\mathbf{x})$, namely, the *generative* approach versus the *discriminative* approach:

Generative approaches: Correspond to the widely used Bayesian perspective, according to which $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$, where $p(\mathbf{x}|\mathbf{y})$ is the likelihood density, accounting for the image features given the label configuration and $p(\mathbf{y})$ is the a priori label configuration density. Correct modeling of these two densities is a serious challenge, compelling model simplifications such as

the conditional independence, which assumes that $p(\mathbf{x}|\mathbf{y})$ has a factorized form. This problem worsens in high dimensional feature spaces, where, usually, $p(\mathbf{x}|\mathbf{y})$ depends on large number of parameters.

Discriminative approaches: In a discriminative framework, the class densities $p(\mathbf{y}|\mathbf{x})$ are modeled directly, thus avoiding the learning of the likelihood densities $p(\mathbf{x}|\mathbf{y})$. The underlying rationale is that learning the class posterior $p(\mathbf{y}|\mathbf{x})$ is equivalent to learn the boundaries among the classification regions, what is expected to be simpler and more flexible than learning the likelihood densities $p(\mathbf{x}|\mathbf{y})$. Examples of discriminative learning in classification include logistic regression, neural networks, the Gaussian process, and generalized additive models. Examples of recent frameworks aimed at discriminative data segmentation are the conditional random fields (CRFs) [83], the discriminative random fields (DRFs) [82], and the Gaussian process [2].

In the past, both discriminative and generative models have been used for hyperspectral image interpretation [6, 25, 27, 28, 30, 34, 37, 38, 53, 90, 91, 132]. We refer to [107] for a seminal view on recent advances in techniques for hyperspectral image processing. However, and mainly because of the special difficulties that arise in hyperspectral data interpretation (including high dimensionality, limited availability of training information, presence of mixed pixels, large datasets, etc.) discriminative models are often preferred since they are widely regarded as less complex than generative models. As a result, discriminative approaches can mitigate the curse of dimensionality introduced by the Hughes effect, because they demand smaller training sets than generative models [13, 102, 134]. Data interpretation based on the use of discriminant functions, which basically encode the boundaries between classes in the feature space, is another effective way of handling very high dimensional data sets [13]. In the following, we provide a brief literature review of state-of-the-art discriminant classification approaches which only consider the spectral information. A more detailed introduction of techniques combining both spatial and spectral information will be provided in subsequent sections of this chapter.

I. Discriminant analysis

Linear Discriminant Analysis (LDA), which is based on optimizing the so-called Fisher's score, has been successfully used in many practical remote sensing classification applications aimed at generating thematic maps in different contexts. For instance, in [122] an investigation based on the use of linear discriminant and profile analyses for airborne thematic mapper data was conducted. In [61], classical LDA was used for recognition of different conifer species using hyperspectral data. In [39], LDA has been used for classification of tropical rain forest tree

species using hyperspectral data at different scales. In [94], the canonical LDA has been used for identifying land cover units in ecology. In [46], a linear constrained distance-based discriminant analysis (LCDA) was introduced which not only maximizes the ratio of inter-distance between classes to intra-distance within classes but also imposes a constraint that all class centers must be aligned along predetermined directions, with practical use in several different applications. In [47], a constrained linear discriminant analysis (CLDA) approach was proposed for hyperspectral image detection and classification as well as its real-time implementation. In [6], the regularized LDA (RLDA) [139] was introduced for hyperspectral hyperspectral classification problems where in comparison with LDA-based classifiers, *i.e.*, standard LDA, penalized LDA [66], orthogonal LDA [140], and uncorrelated LDA [74] are also discussed. In [111], a new kernel discriminant analysis-based projection approach was proposed. In [73], a novel approach based on Fisher discriminant null space was proposed for decomposition of mixed pixels in hyperspectral imagery. Some of these approaches will be used in this thesis as comparative frameworks for evaluating our newly proposed techniques.

II. Support vector machines

Perhaps the most popular discriminative classifier in the remotely sensed hyperspectral image community is the support vector machine (SVM) [21, 121], which is characterized by its ability to effectively deal with large input spaces (and to produce sparse solutions) using limited training samples. This classifier has been successfully used in the context of supervised and semi-supervised hyperspectral classification problems [29, 64, 65, 67, 98]. In [28], a framework for kernel-based methods in the context of hyperspectral image classification applications was presented. Specifically, standard SVMs, regularized radial basis function neural networks (Reg-RBFNN), kernel Fisher discriminant (KFD) analysis, and regularized AdaBoost (Reg-AB) were analyzed and inter-compared. The KFD is in fact another effective discriminative method for hypespectral image classification [50, 99] which benefits from the concept of kernels used in SVMs to obtain nonlinear solutions. In [25], a novel transductive SVMs (TSVMs) was introduced for semi-supervised classification exploiting the unlabeled information based on a weighting strategy. In [30], a framework based on composite kernel machines for enhanced classification of hyperspectral images was proposed which exploits the properties of Mercer's kernels to construct a family of composite kernels that easily combine spatial and spectral information. In [27], a new graph-based semi-supervised algorithm was proposed for hyperspectral image classification problems which efficiently alleviates the curse of dimensionality by exploiting the wealth of unlabeled information through a graph-based methodology. In [60], a Laplacian SVM (LapSVM)

was presented for semi-supervised image classification based on kernel machines where SVMs is regularized with the unnormalized graph Laplacian. In [132], a semi-supervised support vector machine with cluster kernel was presented. As mentioned in some of the aforementioned references, an important observation is that the good classification performance demonstrated by SVMs can be complemented by taking advantage of semi-supervised learning and contextual information. However, the integration of spatial and spectral information is generally done through the combination of dedicated kernels to spectral and contextual information [30]. The desired integration can also be accomplished at the feature extraction level, i.e., by reducing the dimensionality of the input data to a proper subspace in a way that both spatial and spectral information is considered [105]. On the other hand, in semi-supervised learning the wealth of unlabeled data that can be obtained from hyperspectral images is exploited. These novel SVM formulations represent significant developments in which spatial and spectral information can be easily integrated and analyzed by using proper kernel functions. The capability of semi-supervised SVMs to capture the intrinsic information present in the unlabeled data can further mitigate the Hughes phenomenon, and the problems related to the non-stationary behavior of the spectral signatures of classes in the spatial domain [25].

III. Multinomial logistic regression

One of our main contributions in this work is the use of multinomial logistic regression (MLR) discriminative classifiers [16], which exhibit some advantages (under certain circumstances) with regards to previously discussed methods. One of them is the ability to learn the class distributions themselves, which has recently resulted in the successful application of this kind of discriminative classifier to hyperspectral image classification problems [20, 90, 91]. Sparse MLR (SMLR) [80] adopts a Laplacian prior enforcing sparsity and therefore controlling the machine generalization capabilities. Fast sparse MLR (FSMLR) implements an iterative procedure to calculate the MLR regressors that is $O(K^2)$ faster than the original SMLR algorithm in [80] (where K is the number of classes). In [12], the logistic regression via splitting and augmented Lagrangian (LORSAL) algorithm [90] opened the door to processing of hyperspectral images with a very large number of classes. In [20], a Jeffreys prior [71] is adopted to avoid the high computational complexity invovled in estimating the Laplacian regularization parameters. In [90, 91] we showed that very good performance can be obtained by setting (in suboptimal sense) the Laplacian regularization parameter. Therefore, no cross-validation is performed in our work. This has the advantage of reducing the computational cost. Overall, one of the main purposes of this thesis work is to illustrate the advantages that MLR can offer in hyperspectral image

classification and segmentation. These aspects will be emphasized in subsequent sections of this document.

1.2.2 Hyperspectral image segmentation

In order to improve the accuracies obtained by hyperspectral image classification, a recent trend is to integrate the spectral and spatial information in the data interpretation. As shown in the previous section, many hyperspectral image classification techniques are focused on analyzing the data without incorporating information on the spatially adjacent data, i.e., hyperspectral data are usually not treated as images, but as unordered listings of spectral measurements with no particular spatial arrangement. However, the importance of analyzing spatial and spectral patterns simultaneously has been identified as a desired goal by many scientists devoted to multidimensional data analysis [107]. This type of processing has been approached in the past from various points of view. For instance, several possibilities are discussed in [85] for the refinement of results obtained by spectral-based techniques in multispectral imaging through a second step based on spatial context. Such contextual classification, extended also to hyperspectral images [72], accounts for the tendency of certain ground cover classes to occur more frequently in some contexts than in others. In certain applications, however, the integration of spatial and spectral information is mandatory to achieve sufficiently accurate mapping and/or detection results. For instance, urban area mapping requires sufficient spatial resolution to distinguish small spectral classes, such as trees in a park, or cars on a street [7, 42]. This poses two main challenges:

1. We need to manage very high-dimensional data volumes in which the spatial correlation between spectral responses of neighboring pixels can be potentially high. As a result, there is a need to incorporate the spatial arrangement of the data in the development of robust analysis techniques.

2. Processing algorithms need to become more knowledge-based. With finer spatial resolutions, subtle details which can greatly improve scene interpretation may also be misleading in certain applications. This suggests that *a priori* knowledge may be used to improve the characterization of single elements, as well as the whole scene.

At this point, we should recall again that our terminology in this document will be to address hyperspectral image segmentation when we are combining both spatial and spectral information in the analysis of the hyperspectral data. In the following subsections we briefly review recent advances in this area, including some of the aforementioned techniques for spatial-spectral integration as well as mathematical morphology-based approaches and their extension

to hyperspectral image processing [17, 20, 33, 72, 90, 91, 127–130, 135]

I. Extended morphological profiles

Mathematical morphology is a theory for the analysis of spatial structures in image data which has been successfully applied to remotely sensed images [106]. To analyze the structures of an image in a systematic way, the morphological profile was first constructed based on the granulometry principle [106]. Such profiles were then adapted to hyperspectral images by means of extended morphological profiles (EMP) [7], which is built on the morphological profiles (MP) [8] by applying morphological operators of erosion and dilation (and their shape-preserving counterparts: opening and closing by reconstruction) on the components obtained after performing a dimensionality reduction on the original hyperspectral image. EMPs provide an intuitive idea of both the spectral characterization of the pixel vectors in the data and the spatial distribution of such pixels in the scene. In [42], a new method based on the combination of spatial reclassification and mathematical morphology concepts was implemented to process hyperspectral data collected over urban environments. In [53], a joint spectral-spatial classification algorithm was developed for hyperspectral data by using SVMs and morphological profiles. This approach is in fact an extension of the seminal work in [7, 106]. Additional efforts on the integration of spatial and spectral information using mathematical morphology concepts can be found in [127–130].

II. Markov random fields

Another widely used strategy in the literature to integrate spatial information in hyperspectral image classification is through the use of Markov random fields (MRFs), which model the piecewise statistical continuity among neighboring pixels that is expected in real-world scenarios [42]. In this regard, MRFs exploit the continuity, in probability sense, of neighboring labels. MRF is a powerful tool for spatial analysis. Its basic assumption is that, in a hyperspectral image, it is very likely that two neighboring pixels will have the same class label. This simple concept has been explored in [70], in which an adaptive Bayesian contextual classification procedure that utilizes both spectral and spatial inter-pixel dependencies was proposed, where the joint prior probabilities of the classes of each pixel and its spatial neighbors are modeled by an MRF. In [85], spatial characterization and post-processing is performed to the discriminant analysis feature extraction (DAFE) method by modeling the spatial neighborhood of a pixel as a spatially distributed random process. Then, a spatial regularization is performed via the minimization of an energy functional. In [52], a *maximum a posteriori* (MAP)-based framework was proposed in which the class conditional probabilities were learnt by the SVM algorithm and the class prior

probabilities were modeled by MRFs. In [20, 90, 91], an MRF multi-level logistic (MLL) prior [58] is adopted in the Bayesian framework where the MAP estimate is efficiently computed by the α-expansion min-cut-based integer optimization algorithm [23]. This is a crucial step since the MRF is characterized by its computational complexity. In fact, one of the main contributions of this thesis is the integration of MRFs with discriminative classifiers for computationally efficient spatial-spectral segmentation of hyperspectral images.

1.2.3 Semi-supervised and active learning

As mentioned in previous sections of this chapter, in supervised hyperspectral image classification and segmentation we often have to deal with the limited availability of training samples. This is because, normally, labeled samples are often very difficult, expensive or time consuming to collect. With a limited training set, classification accuracy tends to decrease as the number of features increases. In this section, we discuss several approaches to deal with this problem, including semi-supervised learning and active learning, which have become very active research areas in hyperspectral image classification/segmentation.

I. Semi-supervised learning

The performance of hyperspectral image classification and segmentation techniques can be further increased by taking advantage of semi-supervised learning, in which the learning is generally conducted using very few labeled samples (available *a priori*) and a larger amount of so-called *unlabeled* training samples which are automatically generated during the process and with no extra cost. Recently, several semi-supervised methods have become widely popular, including those based on models [5, 15, 56, 103], self-learning strategies [115, 117, 138], co-training [15, 100], multiview learning [24, 119, 126], transductive SVMs [75, 134], and graph-based methods [14, 145]. We refer to [144] for a detailed survey on semi-supervised methods. It should be noted that most available semi-supervised learning algorithms use some type of regularization which encourages that "similar" features are associated to the same class. The effect of this regularization is to push the boundaries between classes towards regions with low data density [32], where a rather usual way of building such regularizers is to associate the vertices of a graph to the complete set of samples and then apply the regularizer to the variables defined on the vertices. This trend has been successfully adopted in several remote sensing studies [25, 27, 91, 107, 132, 143]. Some of the methods developed in this thesis are based on this strategy.

II. Active learning

In order to reduce the cost of acquiring large labeled training sets, another strategy in the literature has been active learning. Active learning is a method of online learning, where a learner strategically selects new training examples that provide maximal information about the unlabeled dataset, resulting in higher classification accuracy for a given training set size as compared to using randomly selected examples. Active learning is most useful when there are sufficient a number of unlabeled samples but it is expensive to obtain class labels. This strategy have been successfully applied in several different classification and segmentation problems. In [96], a mutual information (MI)-based technique for active sampling was proposed for data query selection, which maximizes the the entropy of labels [81]. In [95], an algorithm called breaking ties (BT) was proposed for multi-class SVMs using the one-vs-one approach with a probability approximation which tends to minimize the distance between the first two most probable classes. In [101], another active sampling approach for SVM classifiers was proposed based on the distance of the unlabeled data points from the existing hyperplane. In [113], an active sampling approach which maximizes the Kullback-Leibler divergence of the new label and the training set was developed. In [133], a survey of active sampling approaches was presented in the context of remote sensing classification problems, including: (a) the margin sampling (MS) [26, 120] strategy, which samples the candidates lying within the margin of the current SVM by computing their distance to the dividing hyperplane [101]; (b) the class of active learning methods which relies on the estimation of the posterior probability distribution functions of the classes; (c) the last class of active methods which is based on the query-by-committee paradigm [41, 55, 124]. Among these active learning algorithms, most of them naively select the data point with maximum label entropy, least confidence, or maximum disagreement between multiple learners. In this work, we exploit active learning principles to increase the accuracy of methods for hyperspectral image classification/segmentation at no cost, and further develop a new sampling method which overcomes some of the limitations of the aforementioned techniques for the same purpose.

1.2.4 Thesis contributions

This section summarizes the main topics addressed in the thesis and its main contributions. As indicated in Fig. 1.2, which graphically represents the main contributions in this work and their relationship, we particularly focus on the problem of supervised and semi-supervised hyperspectral image segmentation (i.e., how to partition an image into spatially consistent regions associated to different land-cover classes starting from some –limited– reference information

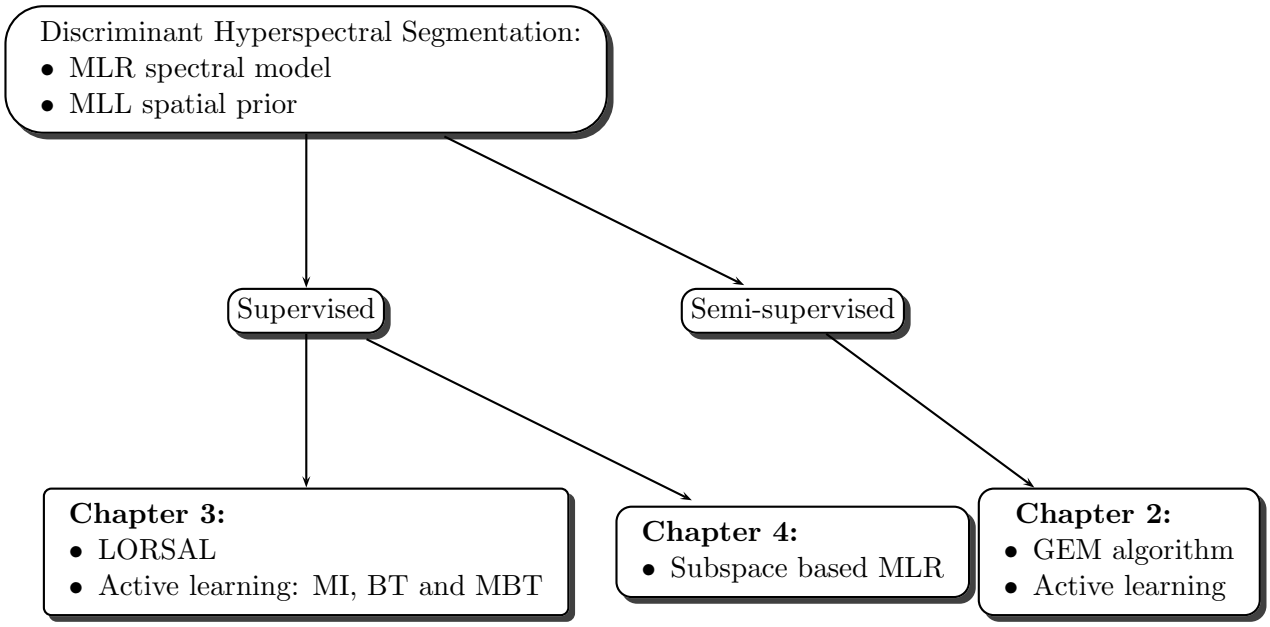


Figure 1.2: Scheme summarizing the thesis organization.

available *a priori*). For this purpose, we introduce several new Bayesian discriminative approaches based on the MLR discriminative model (much less complex compared with generative models since it has much less parameters to learn). In this work, we focus on discriminative approaches based on MLR classifiers for several reasons:

1. First and foremost, MLR classifiers are able to learn directly the posterior class distributions and deal with the high dimensionality of hyperspectral data in a very effective way.

2. Second, we adopt a sparsity inducing prior on the regressors in order to obtain sparse estimates. As a result, most of the components of the regressors are zero. This allows us to control the complexity of our proposed techniques and their generalization capacity.

3. Finally, the MLR provides the class posterior probability. This plays a crucial role in the complete posterior probability which includes spectral and spatial information.

These aspects allowed us to introduce significant innovations in the context of supervised and semi-supervised hyperspectral image segmentation, such as the use of prior probability distribution based on the MRF which promotes piecewise segmentation results with smooth transitions between neighboring class labels. Our probabilistic discriminative framework has

two main advantages. First, it provides a probabilistic interpretation of each label configuration, thus opening the door to compute risk-based segmentations, such as the MAP. Second, the probabilistic setup provides suitable tools to infer model parameters. In our spatial prior, we encourage piecewise smooth segmentations and thus promote solutions in which adjacent pixels are likely to belong to the same class. In the past, the MAP segmentation was very complex to compute. However, with the advent of graph-cut tools [22, 23, 79], we can now compute the MAP estimate efficiently via efficient min-cut based integer optimization techniques.

Another important contribution of the thesis is based on the observation that training samples are limited, expensive and difficult to obtain in real analysis scenarios. To address this common situation, we enhance our proposed supervised techniques (based on labeled training samples) with semi-supervised learning capabilities, thus exploiting unlabeled samples by means of an active learning paradigm. In this regard, the thesis introduces new active sampling strategies based on labeled query selection which are thoroughly discussed and compared with previous developments in the same field. Finally, we also develop innovative strategies to cope with one of the most important problems in hyperspectral image analysis: the presence of mixed pixels (with possibly many participating constituents at a sub-pixel level) due to limited spatial resolution, mixing phenomena happening at different scales, etc. To address this issue we resort to subspace-based techniques that can better discriminate land-cover classes in the presence of heavily mixed pixels.

The remainder of the document has been organized so that the specific contributions listed above and summarized in Fig. 1.2 are presented in different chapters. The chapters are organized according to the following arrangement:

- Chapter 2 presents a new semi-supervised segmentation algorithm, where the posterior class distributions are modeled by the MLR and learnt by a new semi-supervised generative expectation minimization (GEM) algorithm, and the spatial contextual information is modeled by a Markov random field multi-level logistic (MLL) prior, which enforces segmentation results in which neighboring labels belongs to the same class.

- Chapter 3 introduces a new supervised Bayesian approach to hyperspectral image segmentation with active learning, where the posterior class distributions are modeled by the MLR model and learnt by the a recently introduced logistic regression via splitting and augmented Lagrangian (LORSAL) algorithm. Another contribution of this work is the introduction of the modified breaking ties (MBT) active sampling scheme, which is an improvement over the breaking ties (BT) sampling method but with the capacity to provide unbiased samplings.

- Chapter 4 presents a new supervised segmentation algorithm for remotely sensed hyperspectral images which integrates the spectral and spatial information in a Bayesian framework. A multinomial logistic regression (MLR) algorithm is first used to learn the posterior probability distributions in spectral sense, using a subspace projection method to better cope with noise and mixed pixels.

The effectiveness of the proposed innovative techniques is illustrated by comparing their performance with state-of-the-art methods for supervised and semi-supervised hyperspectral image segmentation. The comparison is carried out using both simulated and real hyperspectral data sets. The experiments with simulated images are intended to test the newly proposed techniques in controlled analysis scenarios, in which relevant aspects such as the sensitivity of methods to parameter settings or noise can be quantitatively assessed. The experiments with real images have been conducted using widely standardized data sets in the remote sensing community, with the ultimate goal to provide detailed and rigorous comparisons of our newly developed techniques with other widely used strategies for hyperspectral image segmentation. Combined, these topics intend to illustrate the significant advantages that can be obtained by the proposed techniques, which effectively integrate spatial and spectral information for hyperspectral image segmentation. To conclude this chapter, we present the list of publications that have supported the contributions that will be described in the following chapters of this document.

1.2.5 List of publications

1. J. Li, J. Bioucas-Dias and A. Plaza. Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, submitted, 2010.

2. J. Li, J. Bioucas-Dias and A. Plaza. Hyperspectral image segmentation using a new Bayesian approach with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, accepted, 2010.

3. J. Li, J. Bioucas-Dias and A. Plaza. Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, volume 48, pages 4085−4098, 2010.

4. J. Li, J. Bioucas-Dias, and Antonio Plaza. Supervised hyperspectral image segmentation using active learning. In *IEEE 2nd GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, pages 1−4, 2010.

5. J. Li, J. Bioucas-Dias and A. Plaza. Exploiting spatial information in semi-supervised hyperspectral image segmentation. In *IEEE 2nd GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, pages 1−4, 2010.

6. J. Li, J. Bioucas-Dias, and Antonio Plaza. Semi-supervised hyperspectral image segmentation. In *IEEE 1st GRSS Workshop on Hyperspectral Image and Signal Processing*, pages 1−4, 2009.

7. J. Li, J. Bioucas-Dias, and Antonio Plaza. Semi-supervised hyperspectral classification and segmentation with discriminative learning. In *SPIE Europe Remote Sensing*, volume 7477, 2009.

8. J. Li, J. Bioucas-Dias, and Antonio Plaza. Semi-supervised hyperpsectral image classification based on a Markov random field and sparse multinomial logistic regression. In *IEEE International Geoscience and Remote sensing Symposium*, volume 3, pages III−817−III−820, 2009.

9. J. Li, J. Bioucas-Dias, and Antonio Plaza. Hyperspectral image classification based on a fast Bregman sparse multinomial logistic regression algorithm. In *6th EARSeL SIG IS Workshop*, Tel- Aviv, Israel, 2009.

10. J. Li, J. Bioucas-Dias. Minimum volume simplex analysis: a fast algorithm to unmix hyperspectral data. In *IEEE International Geoscience and Remote sensing Symposium IGARSS*, volume 3, pages III−250−III−253, 2008.

Chapter 2

Semi-Supervised Hyperspectral Image Segmentation Using Multinomial Logistic Regression with Active Learning

Abstract – This chapter presents a new semi-supervised segmentation algorithm, suited to high dimensional data, of which remotely sensed hyperspectral image data sets are an example[1]. The algorithm implements two main steps: (i) semi-supervised learning of the posterior class distributions, followed by (ii) segmentation, which infers an image of class labels from a posterior distribution built on the learnt class distributions, and on a Markov random field (MRF). The posterior class distributions are modeled using multinomial logistic regression (MLR), where the regressors are learnt using both labeled and, through a graph-based technique, unlabeled samples. Such unlabeled samples are actively selected based on the entropy of the corresponding class label. The prior on the image of labels is a multi-level logistic (MLL) model, which enforces segmentation results in which neighboring labels belongs to the same class. The maximum a posteriori (MAP) segmentation is computed by the α-Expansion min-cut based integer optimization algorithm. Our experimental results, conducted using synthetic and real hyperspectral image data sets collected by the Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS) system of NASA Jet Propulsion Laboratory over the regions of Indian Pines, Indiana, and Salinas Valley, California, reveal that the proposed approach can provide classification accuracies which are similar or higher than those achieved by other supervised methods for the considered scenes. Our results also indicate that the use of a spatial prior can greatly improve the final results with respect to a case in which only the learnt class densities are considered, confirming the importance of jointly considering spatial and spectral information in hyperspectral image segmentation.

[1]A preliminary much shorter version of this work appeared in [87].

Index Terms – Hyperspectral image classification, semi-supervised learning, multinomial logistic regression (MLR), Markov random field (MRF), multi-level logistic (MLL) model.

2.1 Introduction

In recent years, several important research efforts have been devoted to remotely sensed hyperspectral image segmentation and classification [85]. Hyperspectral image classification and segmentation are related problems. In order to define these problems in mathematical terms, let $\mathcal{S} \equiv \{1, \cdots, n\}$ denote a set of integers indexing the n pixels of a hyperspectral image. Let $\mathcal{L} \equiv \{1, \cdots, K\}$ be a set of K class labels, and let $\mathbf{x} \equiv (\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ denote an image in which the pixels are d-dimensional feature vectors. Finally, let $\mathbf{y} \equiv (y_1, \cdots, y_n) \in \mathcal{L}^n$ denote an image of class labels. The goal of hyperspectral image classification is, for every image pixel $i \in \mathcal{S}$, to infer the class labels $y_i \in \mathcal{L}$ from the feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ (referred to hereinafter as *spectral vectors*). On the other hand, the goal of hyperspectral image segmentation is to partition the set of image pixels $\mathcal{S}$ into a collection of sets $R_i \subset \mathcal{S}$, for $i = 1, \ldots, K$, sometimes called regions, such that the image pixels in each set R_i be *close* in some sense[2]. Nevertheless, in this chapter, we use the term classification when there is no spatial information and segmentation when the spatial prior is being considered.

Supervised classification (and segmentation) of high dimensional datasets such as hyperspectral images is a difficult endeavor. Obstacles, such as the Hughes phenomenon, arise as the data dimensionality increases, thus fostering the development of advanced data interpretation methods which are able to deal with high dimensional data sets and limited training samples [107].

In the past, both discriminative and generative models have been used for hyperspectral image interpretation. More specifically, techniques based on discriminative models learn directly the posterior class distributions, which are usually far less complex than the class-conditional densities in which generative models are supported. As a consequence, discriminative approaches mitigate the curse of dimensionality because they demand smaller training sets than the generative ones [13, 102, 134]. Data interpretation based on the use of discriminant functions, which basically encode the boundaries between classes in the feature space, is another effective way of handling very high dimensional data sets [13].

Support vector machines (SVMs) [121] and MLR [16] rely, respectively, on discriminant functions and posterior class distributions, based on which many state-of-the-art classification

[2]We recall that a partition of a set $\mathcal{S}$ is a collection of sets $R_i \subset \mathcal{S}$, for $i = 1, \ldots,$ where $\cup_{i=1} R_i = \mathcal{S}$ and $R_i \cap R_j = \emptyset,\ i \neq j$

methods are built. Due to their ability to effectively deal with large input spaces (and to produce sparse solutions), SVMs have been successfully used for supervised classification of hyperspectral image data [28, 53, 107, 109]. In turn, MLR-based techniques have the advantage of being able to model the posterior class distributions, thus supplying (in addition to the boundaries between the classes) a degree of plausibility for such classes. Effective sparse MLR methods are available [80]. These ideas have been recently applied to hyperspectral image classification and segmentation, obtaining promising results [19].

In order to improve the accuracies obtained by SVMs or MLR-based techniques, some efforts have been directed towards the integration of spatial (contextual) information with spectral information in hyperspectral data interpretation [19, 53, 107]. However, due to the supervised nature of these methods, their performance is conditioned by the fact that the acquisition of labeled training data is very costly (in terms of time and finance) in remote sensing applications. In contrast, unlabeled training samples can be obtained easily. This observation has fostered active research on the area of semi-supervised learning, in which classification techniques are trained with both labeled and unlabeled training samples [32, 81]. This trend has been successfully adopted in remote sensing studies [25, 27, 107, 132, 143]. Most semi-supervised learning algorithms use some type of regularization which encourages that "similar" features belong to the same class. The effect of this regularization is to push the boundaries between classes towards regions of low data density [32], where a rather usual way of building such regularizer is to associate the vertices of a graph to the complete set of samples and then build the regularizer depending on variables defined on the vertices.

In this chapter, we introduce a new semi-supervised learning algorithm which exploits both the spatial contextual information and the spectral information in the interpretation of remotely sensed hyperspectral data. The algorithm implements two main steps: (i) semi-supervised learning of the posterior class distributions, implemented by an efficient version of semi-supervised learning algorithm in [81], followed by (ii) segmentation, which infers an image of class labels from a posterior distribution built on the learnt class distributions, and on an MLL prior on the image of labels. The posterior class distributions are modeled using MLR, where the regressors are learnt using both labeled and (through a graph-based technique) unlabeled training samples. For step (i), we use a block Gauss-Seidel iterative method which allows dealing with data sets that, owing to their large size (in terms of labeled samples, unlabeled samples, and number of classes) are beyond the reach of the algorithms introduced in [81]. The spatial contextual information is modeled by means of a MLL prior. The final output of the algorithm is based on an MAP segmentation process which is computed via a very efficient min-cut based integer

optimization technique.

The remainder of the chapter is organized as follows. Section 2.2 formulates the problem and describes the proposed approach. Section 2.3 describes the estimation of the multinomial logistic regressors, including a generalized expectation algorithm to compute their MAP estimate, and a fast algorithm based on the Gauss-Seidel iterative procedure. Section 2.4 gives details about the MLL prior. Section 2.5 addresses the MAP computation of the segmentation via integer optimization techniques based on cuts on graphs. An active method for selecting unlabeled training samples is also introduced. Section 2.6 reports performance results for the proposed algorithm on synthetic and real hyperspectral datasets, and compares such results with those provided by state-of-the-art competitors reported in the literature. The two real hyperspectral scenes considered in our experiments were obtained by the AVIRIS over the regions of Indian Pines, Indiana, and Salinas Valley, California. These scenes have been widely used in the literature and have high-quality ground-truth measurements associated to them, thus allowing a detailed quantitative and comparative evaluation of our proposed algorithm. Finally, Section 2.7 concludes with some remarks and hints at plausible future research avenues.

2.2 Problem formulation and proposed approach

With the notation introduced in Section 2.1 in mind, let us define an image region as $R_k \equiv \{i \in \mathcal{S} \,|\, y_i = k\}$, *i.e.,* R_k is the set of image pixels $i \in \mathcal{S}$ with class labels $y_i = k \in \mathcal{L}$. We note that the collection R_i, for $i = 1, \ldots, K$, is a partition of $\mathcal{S}$ and that the map between vectors $\mathbf{y} \in \mathcal{L}^n$, which we term labelings, and partitions of $\mathcal{S}$, which we term segmentations, is one-to-one. We will, thus, refer interchangeably to labelings and segmentations.

The goal of both image classification (and segmentation) is to estimate $\mathbf{y}$ having observed $\mathbf{x}$, a hyperspectral image made up of d-dimensional pixel vectors. In a Bayesian framework, the estimation $\mathbf{y}$ having observed $\mathbf{x}$ is often carried out by maximizing the posterior distribution[3] $p(\mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$, where $p(\mathbf{x}|\mathbf{y})$ is the likelihood function (*i.e.*, the probability of the features image $\mathbf{x}$ given the labeling $\mathbf{y}$) and $p(\mathbf{y})$ is the prior on the labeling $\mathbf{y}$. Assuming conditional independency of the features given the class labels, *i.e*, $p(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{i=n} p(\mathbf{x}_i|y_i)$, then the posterior $p(\mathbf{y}|\mathbf{x})$, as a function of $\mathbf{y}$, may be written as

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \frac{1}{p(\mathbf{x})} p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \\ &= c(\mathbf{x}) \prod_{i=1}^{i=n} \frac{p(y_i|\mathbf{x}_i)}{p(y_i)} p(\mathbf{y}), \end{aligned} \tag{2.1}$$

[3]To keep the notation simple, we use $p(\cdot)$ to denote both continuous densities and discrete distributions of random variables. The meaning should be clear from the context.

where $c(\mathbf{x}) \equiv \prod_{i=1}^{i=n} p(\mathbf{x}_i)/p(\mathbf{x})$ is a factor not depending on $\mathbf{y}$. The MAP segmentation is then given by

$$\widehat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{L}^n} \left\{ \sum_{i=1}^{n} (\log p(y_i|\mathbf{x}_i) - \log p(y_i)) + \log p(\mathbf{y}) \right\}. \tag{2.2}$$

In the present approach, the densities $p(y_i|\mathbf{x}_i)$ are modeled with the MLR, which corresponds to discriminative model of the discriminative-generative pair for $p(\mathbf{x}_i|y_i)$ Gaussian and $p(y_i)$ multinomial [97], [118]. Notice that $p(y_i)$ can be any distribution, as long as the marginal of $p(\mathbf{y})$ is compatible with such distribution. The estimation of vector of regressors parameterizing the MLR is formulated as in [81], following a semi-supervised approach. To compute the MAP estimate of the regressors, we apply a new Block Gauss-Seidel iterative algorithm. The prior $p(\mathbf{y})$ on the labelings, $\mathbf{y}$, is an MLL Markov random field, which encourages neighboring pixels to have the same label. The MAP labeling/segmentation $\widehat{\mathbf{y}}$ is computed via the α-Expansion algorithm [23], a min-cut based tool to efficiently solve integer optimization problems. All these issues are detailed in the next section.

2.3 Estimation of the logistic regressors

The MLR model is formally given by [16],

$$p(y_i = k|\mathbf{x}_i, \boldsymbol{\omega}) = \frac{\exp(\boldsymbol{\omega}^{(k)}\mathbf{h}(\mathbf{x}_i))}{\sum_{k=1}^{K} \exp(\boldsymbol{\omega}^{(k)}\mathbf{h}(\mathbf{x}_i))}, \tag{2.3}$$

where $\mathbf{h}(\mathbf{x}) \equiv [h_1(\mathbf{x}), \cdots, h_l(\mathbf{x})]^T$ is a vector of l fixed functions of the input, often termed features; $\boldsymbol{\omega}^{(k)}$ is the set of logistic regressors for class k, and $\boldsymbol{\omega} \equiv [\boldsymbol{\omega}^{(1)^T}, \cdots, \boldsymbol{\omega}^{(K-1)^T}]^T$. Given the fact that the density (2.3) does not depend on translations on the regressors $\boldsymbol{\omega}^{(k)}$, we set $\boldsymbol{\omega}^{(K)} = \mathbf{0}$.

Note that the function $\mathbf{h}$ may be linear, *i.e.*, $\mathbf{h}(\mathbf{x}_i) = [1, x_{i,1}, \cdots, x_{i,d}]^T$, where $x_{i,j}$ is the j-th component of $\mathbf{x}_i$ or nonlinear. Kernels [121], *i.e.*, $\mathbf{h}(\mathbf{x}_i) = [1, K_{\mathbf{x},\mathbf{x}_1}, \cdots, K_{\mathbf{x},\mathbf{x}_l}]^T$, where $K_{\mathbf{x}_i,\mathbf{x}_j} = K(\mathbf{x}_i, \mathbf{x}_j)$ and $K(\cdot,\cdot)$ is some symmetric kernel function, are a relevant example of the nonlinear case. Kernels have been largely used because they tend to improve the data separability in the transformed space. In this paper, we use a Gaussian Radial Basis Function (RBF) kernel, $K(\mathbf{x}, \mathbf{z}) \equiv \exp(-\|\mathbf{x} - \mathbf{z}\|^2/(2\rho^2))$, which is widely used in hyperspectral image classification [28]. From now on, d denotes the dimension of $\mathbf{h}(\mathbf{x})$.

In the present problem, learning the class densities amounts to estimating the logistic regressors $\boldsymbol{\omega}$. Since we are assuming a semi-supervised scenario, this estimation is based on a small set of labeled samples, $\mathcal{D}_L \equiv \{(y_1, \mathbf{x}_1), \cdots, (y_L, \mathbf{x}_L)\}$, and a larger set of unlabeled samples,

$\mathcal{X}_U \equiv \{\mathbf{x}_{L+1}, \cdots, \mathbf{x}_{L+U}\}$. Given that our approach is Bayesian, we need to build the posterior density

$$p(\boldsymbol{\omega}|\mathcal{Y}_L, \mathcal{X}_L, \mathcal{X}_U) \propto p(\mathcal{Y}_L|\mathcal{X}_L, \mathcal{X}_U, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathcal{X}_L, \mathcal{X}_U) \tag{2.4}$$

$$= p(\mathcal{Y}_L|\mathcal{X}_L, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathcal{X}_{L+U}), \tag{2.5}$$

where $\mathcal{Y}_L \equiv \{y_1, \cdots, y_L\}$ denotes the set of labels in $\mathcal{D}_L$, $\mathcal{X}_L \equiv \{\mathbf{x}_1, \cdots, \mathbf{x}_L\}$ denotes the set of feature vectors in $\mathcal{D}_L$, and $\mathcal{X}_{L+U}$ stands for $\{\mathcal{X}_L, \mathcal{X}_U\}$. Here, we have used the conditional independence assumption in the right hand side of (2.5).

The MAP estimate of $\boldsymbol{\omega}$ is then given by

$$\widehat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \left\{ l(\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}|\mathcal{X}_{L+U}) \right\}, \tag{2.6}$$

where

$$\begin{aligned} l(\boldsymbol{\omega}) &\equiv \log p(\mathcal{Y}_L|\mathcal{X}_L, \boldsymbol{\omega}) \equiv \log \prod_{i=1}^{L} p(y_i|\mathbf{x}_i, \boldsymbol{\omega}) \\ &\equiv \sum_{i=1}^{L} \left(\mathbf{x}_i^T \boldsymbol{\omega}^{(y_i)} - \log \sum_{j=1}^{K} \exp(\mathbf{x}_i^T \boldsymbol{\omega}^{(j)}) \right) \end{aligned} \tag{2.7}$$

is the log-likelihood function of $\boldsymbol{\omega}$ given the labeled samples $\mathcal{D}_L$ and $p(\boldsymbol{\omega}|\mathcal{X}_{L+U})$ acts as prior on $\boldsymbol{\omega}$. Following the rationale introduced in [81], we adopt the Gaussian prior

$$p(\boldsymbol{\omega}|\boldsymbol{\Gamma}) \propto \exp\left\{ -\frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\Gamma} \boldsymbol{\omega} \right\}, \tag{2.8}$$

where the precision matrix $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\mathcal{X}_{L+U})$ is built in such a way that the density $p(\boldsymbol{\omega}|\boldsymbol{\Gamma})$ promotes vectors $\boldsymbol{\omega}$ leaving "close" labeled and unlabeled features $\mathbf{h}(\mathbf{x})$, for $\mathbf{x} \in \mathcal{X}_{L+U}$, in the same class. The distance between features is defined in terms of a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$, where $\mathcal{V}$ is the set of vertices corresponding to labeled and unlabeled data, $\mathcal{E}$ is a set of edges defined on $\mathcal{V} \times \mathcal{V}$, and $\mathcal{B}$ is a set of weights defined on $\mathcal{E}$. With these definitions in place, the precision matrix writes as

$$\boldsymbol{\Gamma}(\boldsymbol{\lambda}) = \boldsymbol{\Lambda} \otimes (\mathbf{A} + \tau \mathbf{I}),$$

where symbol $\otimes$ denotes the Kronecker product, $\tau > 0$ is a regularization parameter, and

$$\begin{aligned} \boldsymbol{\Lambda} &\equiv \text{diag}(\lambda_1, \cdots, \lambda_{(K-1)}) \\ \mathbf{A} &\equiv \mathbf{X} \boldsymbol{\Delta} \mathbf{X}^T \\ \mathbf{X} &\equiv [\mathbf{h}(\mathbf{x}_1), \cdots, \mathbf{h}(\mathbf{x}_{L+U})] \\ \boldsymbol{\Delta} &\equiv \text{Laplacian of the graph } \mathcal{G}. \end{aligned}$$

Notice that $\mathbf{\Gamma}(\boldsymbol{\lambda})$ is a block diagonal matrix, *i.e.*,

$$\mathbf{\Gamma}(\boldsymbol{\lambda}) = \text{diag}(\lambda_1(\mathbf{A} + \tau\mathbf{I}), \ldots, \lambda_{(K-1)}(\mathbf{A} + \tau\mathbf{I})),$$

where $\text{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_K)$ stands for a block diagonal matrix with diagonal blocks $\mathbf{A}_1, \ldots, \mathbf{A}_K$. In the above definitions, and $\lambda_1, \cdots, \lambda_{(K-1)}$ are non-negative scale factors.

With the definitions above, we have

$$\boldsymbol{\omega}^T \mathbf{\Gamma}(\boldsymbol{\lambda})\, \boldsymbol{\omega} = \sum_{k=1}^{K-1} \lambda_k \left({\boldsymbol{\omega}^{(k)}}^T \mathbf{A} \boldsymbol{\omega}^{(k)} + \tau \|\boldsymbol{\omega}^k\|^2 \right).$$

The quadratic term $\tau\|\boldsymbol{\omega}^k\|^2$ acts as a quadratic regularizer, ensuring that the estimation of $\boldsymbol{\omega}$ is not ill-posed. At the same time, in order to ensure that this quadratic regularizer does not modify the role of matrix $\mathbf{A}$, the value of τ should be much smaller than the largest eigenvalue of $\mathbf{A}$. In order to interpret the role of the quadratic terms ${\boldsymbol{\omega}^{(k)}}^T \mathbf{A} \boldsymbol{\omega}^{(k)}$, let $\mathcal{V} \equiv \{1, \cdots, U+L\}$ and $\mathcal{B} \equiv \{\beta_{ij} \geq 0,\, (i,j) \in \mathcal{E}\}$ denote, respectively, the set of vertices and weights of $\mathcal{G}$. Having in mind the meaning of the Laplacian of a graph, we have

$$\begin{aligned} {\boldsymbol{\omega}^{(k)}}^T \mathbf{A} \boldsymbol{\omega}^{(k)} &= {\boldsymbol{\omega}^{(k)}}^T \mathbf{X}^T \mathbf{\Delta} \mathbf{X} \boldsymbol{\omega}^{(k)} \\ &= \textstyle\sum_{(i,j)\in\mathcal{E}} \beta_{ij} \left[{\boldsymbol{\omega}^{(k)}}^T (\mathbf{h}(\mathbf{x}_i) - \mathbf{h}(\mathbf{x}_j)) \right]^2. \end{aligned}$$

Therefore, the lower values of ${\boldsymbol{\omega}^{(k)}}^T \mathbf{A} \boldsymbol{\omega}^{(k)}$, corresponding to the most probable regressors $\boldsymbol{\omega}^{(k)}$, occur when both features $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same side of the separating hyperplane defined by $\boldsymbol{\omega}^{(k)}$. In this way, the prior acts as a regularizers on $\boldsymbol{\omega}^{(k)}$, promoting those solutions for which the features connected with higher values of weights β_{ij} are given the same label. This implies that the boundaries among the classes tend to be pushed to the regions of low density, with respect to the underlying graph $\mathcal{G}$. In accordance with this rationale, we set in this work

$$\beta_{ij} = e^{-\|\mathbf{h}(\mathbf{x}_i) - \mathbf{h}(\mathbf{x}_j)\|^2}. \tag{2.9}$$

According to a Bayesian point of view, the parameters $\lambda_1, \ldots, \lambda_{(K-1)}$ are random variables and should be integrated out. We assume that they are distributed according to Gamma densities, which are conjugate priors for the inverse of a variances of Gaussian densities [9]. More precisely, we assume they are independent and that

$$\lambda_i \sim \text{Gam}(\alpha, \beta) \quad i = 1, \ldots, K-1, \tag{2.10}$$

where $\text{Gam}(\alpha, \beta)$ stands for a Gamma distribution with shape parameter α and inverse scale parameter β. Noting that λ_i, $i = 1, \ldots, K-1$ are scaling parameters, we set α, β to very small values, thus obtaining a density close to that of Jeffreys prior. We note that the Jeffreys prior, which is non-informative for scale parameters, is obtained by setting to zero the shape and the inverse scale parameters of a Gamma density.

2.3.1 Computing the MAP estimate of the regressors

To compute the MAP estimate of $\boldsymbol{\omega}$, we use an expectation-maximization (EM) algorithm [43], where the scale factors λ_i, for $i = 1 \ldots, K-1$, are the missing variables. The EM algorithm is an iterative procedure that computes, in each iteration, a so-called E-step (for mean value) and the M-step (for maximization). More specifically, at iteration t, these steps are formally given by

E-step:

$$Q(\boldsymbol{\omega}|\boldsymbol{\omega}_t) \equiv E\left[\log p(\boldsymbol{\omega}, \boldsymbol{\lambda}|\mathcal{D}) \,|\, \boldsymbol{\omega}_t\right] \tag{2.11}$$

M-step:

$$\boldsymbol{\omega}_{t+1} \in \arg\max_{\boldsymbol{\omega}} Q(\boldsymbol{\omega}|\boldsymbol{\omega}_t). \tag{2.12}$$

In (2.11), $\mathcal{D} \equiv \{\mathcal{D}_L, \mathcal{X}_U\}$ denotes the set of labeled and unlabeled samples. The most relevant property of the EM algorithm is that the sequence $p(\boldsymbol{\omega}_t|\mathcal{D})$, for $t = 1, 2, \ldots$, is non-decreasing and, under mild assumptions, converges to local optima of the density $p(\boldsymbol{\omega}|\mathcal{D})$.

2.3.2 E-step

From expressions (2.5) and (2.8), we have

$$p(\boldsymbol{\omega}, \boldsymbol{\lambda}|\mathcal{D}) = p(\mathcal{Y}_L|\mathcal{X}_L, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\boldsymbol{\Gamma}(\boldsymbol{\lambda})) p(\boldsymbol{\lambda})\, c^{te}, \tag{2.13}$$

where c^{te} does not depend on $\boldsymbol{\omega}$ and $\boldsymbol{\lambda}$ and $p(\boldsymbol{\lambda}) \equiv \prod_{i=1}^{K-1} p(\lambda_i)$. We have then

$$\begin{aligned} Q(\boldsymbol{\omega}|\boldsymbol{\omega}_t) &= E[\log p(\mathcal{Y}_L|\mathcal{X}_L, \boldsymbol{\omega}) - (1/2)\, \boldsymbol{\omega}^T \boldsymbol{\Gamma}(\boldsymbol{\lambda}) \boldsymbol{\omega} + C|\boldsymbol{\omega}_t] \\ &= \log p(\mathcal{Y}_L|\mathcal{X}_L, \boldsymbol{\omega}) - (1/2)\, \boldsymbol{\omega}^T E[\boldsymbol{\Gamma}(\boldsymbol{\lambda})|\boldsymbol{\omega}_t] \boldsymbol{\omega} + C' \\ &= l(\boldsymbol{\omega}) - (1/2)\, \boldsymbol{\omega}^T \boldsymbol{\Upsilon}(\boldsymbol{\omega}_t) \boldsymbol{\omega} + C', \end{aligned} \tag{2.14}$$

where $l(\boldsymbol{\omega})$ is the log-likelihood function given by (2.7), $\boldsymbol{\Upsilon}(\boldsymbol{\omega}_t) \equiv E[\boldsymbol{\Gamma}(\boldsymbol{\lambda})|\boldsymbol{\omega}_t]$, and C and C' do not depend on $\boldsymbol{\omega}$. Since $\boldsymbol{\Gamma}(\boldsymbol{\lambda})$ is linear on $\boldsymbol{\lambda}$, then $\boldsymbol{\Upsilon}(\boldsymbol{\omega}_t) = \boldsymbol{\Gamma}(E[\boldsymbol{\lambda}|\boldsymbol{\omega}_t])$.

Owing to the use of conjugate Gamma hyper-priors, the expectations $E[\lambda_i|\boldsymbol{\omega}]$ have well-known closed forms [9]. For the present setting, we have

$$\gamma_k \equiv E[\lambda_k|\boldsymbol{\omega}] = (2\alpha + d)[2\beta + (\widehat{\boldsymbol{\omega}}^{(k)})^T(\mathbf{A} + \tau\mathbf{I})\widehat{\boldsymbol{\omega}}^{(k)}]^{-1},$$

for $k = 1, \cdots, K-1$.

2.3.3 M-step

Given the matrix $\boldsymbol{\Upsilon}(\widehat{\boldsymbol{\omega}})$, the M-step amounts to maximize the objective function (2.14), which is a logistic regression problem with a quadratic regularizer. Hereinafter, we adopt the generalized expectation maximization (GEM) [43] approach, which consists in replacing, in the M-step, the objective function $Q(\cdot|\cdot)$ with another one which is simpler to optimize. A necessary condition for GEM still generating a non-decreasing sequence $p(\boldsymbol{\omega}_t|\mathcal{D})$, for $t = 1, 2, \ldots$, is that $Q(\boldsymbol{\omega}_{t+1}|\boldsymbol{\omega}_t) \leq Q(\boldsymbol{\omega}_t|\boldsymbol{\omega}_t)$, for $t = 1, 2, \ldots$ In order to build a simpler objective function, we resort to bound optimization techniques [86], which aim, precisely, at replacing a difficult optimization problem with a series of simpler ones.

Let $\mathbf{g}(\boldsymbol{\omega})$ be the gradient of $l(\boldsymbol{\omega})$ given by

$$\mathbf{g}(\boldsymbol{\omega}) = \sum_{i=1}^{L}(\mathbf{e}_{y_i} - \mathbf{p}_i) \otimes \mathbf{h}(\mathbf{x}_i),$$

where $\mathbf{e}_k$ is the kth column of the identity matrix of size K and

$$\mathbf{p}_i \equiv [p(y = 1|\mathbf{x}_i, \boldsymbol{\omega}), p(y = 2|\mathbf{x}_i, \boldsymbol{\omega}), \ldots, p(y = K|\mathbf{x}_i, \boldsymbol{\omega})]^T. \tag{2.15}$$

Let us define the non-positive definite matrix as

$$\mathbf{B} \equiv -\frac{1}{2}\left[\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{K-1}\right] \otimes \sum_{i=1}^{L}\mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)^T, \tag{2.16}$$

where $\mathbf{1}$ denotes a column vector of 1s and $\mathbf{1}^T$ is the transpose of such column vector. We now define the following quadratic majorizer for function Q resulting from the E-step stated in Eq. (2.14)

$$Q_B(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}}) \equiv l(\widehat{\boldsymbol{\omega}}) + (\boldsymbol{\omega} - \widehat{\boldsymbol{\omega}})^T\mathbf{g}(\widehat{\boldsymbol{\omega}}) + [(\boldsymbol{\omega} - \widehat{\boldsymbol{\omega}})^T\mathbf{B}(\boldsymbol{\omega} - \widehat{\boldsymbol{\omega}}) - \boldsymbol{\omega}^T\boldsymbol{\Gamma}(\widehat{\boldsymbol{\omega}})\boldsymbol{\omega}]/2.$$

Let $\mathbf{H}(\boldsymbol{\omega})$ be the Hessian of $l(\boldsymbol{\omega})$. Matrix $\mathbf{H} - \mathbf{B}$ is semi-positive definite [16], *i.e.,* $\mathbf{H}(\boldsymbol{\omega}) \succeq \mathbf{B}$

for any $\boldsymbol{\omega}$. It is then easy to show that

$$Q(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}}) \geq Q_B(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}})$$

with equality if and only if $\boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}$. Thus, $Q_B(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}})$ is a valid surrogate function for $Q(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}})$. That is, by replacing Q with Q_B in (2.11), the inequality $Q(\boldsymbol{\omega}_{t+1}|\boldsymbol{\omega}_t) \geq Q(\boldsymbol{\omega}_t|\boldsymbol{\omega}_t)$ for $t = 1, 2, \ldots$ still holds, which implies $p(\boldsymbol{\omega}_t|\mathcal{D}) \leq p(\boldsymbol{\omega}_t|\mathcal{D})$, for $t = 1, 2, \ldots$

The maximizer of $Q_B(\boldsymbol{\omega}|\boldsymbol{\omega}_t)$ with respect to $\boldsymbol{\omega}$ is

$$\boldsymbol{\omega}_{t+1} = (\mathbf{B} - \boldsymbol{\Gamma}(\boldsymbol{\omega}_t))^{-1}(\mathbf{B}\boldsymbol{\omega}_t - \mathbf{g}(\boldsymbol{\omega}_t)),$$

which amounts to solving a linear system with $d(K-1)$ unknowns, thus with $O((d(K-1))^3)$ complexity. This complexity may be unbearable, even for middle-sized data sets. To tackle this difficulty, a sequential approach in which the algorithm only maximizes Q_B with respect to one element of $\boldsymbol{\omega}$ at a time is proposed in [81]. Here, the complexity of a complete scanning of all elements of $\boldsymbol{\omega}$ is $O(Kd(L+d))$, much lighter than $O((d(K-1))^3)$. What we have found out, however, is that the convergence rate of this algorithm is too small, a factor that rules out its application in realistic hyperspectral imaging applications.

In order to increase the convergence rate and to handle systems of reasonable size, we implement a Block Gauss-Seidel iterative procedure in which the blocks are the regressors of each class. Thus, in each iteration, we solve $K-1$ systems of dimension d. Furthermore, we have observed that just one iteration before recomputing the precision matrix $\boldsymbol{\Gamma}$ is nearly the best choice. Notice that, even with just one Gauss-Seidel iteration, the algorithm is still a GEM. The improvement in complexity with respect to the exact solution is given by $O((K-1)^2)$, which makes a difference when there are many class labels, as it is indeed the case in most hyperspectral imaging applications.

The pseudo-code for the GEM algorithm to compute the MAP estimate of $\boldsymbol{\omega}$ is shown in Algorithm 2.1, where GEMiters denotes the maximum number of GEM iterations and BSGiters denotes the number of Block Gauss-Seidel iterations. The notation $(\cdot)^{(k)}$ stands for the block column vectors corresponding to regressors $\boldsymbol{\omega}^{(k)}$.

2.4 The Multi-Level logistic spatial prior

In segmenting real world images, it is very likely that neighboring pixels belong to the same class. The exploitation of this (seemingly naive) contextual information improves, in some cases dramatically, the classification performance. In this work, we integrate the contextual

Algorithm 2.1 GEM algorithm to estimate the MLR regressors $\boldsymbol{\omega}$

Require: $\boldsymbol{\omega}_0$, $\mathcal{D}_L$, $\mathcal{X}_U$, α, β, τ, GEMiters, BSGiters
Ensure: $u_{k,l} \equiv [\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathbf{T}}/(K-1)]_{k,l}$
  $\mathbf{R} \equiv \sum_{i=1}^{L} \mathbf{h}(\mathbf{x_i})\mathbf{h}(\mathbf{x_i})^T$, $\mathbf{X} \equiv [\mathbf{h}(\mathbf{x}_1), \cdots, \mathbf{h}(\mathbf{x}_{L+U})]$
  $\mathcal{B} := \mathcal{B}(\mathbf{X})$($*$ build the graph weights according to (2.9) $*$)
  $\Delta := \Delta(\mathcal{B})$ ($*$ Δ is the Laplacian of graph $\mathcal{G}$ $*$)
  $i := 1$
  $\mathbf{A} := \mathbf{X}\Delta\mathbf{X}^T$
  while $i \leq$ GEMiter **or** stopping criterion is not satisfied **do**
    $\lambda_k := (2\alpha + d)[2\beta + (\boldsymbol{\omega}_i^{(k)})^T(\mathbf{A} + \tau\mathbf{I})\boldsymbol{\omega}_i^{(k)}]^{-1}, \;\; k = 1, \ldots, K-1$
    $\mathbf{z} := \mathbf{B}\boldsymbol{\omega}_{i-1} - \mathbf{g}(\boldsymbol{\omega}_{i-1})$
    $\mathbf{C}_{k,l} := u_{k,l}\mathbf{R} - \lambda_l(\mathbf{A} + \tau\mathbf{I})$
    for $j := 1$ to BSGiters **do**
      for $k := 1$ to $K-1$ **do**
        $\boldsymbol{\omega}_{(i)}^{(k)} =$ solution $\{\mathbf{C}_{k,k}\boldsymbol{\omega}^{(k)} = \mathbf{z}^{(k)} - \sum_{l=1, l\neq k}^{K-1} \mathbf{C}_{k,l}\boldsymbol{\omega}_i^{(l)}\}$
      end for
    end for
  end while

information with spectral information by using an isotropic MLL prior to model the image of class labels $\mathbf{y}$. This prior, which belongs to the MRF class, encourages piecewise smooth segmentations and thus promotes solutions in which adjacent pixels are likely to belong the same class. The MLL prior is a generalization of the Ising model [58] and has been widely used in image segmentation problems [92].

According to the Hammersly-Clifford theorem [10], the density associated with a MRF is a Gibbs's distribution [58]. Therefore, the prior model for segmentation has the following structure

$$p(\mathbf{y}) = \frac{1}{Z}\, e^{\left(-\sum_{c\in\mathcal{C}} V_c(\mathbf{y})\right)}, \tag{2.17}$$

where Z is a normalizing constant for the density, the sum in the exponent is over the so-called prior potentials $V_c(\mathbf{y})$ for the set of cliques[4] $\mathcal{C}$ over the image, and

$$-V_c(\mathbf{y}) = \begin{cases} \upsilon_{y_i}, & \text{if } |c| = 1 \text{ (single clique)} \\ \mu_c, & \text{if } |c| > 1 \text{ and } \forall_{i,j\in c} y_i = y_j \\ -\mu_c, & \text{if } |c| > 1 \text{ and } \exists_{i,j\in c} y_i \neq y_j \end{cases} \tag{2.18}$$

where μ_c is a non-negative constant.

The potential function in (2.18) encourages neighbors to have the same label. By varying the set of cliques and the parameters υ_{y_i} and μ_c, the MLL prior offers a great deal of flexibility.

[4] A clique is a single term or either a set of pixels that are neighbors of one another.

For example, the model generates texture-like regions if μ_c depends on c and blob-like regions otherwise [93]. By taking $\mu_c = \frac{1}{2}\mu > 0$, the Eq. (2.17) can be rewritten as

$$p(\mathbf{y}) = \frac{1}{Z}\, e^{\sum_{i\in\mathcal{S}} \upsilon_{y_i} + \mu \sum_{(i,j)\in\mathcal{C}} \delta(y_i - y_j)} \tag{2.19}$$

where $\delta(y)$ is the unit impulse function[5]. This choice gives no preference to any direction. The unary cliques υ_{y_i} are defined by the marginal $p(y_i)$ in the following sense:

$$p(y_i) = \sum_{j=1,\ldots,n, j\neq i} p(y_j).$$

Herein, we assume $p(y_i) = 1/K$, *i.e.*, equiprobable classes. In this case, a simple computation leads to $\upsilon_{y_i} = c^{te}$. Notice that the pairwise interaction terms $\delta(y_i - y_j)$ attach higher probability to equal neighboring labels than the other way around. In this way, the MLL prior promotes piecewise smooth segmentations. The level of smoothness is controlled by parameter μ.

In this paper, we consider only first and second order neighborhoods; *i.e.*, considering that pixels are arranged in a square grid where the distance between horizontal or vertical neighbors is defined to be 1, the cliques corresponding to first and second order neighborhoods are, respectively, $\{(i,j) \in \mathcal{C} \,|\, d(i,j) \leq 1,\, i,j \in \mathcal{S}\}$ and $\{(i,j) \in \mathcal{C} \,|\, d(i,j) \leq \sqrt{2},\, i,j \in \mathcal{S}\}$, where $d(i,j)$ is the distance between pixels $i, j \in \mathcal{S}$.

2.5 Computing the MAP estimate via graph-cuts

Based on the posterior class densities $p(y_i|\mathbf{x}_i)$ and on the MLL prior $p(\mathbf{y})$, assuming equiprobable classes and according to (2.2), the MAP segmentation is finally given by

$$\widehat{\mathbf{y}} \;=\; \arg\min_{\mathbf{y}\in\mathcal{L}^n} \sum_{i\in\mathcal{S}} -\log p(y_i|\widehat{\boldsymbol{\omega}}) - \mu \sum_{i,j\in C} \delta(y_i - y_j), \tag{2.20}$$

where $p(y_i|\widehat{\boldsymbol{\omega}}) \equiv p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$, computed at $\widehat{\boldsymbol{\omega}}$. Minimization of expression (2.20) is a combinatorial optimization problem, involving unary and pairwise interaction terms. The exact solution for $K = 2$ was introduced in [63] by mapping the problem into the computation of a min-cut on a suitable graph. This line of attack was reintroduced in the beginning of this century, and has been intensely researched since then (see, e.g, [4, 22, 23, 79]). As a result of this research, the number integer optimization problems that can now be solved exactly (or with a very good approximation) has increased substantially. A key element in graph-cut based

[5] *i.e.*, $\delta(0) = 1$ and $\delta(y) = 0$, for $y \neq 0$

approaches to integer optimization is the so-called sub-modularity of the pairwise terms: a pairwise term $V(y_i, y_j)$ is said to be submodular (or graph-representable) if $V(y_i, y_i)+V(y_j, y_j) \leq V(y_i, y_j) + V(y_j, y_i)$, for any $y_i, y_j \in \mathcal{L}$. This is the case of our binary term $-\mu\delta(y_i - y_j)$. In this case, the α-Expansion algorithm [23] can be applied. It yields very good approximations to the MAP segmentation problem and is efficient from a computational point of view, being its practical computational complexity $O(n)$.

2.5.1 Semi-supervised algorithm

Let $\mathcal{X}_{\overline{L+U}} \equiv \{\mathbf{x}_{U+1}, \cdots, \mathbf{x}_n\}$ denote the unlabeled set in $\mathbf{x}$. The pseudo-code for the proposed semi-supervised segmentation algorithm with discriminative class learning MLL prior is shown in Algorithm 2.2.

Algorithm 2.2 Semi-supervised segmentation algorithm

Require: $\mathcal{D}_L$, $\mathcal{X}_U$, $\mathcal{X}_{L+U}$, $\mathcal{X}_{\overline{L+U}}$, GEMiters, BSGiters, α, β, τ, m

1: **while** stopping criterion is not satisfied **do**
2: $\widehat{\boldsymbol{\omega}} :=$ GEM($\mathcal{D}_L, \mathcal{X}_U, \alpha, \beta, \tau$, GEMiters, BSGiters)
3: $\widehat{\mathbf{P}} := \widehat{\mathbf{p}}(\mathbf{x}_i, \widehat{\boldsymbol{\omega}})$, $\mathbf{x}_i \in \mathcal{X}_{\overline{L+U}}$
4: ($* \widehat{\mathbf{P}}$ collects the MLR probabilities (2.15) for all feature vectors in $\mathcal{X}_{\overline{L+U}} *$)
5: $\mathcal{X}_{new} := \varphi(\widehat{\mathbf{P}}, m)$
6: ($*\varphi(\widehat{\mathbf{P}}, m)$ selects m unlabeled samples from $\mathcal{X}_{\overline{L+U}}$. See explanation $*$)
7: $\mathcal{X}_{L+U} := \mathcal{X}_{L+U} + \mathcal{X}_{new}$
8: $\mathcal{X}_{\overline{L+U}} := \mathcal{X}_{\overline{L+U}} - \mathcal{X}_{new}$
9: **end while**
10: $\widehat{\mathbf{P}} := \widehat{\mathbf{p}}(\mathbf{x}_i, \widehat{\boldsymbol{\omega}})$, $i \in \mathcal{S}$
11: $\widehat{\mathbf{y}} := \alpha$-Expansion($\widehat{\mathbf{P}}, \mu$, neighborhood)

Lines 2, 10, and 11 of Algorithm 2.2 embody the core of our proposed algorithm. Specifically, line 2 implements the semi-supervised learning of the MLR regressors through the GEM procedure described in Algorithm 2.1. It uses both the labeled and unlabeled samples. Line 10 computes the multinomial probabilities for the complete hyperspectral image. Line 11 computes the MAP segmentation efficiently by applying the α-Expansion graph-cut based algorithm. The neighborhood parameter for the α-Expansion determines the strength of the spatial prior. For illustrative purposes, Figure 2.1 sketches the most relevant components of the proposed segmentation algorithm in a flow chart.

2.5.2 Active selection of unlabeled samples

Lines 3-8 in Algorithm 2.2 implement the procedure for active selection of unlabeled training samples. The objective is to select sets of unlabeled samples, based on the actual results provided by the classifier, that hopefully lead to the best performance gains for the classifier. Contrarily

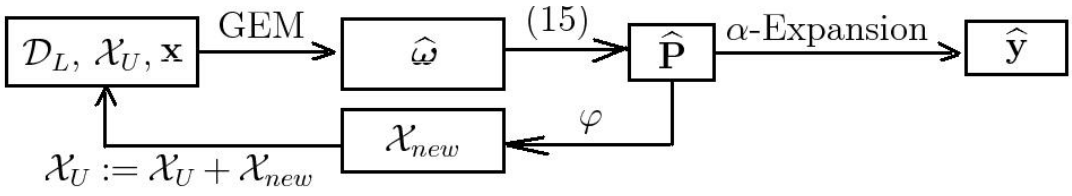


Figure 2.1: Block scheme of Algorithm 2.2.

to active selection of labeled samples [44, 96, 133], the selection on unlabeled samples has not been studied in detail in the literature. These samples are inexpensive and, thus, the question of how many unlabeled samples should be used in hyperspectral data classification arises. In the context of the proposed methodology, however, the complexity of the learning process increases significantly with the incorporation of unlabeled samples, leading to cubic complexity when all samples (labeled and unlabeled) are used for classification. In turn, active selection of a limited number of unlabeled samples allows us to reduce computational complexity significantly and to achieve overall performances that, otherwise, would be only reached with a much larger number of samples.

In this work, we have considered two strategies for the selection criterion implemented by function φ shown in line 5 of Algorithm 2.2, namely, the following:

(i) randomly: in step 5, these m unlabeled samples are randomly selected from $\mathcal{X}_{\overline{L+U}}$.

(ii) maximum entropy: in step 5, these m unlabeled samples have the maximum entropy $\mathbf{HI}(\mathbf{x}_i) = [\widehat{p}^{(1)}, \ldots, \widehat{p}^{(K)}],\ \mathbf{x}_i \in \mathcal{X}_{\overline{L+U}}$, which correspond to the samples near the classifier boundaries.

In the literature, active selection studies for the labeled samples give evidence that, maximum entropy yields very good performance [81, 133]. However, our research is different as we use active selection for the set of unlabeled samples. Nevertheless, we still consider this criterion for our approach. In the next section, we will justify the good behavior of this criterion in the case of active selection of unlabeled samples.

2.5.3 Overall complexity

The complexity of Algorithm 2.2 is dominated by the semi-supervised learning stage of the MLR regressors implemented through the GEM process in Algorithm 2.1, which has computational complexity $O(d^3(K-1))$ as described in Section 2.3.1, and also by the α-Expansion algorithm used to determine the MAP segmentation, which has practical complexity $O(n)$ as described in Section 2.5. Since in most applications $d^3(K-1) > n$, the overall complexity is dominated by that of the GEM process in Algorithm 2.1, which is used to learn the MLR regressors.

As already referred, compared with the semi-supervised algorithm presented in [81], the proposed semi-supervised algorithm is $(K-1)^2$ faster. For a problem with 500 labeled pixels, 224 bands, and 10 classes on a 2.31GHz PC, with only the first 20 iterations, the proposed algorithm took 10.53 seconds, whereas the algorithm in [81] took 106.77 seconds.

2.6 Experimental results

In this section, we evaluate the performance of the proposed algorithm using both simulated and real hyperspectral data sets. The main objective in running experiments with simulated data is the assessment and characterization of the algorithm in a controlled environment, whereas the main objective in running experiments with real data sets is comparing its performance with that reported for state-of-the-art competitors with the same scenes.

This section is organized as follows. Section 2.6.1 reports experiments with simulated data, and contains the following experiments. In Subsection 2.6.1.I, we conduct an evaluation of the impact of the spatial prior on the analysis of simulated data sets. Subsection 2.6.1.II performs an evaluation of the impact of incorporating unlabeled samples to the analysis. Finally, Subsection 2.6.1.III conducts an experimental evaluation of the increase in classification results after including the active selection methodology. On the other hand, Section 2.6.2 evaluates the performance of the proposed algorithm using two real hyperspectral scenes collected by AVIRIS over agricultural fields located at Indian Pines, Indiana [85], and the Valley of Salinas, California [85]. In this section, the algorithm is compared with state-of-the-art competitors.

It should be noted that, in all experiments other than those related with the evaluation of the impact of the spatial prior, we use RBF Kernels $K(\mathbf{x},\mathbf{z}) = \exp(-\|\mathbf{x}-\mathbf{z}\|^2/(2\rho^2))$ to normalize data[6]. The scale parameter of the RBF Kernel is set to $\rho = 0.6$. In our experiments, we use all of the available spectral bands without applying any feature selection strategy. Since we use RBF kernels, the overall complexity only depends on the total number of labeled and unlabeled samples. Thus, the application of feature selection techniques makes no significant differences in this particular scenario. Although this setting is not optimal for all experiments, we have observed that it yields very good results in all experiments. In all cases, the reported values of the overall accuracy (OA) are obtained as the mean values after 10 Monte Carlo runs, with respect to the labeled samples $\mathcal{D}_L$, except for the results over the AVIRIS Salinas dataset, which are obtained with 5 Monte Carlo runs. The labeled samples for each Monte Carlo simulation are obtained by resampling a much larger set of labeled samples. Finally,

[6]The normalization is $\mathbf{x}_i := \frac{\mathbf{x}_i}{(\sqrt{\sum \|\mathbf{x}_i\|^2})}$, for $i = 1,\ldots,n$, where $\mathbf{x}_i$ is a spectral vector and $\mathbf{x}$ is the collection of all image spectral vectors.

it is important to emphasize that in this section we will frequently refer to classification and segmentation results, respectively, when addressing the results provided by the MLR (spectral-based classification) and the complete algorithm (which introduces contextual information to provide a final segmentation).

2.6.1 Experiments with simulated data

In this section, a simulated hyperspectral scene is used to evaluate the proposed semi-supervised algorithm, mainly to analyse the impact of the smoothness parameter μ. For this purpose, we generate images of labels, $y \in \mathcal{L}^n$, sampled from a 128×128 MLL distribution with $\mu = 2$. The feature vectors are simulated according to:

$$\mathbf{x}_{y_i} = \mathbf{m}_{y_i} + \mathbf{n}_{y_i}, \quad i \in \mathcal{S}, \;\; y_i \in \mathcal{L}^n \tag{2.21}$$

where $\mathbf{x}_{y_i}$ denotes the spectral vector, $\mathbf{m}_{y_i}$ denotes a known vector, and $\mathbf{n}_{y_i}$ denotes zero-mean Gaussian noise with covariance $\sigma^2\mathbf{I}$, *i.e.*, $\mathbf{n}_{y_i} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

In Subsection 2.6.1.I, we address a binary classification problem, *i.e.*, $K = 2$, with $\mathbf{x}_{y_i} \in \mathbb{R}^{50}$, $\mathbf{m}_{y_i} = \xi_i\phi$, $\|\boldsymbol{\phi}\| = 1$, and $\xi_i = \pm 1$. The image of class labels $\mathbf{y}$ is shown in Figure 2.2(a), where labels $y_i = 1, 2$ corresponds to $\xi_i = -1, +1$, respectively. In this problem, the theoretical OA, given by $\mathrm{OA}_{opt} \equiv 100\,(1 - \mathrm{P}_e)\%$ and corresponding to the minimal probability of error [49] is

$$\mathrm{P}_e = \frac{1}{2}\mathrm{erfc}\left(\frac{1+\lambda_0}{\sqrt{2}\,\sigma}\right) p_0 + \frac{1}{2}\mathrm{erfc}\left(\frac{1-\lambda_0}{\sqrt{2}\,\sigma}\right) p_1, \tag{2.22}$$

where $\lambda_0 = (\sigma^2/2)\ln(p_0/p_1)$ and p_0 and p_1 are the a priori class labels.

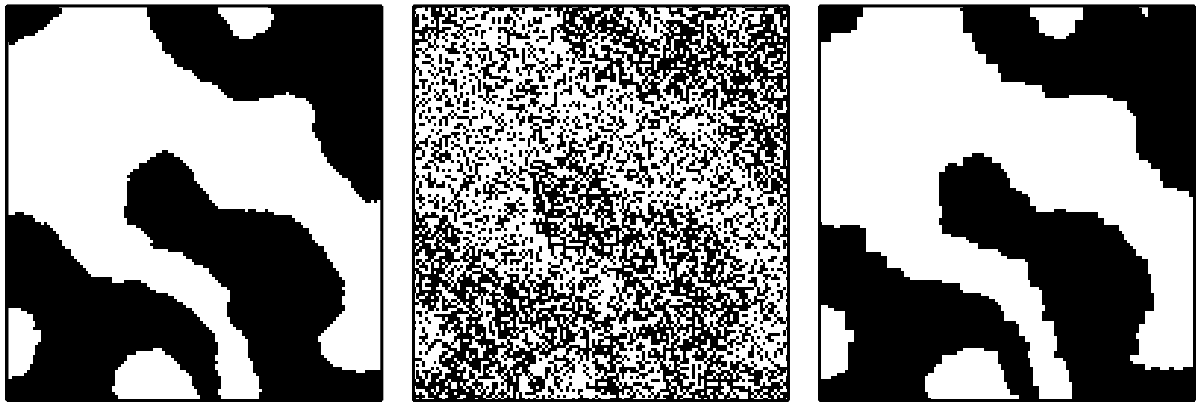
In Subsection 2.6.1.II, the images of class labels are generated with $K = 10$ and $\mathbf{m}_{y_i} = \mathbf{s}_{y_i}$, for $i \in \mathcal{S}$, where $\mathbf{s}_k$, for $k \in \mathcal{L}$, are spectral signatures obtained from the U.S. Geological Survey (USGS) digital spectral library[7]. For a multi-class classification problem, because the probability of error is difficult to compute, we use the error bound

$$\mathrm{P}_e \leq \frac{K-1}{2}\mathrm{erfc}\left(\frac{\mathrm{dist}_{\min}}{2\sigma}\right), \tag{2.23}$$

where $\mathrm{dist}_{\min}$ denotes the minimum distance between any point of mean vectors, *i.e.*, $\mathrm{dist}_{\min} = \min_{i \neq j}\|\mathbf{m}_{y_i} - \mathbf{m}_{y_j}\|$, for any $y_i, y_j \in \mathcal{L}$. This is the so-called union bound [13], which is widely used in multi-class classification problems.

Finally, in Subsection 2.6.1.III we use the same experimental setting as in Subsection 2.6.1.I except for the number of spectral band, which is set to 200, *i.e.*, $\mathbf{x}_{y_i} \in \mathbb{R}^{200}$.

[7]The USGS library of spectral signatures is available online: http://speclab.cr.usgs.gov



(a) (b) (c)

Figure 2.2: Classification and segmentation results obtained after applying the proposed method on a simulated hyperspectral scene representing a binary classification problem. (a) Ground-truth class labels. (b) Classification result (OA=66.94%, with $\text{OA}_{opt} = 75.95\%$). (c) Segmentation result(OA=96.41%).

I. Impact of including a spatial prior

In this example, we use a linear kernel in the characterization of the simulated hyperspectral scene because it yields the correct discriminative density for the Gaussian observations with equal covariance matrix. The number of unlabeled samples is set to zero in this experiment, mainly because our focus is to analyze the effect of the spatial prior independently of other considerations. Figure 2.3 (a) illustrates the OA results as a function of the smoothness parameter μ. It should be noted that the segmentation performance is almost insensitive to μ with $\mu \geq 1$ for the considered problem. In the following experiments, we empirically set $\mu = 1$. Again, although this setting might not be optimal, it leads to good and stable results in our experiments.

On the other hand, Figure 2.3(b-d) presents the OA results with 5, 50 and 500 labeled samples per class, respectively, as a function of the noise standard deviation σ. As shown in the plots, it can be observed that the classification OA approaches the optimal value OA_{opt} as the number of labeled samples is increased, but it is also clear that the number of labeled samples needs to be relatively high in order to obtain classification accuracies which are close to optimal. In turn, it can also be observed in Figure 2.3 that the inclusion of the spatial prior provides much higher segmentation accuracies than those reported for the classification stage (superior in all cases to the values of OA_{opt}). Further, the sensitivity of these results to the amount of noise in the simulated hyperspectral image can be compensated by increasing the number of labeled samples, but accurate values of segmentation OA can be obtained using very few labeled samples, in particular, when the amount of simulated noise is not very high. This experiment confirms our introspection that the inclusion of a spatial prior can significantly improve the classification

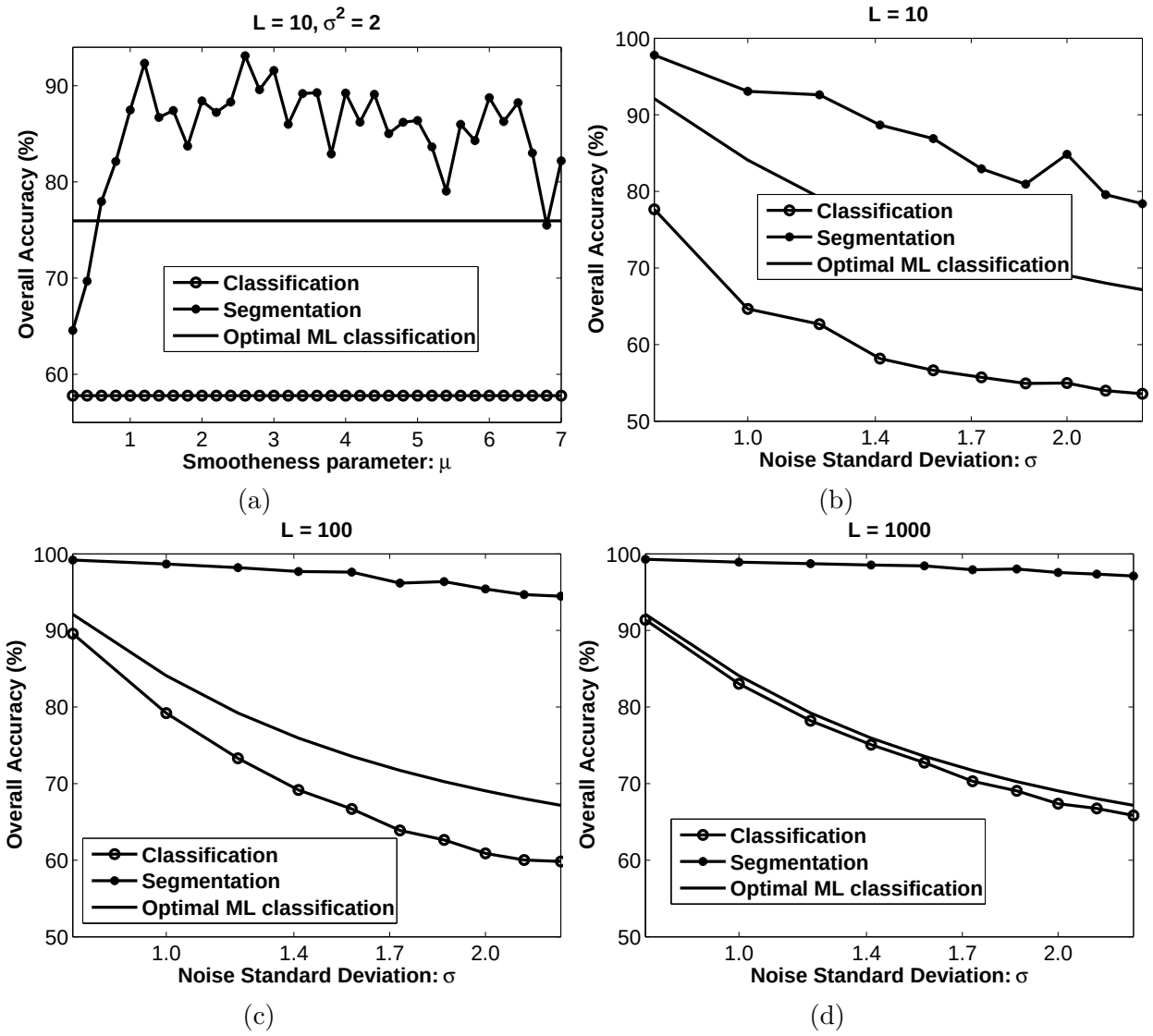
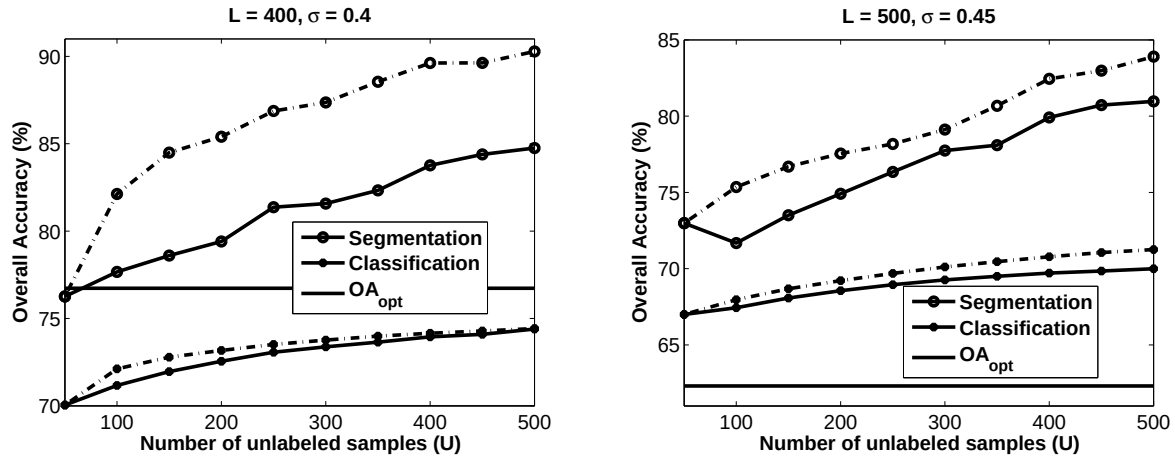


Figure 2.3: (a), OA results as a function of the spatial prior parameter μ with $L = 10, \sigma^2 = 2$. (b), (c) and (d), OA results as a function of the standard deviation σ of the noise introduced in the simulated hyperspectral image, considering different numbers of labeled training samples.



(a) (b)

Figure 2.4: OA results as a function of the number of unlabeled samples. (a) Analysis scenario based on a fixed number of $L = 400$ (40 labeled training samples per class) and $\sigma = 0.4$. (b) Analysis scenario based on a fixed number of $L = 500$ (50 labeled training samples per class) and $\sigma = 0.45$. Solid and dash-dot lines represent random selection and maximum entropy-based active selection, respectively.

results provided by using only spectral information. For illustrative purposes, Figs. 2.2(b) and 2.2(c) show the classification and segmentation maps respectively obtained with $\sigma^2 = 2$ and $L = 100$. In this example, the increase in OA introduced by incorporating the spatial prior with regards to the optimal classification that can be achieved ($\mathrm{OA}_{opt} = 75.95\%$) is clearly noticeable (about 20.46%), thus revealing the importance of including the spatial prior after classification.

II. Impact of incorporating unlabeled samples

In this subsection, we analyze the impact of including unlabeled samples via an active selection strategy in the analysis of simulated hyperspectral data. Specifically, we consider two selection strategies for unlabeled samples: (i) random, and (ii) maximum entropy-based. The latter corresponds to selecting unlabeled samples close to the boundaries between regions in feature space. Figure 2.4 shows the OA results obtained for the proposed algorithm as a function of the number of unlabeled samples for two different analysis scenarios: (a) fixed number of labeled training samples, $L = 400$ (40 per class) and noise standard deviation $\sigma = 0.4$, and (b) fixed $L = 500$ (50 per class) and $\sigma = 0.45$. The theoretical OA, termed as $\mathrm{OA}_{opt} \equiv 100\,(1 - \mathrm{P}_e)\%$, where P_e denotes the union bound in this problem, is also plotted. After analyzing the results reported in Figure 2.4, the following general observations can be made:

- The inclusion of a spatial prior improves the classification OA.
- The inclusion of unlabeled samples improves the segmentation OA by roughly 15% in Figure 2.4(a) and in approximately 10% in Figure 2.4 (b). This effect is observed for all

Table 2.1: OA (%) as a function of the number of unlabeled samples in the toy example illustrated in Figure 2.5(b).

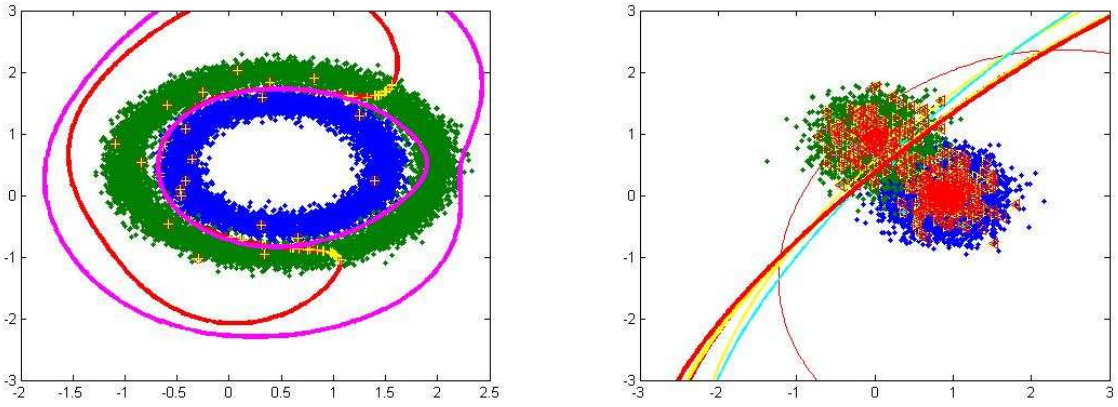
U	0	50	100	150	200	250	300	350	400	450
OA	55.78	86.19	89.29	87.30	88.17	87.73	89.45	90.13	90.45	91.05

considered numbers of unlabeled samples.

- Finally, it is clear from Figure 2.4 that maximum entropy-based active selection performs uniformly better than random selection in terms of OAs.

III. Impact of the considered active selection approach

The main objective of this subsection is to provide an informal justification about why the proposed method for maximum entropy-based active selection of unlabeled samples performs accurately in experiments. Figure 2.5, with 20 labeled samples (10 per class), illustrates the improvements in the separation boundaries established by our proposed classifier as the number of unlabeled samples increases using a toy example. In Figure 2.5(a), in which the noise standard deviation is set to $\sigma = 0.1$, red circles denote the labeled samples. The red line is the classifier boundary defined without unlabeled samples. An OA of 79.32% was obtained in this case. The yellow plus signs (a total of $U = 50$) represent the unlabeled samples. Since we have selected the unlabeled samples with maximum entropy, and the entropy of a sample increases as it approaches the boundary, the selected unlabeled samples are over the contour and located in the area of higher density. The inclusion of these samples have pushed the contour outwards, thus ensuring that all of them stay in the same classification region. Of course, the movement of the boundary in the opposite direction would have also left all the unlabeled samples in the same side of the boundary but would have decreased too much the likelihood term associated with the labeled samples. In this example, the final OA after including unlabeled samples is 98.6%. A similar phenomenon is observed in Figure 2.5(b), in which $\sigma = 0.3$ is considered. For illustrative purposes, Table 2.1 shows the OA results as a function of the number of unlabeled samples for the example reported in Figure 2.5(b). Each column of Table 2.1 corresponds to a different type of color/thickness in 2.5(b), from the thin red line to the thick red line. It is clear that, as the number of unlabeled samples increases, the definition of the separating boundary improves along with the overall performance of the classifier.



(a) (b)

Figure 2.5: Changes in the boundary by the proposed classifier in a binary classification problem as the number of unlabeled samples (selected using a maximum entropy-based criterion) is increased.

2.6.2 Experiments with real hyperspectral data

In order to further evaluate and compare the proposed algorithm with other state-of-the-art techniques for classification and segmentation, in this section we use two real hyperspectral data sets collected by the AVIRIS instrument operated by NASA/JPL:

- The first data set used in experiments was collected over the Valley of Salinas, in 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 meters. The data includes vegetables, bare soils and vineyard fields. The upper-leftmost part of Figure 2.6 shows the entire scene (with overlaid ground-truth areas) and a sub-scene of the dataset (called hereinafter Salinas A), outlined by a red rectangle. The Salinas A sub-scene comprises 83×86 pixels and is known to represent a difficult classification scenario with highly mixed pixels [108], in which the lettuce fields can be found at different weeks since planting. The upper-rightmost part of Figure 2.6 shows the available ground-truth regions for the scene, and the bottom part of Figure 2.6 shows some photographs taken in the field for the different agricultural fields at the time of data collection.

- The second data set used in experiments is the well-known AVIRIS Indian Pines scene, collected over Northwestern Indiana in June of 1992 [85]. This scene, with a size of 145×145 pixels, was acquired over a mixed agricultural/forest area, early in the growing season. The scene comprises 224 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

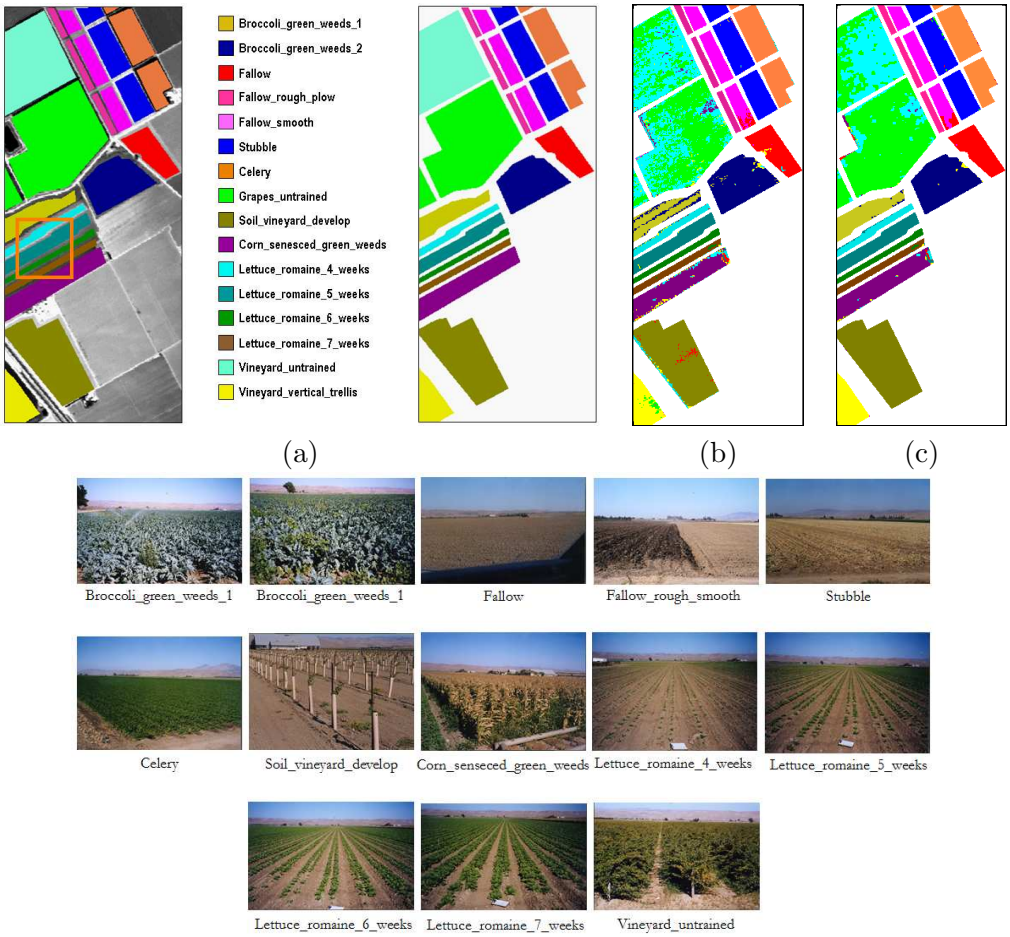


Figure 2.6: AVIRIS Salinas data set along with the classification maps by using $L = 128$, $U = 256$. Upper part: (a), right: original image at 488 nm wavelength with a red rectangle indicating a sub-scene called Salinas A; left, ground truth map containing 16 mutually exclusive land-cover classes. (b) Classification map (OA = 82.55%). (c) Segmentation map (OA = 91.14%). Bottom part: Photographs taken at the site during data collection.

pixel. For illustrative purposes, Figure 2.7(a) shows the ground-truth map available for the scene, displayed in the form of a class assignment for each labeled pixel, with 16 mutually exclusive ground-truth classes. These data, including ground-truth information, are available online[8], a fact which has made this scene a widely used benchmark for testing the accuracy of hyperspectral data classification and segmentation algorithms.

I. Experiments with the full AVIRIS Salinas data set

Table 2.2 reports the segmentation and classification scores achieved for the proposed method with the full AVIRIS Salinas data set, in which the accuracy results are displayed for different numbers of labeled samples (ranging from 5 to 15 per class) and considering also unlabeled samples in a range from $U = 0$ (no unlabeled samples) to $U = 2 \times L$. As shown in Table 2.2,

[8]http://cobweb.ecn.purdue.edu/ biehl/MultiSpec/

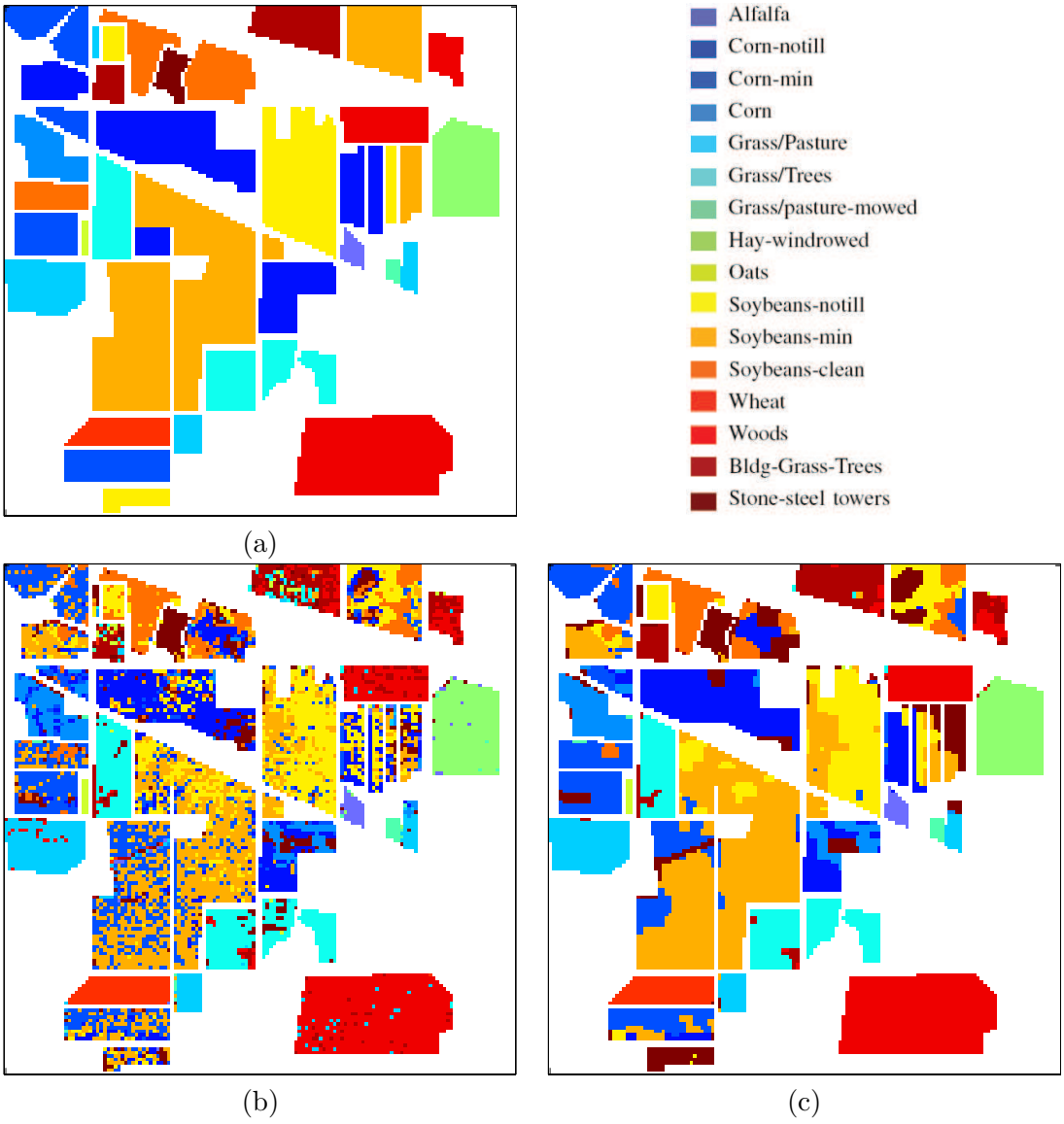


Figure 2.7: AVIRIS Indian Pines scene along with the classification and segmentation maps by using $L = 160$, $U = 288$. (a) Ground truth-map containing 16 mutually exclusive land-cover classes. (b) Classification map (OA = 62.98%). (c) Segmentation map (OA = 74.98%).

Table 2.2: Classification (in the parentheses) and segmentation OAs [%] achieved after applying the proposed algorithm to the full AVIRIS Salinas data set using different numbers of labeled training samples (L). The number of unlabeled samples U is set to $U = 0, L$ and $2 \times L$. Each value of OA reported in the table was obtained with 5 Monte Carlo runs.

	Number of total labeled samples for all classes (L)				
U	80	128	160	192	240
0	86.74 (80.75)	88.94 (81.97)	91.30 (84.47)	92.22 (84.63)	93.87 (85.85)
L	87.20 (80.98)	89.54 (82.39)	92.31 (84.85)	92.42 (84.81)	94.70 (86.21)
$2L$	87.21 (81.14)	89.61 (82.40)	92.93 (85.07)	92.85 (84.84)	95.13 (86.49)

the proposed algorithm obtains very good OAs with limited training samples. Specifically, with only 240 labeled pixels (15 per class), the OA obtained is 93.87% ($U = 0$), 94.70% ($U = L$) and 95.13% ($U = 2 \times L$), which are better than the best result reported in [109] for a set of SVM-based classifiers applied to the same scene with a comparatively much higher number of training samples. Specifically, the SVM classifier in [109] was trained with 2% of the available ground-truth pixels, which means a total of around 1040 labeled samples (about 65 per class). The results reported in this work are only slightly lower than those reported in [108] using a multi-layer perceptron (MLP) neural network classifier, trained with 2% of the available ground-truth pixels, and with multi-dimensional morphological feature extraction prior to classification (the maximum OA reported in [108] for the full AVIRIS Salinas scene was 95.27%, but this result again used a comparatively much higher number of training samples).

On the other hand, it can also be seen from Table 2.2 that the inclusion of a spatial prior significantly improves the results obtained by using the spectral information only (approximately in the order of 6% increase in OA). Furthermore, the inclusion of unlabeled samples in the proposed approach increases the OA in approximately 1% or 2% with regards to the case in which only labeled samples are used. The above results confirm our introspection (already reported in the simulated data experiments) that the proposed approach can greatly benefit from the inclusion of a spatial prior and unlabeled samples in order to increase the already good classification accuracies obtained using the spectral information only. Figure 2.6 (b) and (c) plot the classification and segmentation maps. Effective results can be seen in these maps.

II. Experiments with the AVIRIS Salinas A Sub-Scene

In this experiment, we use a sub-scene of Salinas dataset, which comprises 83×86 pixels and 6 classes. As mentioned above, this sub-scene is known to represent a challenging classification scenario due to the similarity of the different lettuce classes comprised by the sub-scene, which

Table 2.3: Segmentation OAs [%] achieved after applying the proposed algorithm to the AVIRIS Salinas A sub-scene using different numbers of labeled training samples (L). The number of unlabeled samples U is set in a range between $U = 0$ and $U = 5 \times L$. The classification results obtained by the proposed method without the spatial prior are also reported. Each value of OA reported in the table was obtained with 10 Monte Carlo runs.

	L			
U	18	30	48	60
0	93.64	97.76	98.00	99.68
$2L$	95.71	98.45	98.76	99.68
$3L$	95.52	98.71	99.40	99.58
$4L$	96.70	99.28	99.70	99.52
$5L$	96.74	99.66	99.62	99.70
Class.(U=5L)	90.86	95.01	96.74	97.47

are at different weeks since planting and hence have similar spectral features only distinguished by the fraction of lettuce covering the soil in each of the 3.7 meter pixels of the scene. Table 2.3 reports the segmentation (with spatial prior) scores achieved for the proposed method with the AVIRIS Salinas A sub-scene, in which the accuracy results are displayed for different numbers of labeled samples (ranging from 3 to 10 per class) and considering also unlabeled samples in a range from $U = 0$ (no unlabeled samples) to $U = 5 \times L$. The classification results (obtained without using the spatial prior and for $U = 5L$) are also displayed in Table 2.3. As shown in Table 2.3, the proposed algorithm achieved a segmentation OA of up to 99.28% for $U = 4 \times L$ and only 5 labeled samples per class (30 labeled samples in total). This represents an increase of approximately 4.27% OA with respect to the same configuration for the classifier but without using the spatial prior. These results are superior to those reported in [109] and [108] for the classes included in the AVIRIS Salinas A sub-scene using an SVM-based classifier and an MLP-based classifier with multi-dimensional morphological feature extraction, respectively.

III. Experiments with the AVIRIS Indian Pines data set

Table 2.4 reports the segmentation and classification scores achieved for the proposed method with the AVIRIS Indian Pines data set, in which the accuracy results are displayed for different numbers of labeled samples (ranging from 5 to 15 per class) and considering also unlabeled samples in a range from $U = 0$ (no unlabeled samples) to $U = 32 \times k$, with $k = 0, 1, \cdots, 9$. As in previous experiments, the number of labeled samples in Table 2.4 represents the total number of samples selected across the different classes, with approximately the same amount of labeled

samples selected for each class. After a detailed analysis of the experimental results reported on Table 2.4, it is clear that the proposed segmentation method (with spatial prior) provides competitive results for a limited number of labeled samples, outperforming the same classifier without spatial prior in all cases by a significant increase in OA (the increase is always in the order of 10% or higher).

Further, the use of unlabeled samples significantly increases the OA scores reported for the proposed segmentation algorithm. Just as an example, if we assume that 8 labeled samples are used per class, increasing the number of unlabeled samples from 0 to 288 results in an OA increase of approximately 5%, indicating that the proposed approach can greatly benefit not only from the inclusion of a spatial prior, but also from the incorporation of an active learning strategy in order to provide results which are competitive with other results reported in the literature with the same scene. For instance, the proposed algorithm yields better results in terms of OA than the semi-supervised cluster SVMs introduced in [132]. Specifically, when 128 labeled samples (8 samples per class) are used by our proposed method, the OA of the proposed approach is 69.79% ($U = 288$, obtained by active selection), which is comparable to the best result 69.82% reported in [132] (using 519 labeled samples). For illustrative purposes, Figs. 2.7(b) and 2.7(c) show the classification and segmentation maps, respectively. These figures indicate effective results without severe block artifacts. Notice that the results plotted in Figure 2.6 and Figure 2.7 are obtained with just 8 and 10 samples per class, respectively. To give an idea of the quality of this result, we note that the recent semi-supervised technique [132] takes, approximately, 2 times more training samples to achieve a comparable performance, if we take into account only classification results, and 4 times more, if we use spatial information (see Table 2.4).

At this point, we want to call attention for the "good" performance of the proposed algorithm, including the active selection procedure, in the four small size classes, namely "Alfalfa (54 samples)", "Grass/pasture-mowed (26 samples)", "Oats (20 samples)", and "Stone-steel towers (95 samples)". Without going into deep details, this performance is essentially a consequence of having decent estimates for the regressors $\boldsymbol{\omega}$ given by (2.6), condition without which the active selection would fail to provide good results [96].

2.7 Conclusions and future lines

In this paper, we have introduced a new semi-supervised classification/segmentation approach for remotely sensed hyperspectral data interpretation. Unlabeled training samples (selected by means of an active selection strategy based on the entropy of the samples) are used to

Table 2.4: Classification (in parentheses) and segmentation OAs [%] achieved after applying the proposed algorithm to the full AVIRIS Indian Pines data set using different numbers of labeled training samples (L). The number of unlabeled samples U is set in a range between $U = 0$ and $U = 32 \times k$, with $k = 0, 1, \cdots, 9$. The classification results obtained by the proposed method without the spatial prior are also reported. Each value of OA reported in the table was obtained with 10 Monte Carlo runs.

	Number of total labeled samples for all classes (L)				
U	80	128	160	192	240
0	59.09 (52.94)	64.92 (58.65)	70.85 (63.19)	73.88 (66.51)	78.92 (69.09)
32	61.32 (53.07)	65.34 (58.60)	75.60 (63.44)	79.78 (66.44)	76.52 (68.83)
64	59.32 (53.02)	67.47 (58.32)	72.48 (63.33)	75.79 (66.31)	77.47 (68.51)
96	60.37 (52.85)	67.05 (58.25)	74.43 (63.27)	79.11 (66.23)	79.85 (68.42)
128	61.47 (52.87)	67.26 (57.98)	73.92 (63.11)	76.01 (66.15)	75.63 (68.30)
160	60.71 (52.78)	72.14 (57.98)	73.37 (63.01)	78.27 (66.06)	79.10 (68.32)
192	60.40 (52.77)	69.85 (57.96)	73.53 (62.91)	76.83 (65.96)	79.10 (68.22)
224	61.11 (52.72)	67.18 (57.93)	72.14 (62.91)	77.48 (65.99)	78.01 (68.16)
256	61.59 (52.74)	71.33 (57.85)	74.42 (62.82)	73.92 (65.94)	78.15 (68.08)
288	60.71 (52.65)	69.79 (57.94)	73.02 (62.82)	77.16 (65.84)	79.90 (68.04)

improve the estimation of the class distributions. By adopting a spatial multi-level logistic prior and computing the maximum a posteriori segmentation with the α-expansion graph-cut based algorithm, it has been observed that the overall segmentation accuracy achieved by our proposed method in the analysis of simulated and real hyperspectral scenes collected by the AVIRIS imaging spectrometer improves significantly with respect to the classification results proposed by the same algorithm using only the learnt class distributions in spectral space. This demonstrates the importance of considering not only spectral but also spatial information in remotely sensed hyperspectral data interpretation. The obtained results also suggest the robustness of the method to analysis scenarios in which limited labeled training samples are available *a priori*. In this case, the proposed method resorts to intelligent mechanisms for automatic selection of unlabeled training samples, thus taking advantage of an active learning strategy in order to enhance the segmentation results. A comparison of the proposed method with other state-of-the-art classifiers in the considered (highly representative) hyperspectral scenes indicates that the proposed method is very competitive in terms of the (good) overall accuracies obtained, and the (limited) number of training samples (both labeled and unlabeled) required to achieve such accuracies. Further work will be directed towards testing the proposed segmentation approach in different analysis scenarios dominated by the limited availability of training samples *a priori*.

Chapter 3

Hyperspectral Image Segmentation Using a New Bayesian Approach with Active Learning

Abstract – This paper introduces a new supervised Bayesian approach to hyperspectral image segmentation with active learning, which consists of two main steps. First, we use a multinomial logistic regression (MLR) model to learn the class posterior probability distributions. This is done by using a recently introduced logistic regression via splitting and augmented Lagrangian (LORSAL) algorithm. Second, we use the information acquired in the previous step to segment the hyperspectral image using a multi-level logistic prior that encodes the spatial information. In order to reduce the cost of acquiring large training sets, active learning is performed based on the MLR posterior probabilities. Another contribution of this work is the introduction of a new active sampling approach, called modified breaking ties (MBT), which provides an unbiased sampling. Further, we have implemented our proposed method in an efficient way. For instance, in order to obtain the time-consuming maximum *a posteriori* segmentation, we use the α-Expansion min-cut based integer optimization algorithm. The state-of-the-art performance of the proposed approach is illustrated using both simulated and real hyperspectral data sets in a number of experimental comparisons with recently introduced hyperspectral image analysis methods.

Index Terms – Hyperspectral image segmentation, sparse multinomial logistic regression, ill-posed problems, graph cuts, integer optimization, mutual information, active learning.

3.1 Introduction

With the recent developments in remote sensing instruments, hyperspectral images are now widely used in different application domains [107]. The special characteristics of hyperspectral data sets bring difficult processing problems. Obstacles, such as the Hughes phenomenon [68],

come out as the data dimensionality increases. These difficulties have fostered the development of new classification methods, which are able to deal with ill-posed classification problems. For instance, several machine learning techniques are applied to extract relevant information from hyperspectral data sets [19, 30, 36]. However, although many contributions have been made to this area, the difficulty in learning high dimensional densities from a limited number of training samples (an ill-posed problem) is still an active area of research.

Discriminative approaches, which learn the class distributions in high dimensional spaces by inferring the boundaries between classes in feature space [13, 102, 134], tackle effectively the above mentioned difficulties. Specifically, support vector machines (SVMs) [121] are among the state-of-the-art discriminative techniques that can be applied to solve ill-posed classification problems. Due to their ability to deal with large input spaces efficiently and to produce sparse solutions, SVMs have been used successfully for supervised and semi-supervised classification of hyperspectral data using limited training samples [25, 28, 34–37, 53, 107]. On the other hand, multinomial logistic regression (MLR) [16] is an alternative approach to deal with ill-posed problems, which has the advantage of learning the class probability distributions themselves. This is crucial in the image segmentation step. As a discriminative classifier, MLR models directly the posterior densities instead of the joint probability distributions. The distinguishing features of discriminative classifiers have been reported in the literature before [13, 102, 118]. For instance, effective sparse MLR (SMLR) methods are available in the literature [80]. These ideas have been applied to hyperspectral image classification [18, 19, 91] yielding good performance.

Another well-known difficulty in supervised hyperspectral image classification is the limited availability of training data, which are difficult to obtain in practice as a matter of cost and time. In order to effectively work with limited training samples, several methodologies have been proposed, including feature extraction methods such as principal component analysis (PCA), linear discriminant analysis (LDA), discriminant analysis feature extraction (DAFE), multiple classifiers and decision fusion [112], among many others [107]. Active learning, which is another active research topic, has been widely studied in the literature [40, 44, 76, 81, 96, 113, 133]. These studies are based on different principles, such as the evaluation of the disagreement between a committee of classifiers [133], the use of hierarchical classification frameworks [76, 113], unbiased query by bagging [40], or the exploitation of a local proximity-based data regularization framework [44].

In this work, we use active learning to construct small training sets with high training utility, with the ultimate goal of systematically achieving noticeable improvements in classification results with regards to those found by randomly selected training sets of the same size. Since

active learning is intrinsically biased sampling, an issue to be investigated in our experiments is whether the considered classifier (in this work, the MLR) would be able to cope with the class imbalance problem that might be inferred during the active learning strategy. Another trend to improve classification accuracy is to integrate spatial contextual information with spectral information for hyperspectral data interpretation [19, 53, 107, 130]. These methods exploit, in a way or another, the continuity (in probability sense) of neighboring labels. In other words, it is very likely that, in a hyperspectral image, two neighboring pixels have the same label.

In this chapter, we introduce a new supervised Bayesian segmentation approach which exploits both the spectral and spatial information in the interpretation of remotely sensed hyperspectral data sets. The algorithm implements two main steps: (a) learning stage, using the multinomial logistic regression via variable splitting and augmented Lagrangian (LORSAL)[12] algorithm to infer the class distributions; (b) segmentation stage, which infers the labels from a posterior distribution built on the learnt class distributions and on a multi-level logistic (MLL) prior [93]. The computation of the maximum *a posteriori* (MAP) segmentation amounts at maximizing the posterior distribution of class labels. This is a hard integer optimization problem, which we solve by using the powerful graph-cut based α-Expansion algorithm [22]. It yields exact solutions in the binary case and very good approximations when there are more than two classes. Furthermore, we aim at significantly exploiting the efficiency of the labeled samples by means of active learning, thus reducing the size of the required training set and taking full advantage of the MLR posterior probabilities. In this work, different strategies are used to implement active learning in addition to random sampling (RS): (a) the mutual information (MI) between the MLR regressors and the class labels[81, 96]; (b) a criterion called breaking ties (BT) [95]; and (c) our proposed version called modified breaking ties (MBT), which is also intended to guarantee unbiased samplings among the classes.

The remainder of the chapter is organized as follows. Section 3.2 formulates the hyperspectral image segmentation problem. Section 3.3 describes the proposed approach. Section 3.4 presents the active learning algorithms considered in this work. Section 3.5 reports segmentation results based on both simulated and real hyperspectral datasets in several ill-posed scenarios. Comparisons with state-of-the-art algorithms are also included and thoroughly described in this section. Finally, Section 4.5 concludes with a few remarks and hints at plausible future research lines.

3.2 problem formulation

Let $\mathcal{S} \equiv \{1, \ldots, n\}$ denote a set of integers indexing the n pixels of a hyperspectral image; let $\mathcal{L} \equiv \{1, \ldots, K\}$ be a set of K labels; let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ denote an image of d-dimensional feature vectors; let $\mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{L}^n$ be an image of labels; and let $\mathcal{D}_L \equiv \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_L, y_L)\} \in \left(\mathbb{R}^d \times \mathcal{L}\right)^L$ be a training set where L denotes the total number of available labeled samples. With the above definitions in place, the goal of classification is to assign a label $y_i \in \mathcal{L}$ to each pixel $i \in \mathcal{S}$, based on the vector $\mathbf{x}_i$, resulting in an image of class labels $\mathbf{y}$. We call this assignment a *labeling*. On the other hand, the goal of segmentation is to compute, based on the observed image $\mathbf{x}$, a partition $\mathcal{S} = \cup_i \mathcal{S}_i$ of the set $\mathcal{S}$ such that the pixels in each element of the partition share some common properties (i.e., they represent the same type of land cover). Notice that, given a labeling $\mathbf{y}$, the collection $\mathcal{S}_k = \{i \in \mathcal{S} \,|\, y_i = k\}$ for $k \in \mathcal{L}$, is a partition of $\mathcal{S}$. Also, given the segmentation $\mathcal{S}_k$ for $k \in \mathcal{L}$, the image $\{y_i \,|\, y_i = k \text{ if } i \in \mathcal{S}_k,\ i \in \mathcal{S}\}$ is a labeling. Therefore, we can assume that there is a one-to-one relationship between labelings and segmentations. Nevertheless, in this paper we will refer to the term *classification* when there is no spatial information involved in the processing stage, while we will refer to *segmentation* when the spatial prior is being considered.

In a Bayesian framework, inference is often carried out by maximizing the posterior distribution:

$$p(\mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y})p(\mathbf{y}), \tag{3.1}$$

where $p(\mathbf{x}|\mathbf{y})$ is the likelihood function (*i.e.*, the probability of the feature image given the labels) and $p(\mathbf{y})$ is the prior over the labels in $\mathbf{y}$. Assuming conditional independency of the features given the labels, *i.e.*, $p(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{i=n} p(\mathbf{x}_i|y_i)$, the posterior $p(\mathbf{y}|\mathbf{x})$ may be written as a function of $\mathbf{y}$ as follows:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \frac{1}{p(\mathbf{x})} p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \\ &= \alpha(\mathbf{x}) \prod_{i=1}^{i=n} \frac{p(y_i|\mathbf{x}_i)}{p(y_i)} p(\mathbf{y}), \end{aligned} \tag{3.2}$$

where $\alpha(\mathbf{x}) \equiv \prod_{i=1}^{i=n} p(\mathbf{x}_i)/p(\mathbf{x})$ is a factor not depending on $\mathbf{y}$. The MAP segmentation is then given by:

$$\widehat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{L}^n} \left\{ \sum_{i=1}^{n} \left(\log p(y_i|\mathbf{x}_i) - \log p(y_i)\right) + \log p(\mathbf{y}) \right\}. \tag{3.3}$$

In the present approach, the densities $p(y_i|\mathbf{x}_i)$ are modeled as MLRs [16], whose regressors are learnt via the LORSAL algorithm [12]. As prior $p(\mathbf{y})$ on the labelings, $\mathbf{y}$, we adopt an MLL Markov random field (MRF) [93], which encourages neighboring pixels to have the same label. The MAP labeling/segmentation $\widehat{\mathbf{y}}$ is computed via the α-Expansion algorithm [23], a

min-cut based tool to efficiently solve a class of integer optimization problems of which the MAP segmentation in Eq. (3.3) is an example.

3.3 Proposed approach

As mentioned in the previous section, in this work we model the posterior densities $p(y_i|\mathbf{x}_i)$ using a MLR, which is formally given by [16]:

$$p(y_i = k|\mathbf{x}_i, \boldsymbol{\omega}) \equiv \frac{\exp(\boldsymbol{\omega}^{(k)}\mathbf{h}(\mathbf{x}_i))}{\sum_{k=1}^{K}\exp(\boldsymbol{\omega}^{(k)}\mathbf{h}(\mathbf{x}_i))}, \tag{3.4}$$

where $\mathbf{h}(\mathbf{x}) \equiv [h_1(\mathbf{x}), ..., h_l(\mathbf{x})]^T$ is a vector of l fixed functions of the input, often termed *features*, and $\boldsymbol{\omega} \equiv [\boldsymbol{\omega}^{(1)^T}, ..., \boldsymbol{\omega}^{(K)^T}]^T$ denotes the logistic regressors. Since the density in Eq. (3.4) does not depend on translations of the regressors $\boldsymbol{\omega}^{(K)}$, we take $\boldsymbol{\omega}^{(K)} = \mathbf{0}$ and remove it from $\boldsymbol{\omega}$, *i.e.*, $\boldsymbol{\omega} \equiv [\boldsymbol{\omega}^{(1)^T}, ..., \boldsymbol{\omega}^{(K-1)^T}]^T$.

It should be noted that function $\mathbf{h}$ may be linear, *i.e.*, $\mathbf{h}(\mathbf{x}_i) = [1, x_{i,1}, ..., x_{i,d}]^T$, where $x_{i,j}$ is the j-th component of $\mathbf{x}_i$. Alternatively, $\mathbf{h}$ can also be nonlinear. For the nonlinear case, kernels are a relevant example and can be expressed by $\mathbf{h}(\mathbf{x}_i) = [1, K_{\mathbf{x}_i,\mathbf{x}_1}, ..., K_{\mathbf{x}_i,\mathbf{x}_l}]^T$, where $K_{\mathbf{x}_i,\mathbf{x}_j} \equiv K(\mathbf{x}_i, \mathbf{x}_j)$ and $K(\cdot,\cdot)$ is some symmetric kernel function. Kernels have been largely used in this context because they tend to improve the data separability in the transformed space. In this paper, we present results only for the Gaussian Radial Basis Function (RBF) kernel, given by $K(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2/(2\rho^2))$. The RBF kernel has been widely used in hyperspectral image classification [28]. If we denote by γ the dimension of $\mathbf{h}(\mathbf{x})$, then we have $\gamma = d + 1$ for the linear case and $\gamma = L + 1$ for the RBF kernel (recall that L is the number of samples in the training set $\mathcal{D}_L$). In addition to the Gaussian RBF kernel, we have considered other alternative kernels such as the polynomial one. However, we have experimentally tested that the results obtained are very similar in both cases. Hence, in the following we adopt the Gaussian RBF kernel as a baseline for simplicity.

3.3.1 LORSAL

In our context, learning the class densities amounts to estimating the logistic regressors $\boldsymbol{\omega}$. Following the principles of the SMLR algorithm [80], the estimation of $\boldsymbol{\omega}$ amounts to computing the MAP estimate:

$$\widehat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \quad \ell(\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}), \tag{3.5}$$

where $\ell(\boldsymbol{\omega})$ is the log-likelihood function given by:

$$\ell(\boldsymbol{\omega}) \equiv \log \prod_{i=1}^{L} p(y_i|\mathbf{x}_i, \boldsymbol{\omega}), \tag{3.6}$$

and

$$p(\boldsymbol{\omega}) \propto \exp(-\lambda \|\boldsymbol{\omega}\|_1) \tag{3.7}$$

is a Laplacian prior promoting the sparsity on $\boldsymbol{\omega}$ ($\|\boldsymbol{\omega}\|_1$ denotes the l_1 norm of $\boldsymbol{\omega}$) with λ acting as a regularization parameter. The prior $p(\boldsymbol{\omega})$ forces many components of $\boldsymbol{\omega}$ to be zero. Thus, the Laplacian prior selects just a few kernel functions. The sparseness imposed on the regression vector controls the MLR classifier complexity and, consequently, enhances its generalization capacity.

Solving the convex problem in Eq. (3.5) is difficult because the term $\ell(\boldsymbol{\omega})$ is non-quadratic and the term $\log p(\boldsymbol{\omega})$ is non-smooth. A majorization-minimization framework[69] has recently been used in [80, 81, 89, 91] to decompose the problem in Eq. (3.5) into a sequence of quadratic problems. The computational cost of the SMLR algorithm used for solving each quadratic problem is $O((\gamma K)^3)$, which is prohibitive when dealing with datasets with a large number of features, with a large number of classes, or both. The fast sparse multinomial logistic regression (FSMLR) [18] estimates the sparse regressors in an efficient way by implementing a block-based Gauss-Seidel iterative procedure to calculate $\boldsymbol{\omega}$. This procedure is on the order of K^2 faster than the original SMLR algorithm. Thus, the FSMLR algorithm extends the capability of SMLR to handle data sets with a large number of classes. However, with an overall complexity of $O(\gamma^3 K)$, the complexity of FSMLR is still unbearable in many cases, in particular, for hyperspectral data sets with high-dimensional features.

In this paper, we resort to the recently introduced LORSAL algorithm [12] to learn the MLR regressors given by Eq. (3.5). By replacing the $\log p(\boldsymbol{\omega})$ in Eq. (3.5) with $\log p(\boldsymbol{\nu})$, approximating $\ell(\boldsymbol{\omega})$ with a quadratic majorizer, and introducing the constraint $\boldsymbol{\omega} = \boldsymbol{\nu}$, the LORSAL algorithm replaces a difficult non-smooth convex problem with a sequence of quadratic plus diagonal l_2-l_1 problems which are easier to solve. For additional details see the Appendix located at the end of this paper. In practice, the total cost of the LORSAL algorithm is $O(\gamma^2 K)$ per iteration, which contrasts with the $O((\gamma K)^3)$ and $O(\gamma^3 K)$ complexities of SMLR and FSMLR, respectively. As a result, the reduction of computational complexity is on the order of γK^2 and γ, respectively.

3.3.2 The multi-level logistic (MLL) spatial prior

In order to encourage piecewise smooth segmentations and promote solutions in which adjacent pixels are likely to belong to the same class, we include spatial-contextual information in our proposed method by adopting an isotropic MLL prior to model the image of class labels $\mathbf{y}$. This prior, which belongs to the MRF class, is a generalization of the Ising model [58] and has been widely used in image segmentation problems (see *e.g.*, [19, 89, 91, 92]).

According to the Hammersly-Clifford theorem [10], the density associated with an MRF is a Gibbs' distribution [58]. Thus, the prior model has the structure:

$$p(\mathbf{y}) = \frac{1}{Z}\, e^{\left(-\sum_{c\in\mathcal{C}} V_c(\mathbf{y})\right)}, \tag{3.8}$$

where Z is a normalizing constant for the density, the sum in the exponent is over the so-called prior potentials $V_c(\mathbf{y})$ for the set of cliques $\mathcal{C}$ over the image, and:

$$-V_c(\mathbf{y}) = \begin{cases} \upsilon_{y_i}, & \text{if } |c| = 1 \text{ (single clique)} \\ \mu_c, & \text{if } |c| > 1 \text{ and } \forall_{i,j\in c} y_i = y_j \\ -\mu_c, & \text{if } |c| > 1 \text{ and } \exists_{i,j\in c} y_i \neq y_j, \end{cases} \tag{3.9}$$

where μ_c is a non-negative constant.

The potential function in Eq. (3.9) encourages neighbors to have the same class label. The considered MLL prior offers great flexibility in this task by varying the set of cliques and the parameters υ_{y_i} and μ_c. For example, the model generates texture-like regions if μ_c depends on c and blob-like regions otherwise [93]. The single clique term υ_{y_i} determines the marginals $p(y_i)$, *i.e.*, the prior class distributions. In this work, we assume equiprobable classes and this implies that υ_{y_i} is constant. We note, however, that any other distribution can be modeled by a suitable choice of the term υ_{y_i}. Then Eq. (3.8) can be rewritten as:

$$p(\mathbf{y}) = \frac{1}{Z}\, e^{\mu \sum_{\{i,j\}\in\mathcal{C}} \delta(y_i - y_j)}, \tag{3.10}$$

where $\delta(y)$ is the unit impulse function. This choice gives no preference to any direction. Notice that the pairwise interaction terms $\delta(y_i - y_j)$ attach higher probability to equal neighboring labels than the other way around. In this way, the MLL prior promotes piecewise smooth segmentations, where μ controls the degree of smoothness.

3.3.3 Computing the MAP estimate via graph-cuts

Using the LORSAL algorithm to learn $p(y_i|\mathbf{x}_i)$ and the MLL prior $p(\mathbf{y})$, and according to Eq. (3.3), the MAP segmentation is finally given by:

$$\widehat{\mathbf{y}} \quad = \quad \arg\min_{\mathbf{y}\in\mathcal{L}^n} \left\{ \sum_{i\in\mathcal{S}} -\log p(y_i|\widehat{\boldsymbol{\omega}}) - \mu \sum_{i,j\in C} \delta(y_i - y_j) \right\}, \tag{3.11}$$

where $p(y_i|\widehat{\boldsymbol{\omega}}) \equiv p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$, computed at $\widehat{\boldsymbol{\omega}}$. Minimization of Eq. (3.11) is a combinatorial optimization problem involving unary and pairwise interaction terms, which is very difficult to compute. Recently developed energy minimization algorithms like graph-cuts [22, 23, 79], loopy belief propagation [141, 142], and tree-reweighed message passing [78] are efficient tools to tackle this class of optimization problems. In this work, we use the α-Expansion algorithm [23] to solve our integer optimization problem [4]. This algorithm yields very good approximations to the MAP segmentation and is quite efficient from a computational point of view, being the practical computational complexity of this algorithm $O(n)$. The pseudo-code for the proposed supervised segmentation algorithm with discriminative class learning and MLL prior is shown in Algorithm 3.1.

Algorithm 3.1 Supervised segmentation algorithm (LORSAL-MLL)

Require: $\mathcal{D}_L$, λ, β

1: $\widehat{\boldsymbol{\omega}} := \text{LORSAL}(\mathcal{D}_L, \lambda, \beta)$
2: $\widehat{\mathbf{P}} := \widehat{\mathbf{p}}(\mathbf{x}_i, \widehat{\boldsymbol{\omega}})$, $i \in \mathcal{S}$
3: $\widehat{\mathbf{y}} := \alpha\text{-Expansion}(\widehat{\mathbf{P}}, \mu)$

3.3.4 Overall complexity

The overall complexity of our proposed approach is dominated by the supervised learning of the MLR regressors through the LORSAL algorithm, shown in Algorithm 3.4 (see Appendix), which has a complexity of $O(\gamma^2 K)$, and by the α-Expansion algorithm used to determine the MAP segmentation, which has a practical complexity of $O(n)$. In conclusion, if $\gamma^2 K \gg n$ (*e.g.*, $\mathbf{h}(\mathbf{x})$ are kernels and the number of classes is large), then the algorithm's complexity is dominated by the computation of the MLR regressors, whereas if $\gamma^2 K \ll n$, then the algorithm's complexity is dominated by the $\alpha-$Expansion algorithm.

3.4 Active learning

In this work, we use active learning to reduce the need for large amounts of labeled samples. The basic idea of active learning is to iteratively enlarge the training set by requesting an expert

to label new samples from the unlabeled set $\{\mathbf{x}_i, i \in \mathcal{S}_U\}$ in each iteration, where $\mathcal{S}_U$ is the set of unlabeled feature vectors, *i.e.*, spectral vectors in the observed context. The relevant question is, of course, what vectors in $\mathcal{S}_U$ are most informative and should be chosen as new samples. In this paper, we take advantage of the MLR model, which provides the exact posterior probabilities. Therefore, three different sampling schemes, based on the spectral information (more specifically, on the MLR posterior probabilities just provided by the LORSAL algorithm) are implemented: (a) MI-based criterion [81, 96]; (b) BT algorithm [95]; and (c) our proposed MBT scheme.

3.4.1 MI-based active learning

The first active learning scheme considered is an MI-based criterion [81, 96] that maximizes the mutual information between the MLR regressors and the class labels. Let $\mathrm{I}(\boldsymbol{\omega}; y_i|\mathbf{x}_i)$ denote the MI between the MLR regressors and the class label y_i. Following [96], the new vector $\mathbf{x}_i$ is selected according to:

$$\widehat{\mathbf{x}}_i^{\mathrm{MI}} = \arg \max_{\mathbf{x}_i,\, i \in \mathcal{S}_U} \mathrm{I}(\boldsymbol{\omega}; y_i|\mathbf{x}_i), \tag{3.12}$$

where (see [96] for more details)

$$\mathrm{I}(\boldsymbol{\omega}; y_i|\mathbf{x}_i) \;\; = \;\; (1/2) \log(|\mathbf{H}^{\mathrm{MI}}|/\mathbf{H}). \tag{3.13}$$

Here, $\mathbf{H}$ is the posterior precision matrix, *i.e.*, the Hessian of minus the log-posterior [131]

$$\mathbf{H} \equiv \nabla^2(-\log p(\widehat{\boldsymbol{\omega}}|\mathcal{D}_L)),$$

and $\mathbf{H}^{\mathrm{MI}}$ is the posterior precision matrix after including the new sample $\mathbf{x}_i$. In the proposed approach, we use a Laplacian approximation of the posterior to model $p(\boldsymbol{\omega}|\mathcal{D}_L)$, such that $p(\boldsymbol{\omega}|\mathcal{D}_L) \simeq \mathcal{N}(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}}, \mathbf{H}^{-1})$, which assumes that the MAP estimate $\widehat{\boldsymbol{\omega}}$ remains unchanged after including the new sample. If the size of the initial training sample is "small", this assumption may not hold at the beginning of the active learning procedure. Nevertheless, it has been empirically observed that it leads to a very good approximation [81, 88]. Under this assumption, we can compute $\mathbf{H}^{\mathrm{MI}}$ as follows:

$$\mathbf{H}^{\mathrm{MI}} = \mathbf{H} + \left(\mathrm{diag}(\mathbf{p}_i(\widehat{\boldsymbol{\omega}})) - \mathbf{p}_i(\widehat{\boldsymbol{\omega}})\mathbf{p}_i(\widehat{\boldsymbol{\omega}})^T\right) \otimes \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)^T, \tag{3.14}$$

where $\mathbf{p}_i(\widehat{\boldsymbol{\omega}}) \equiv [p_{i,1}, \ldots, p_{i,K}]^T$, $p_{i,k} \equiv p(y_i = k|\mathbf{x}_i, \widehat{\boldsymbol{\omega}})$ for $k = 1, \ldots, K$, and $\otimes$ is the Kronecker

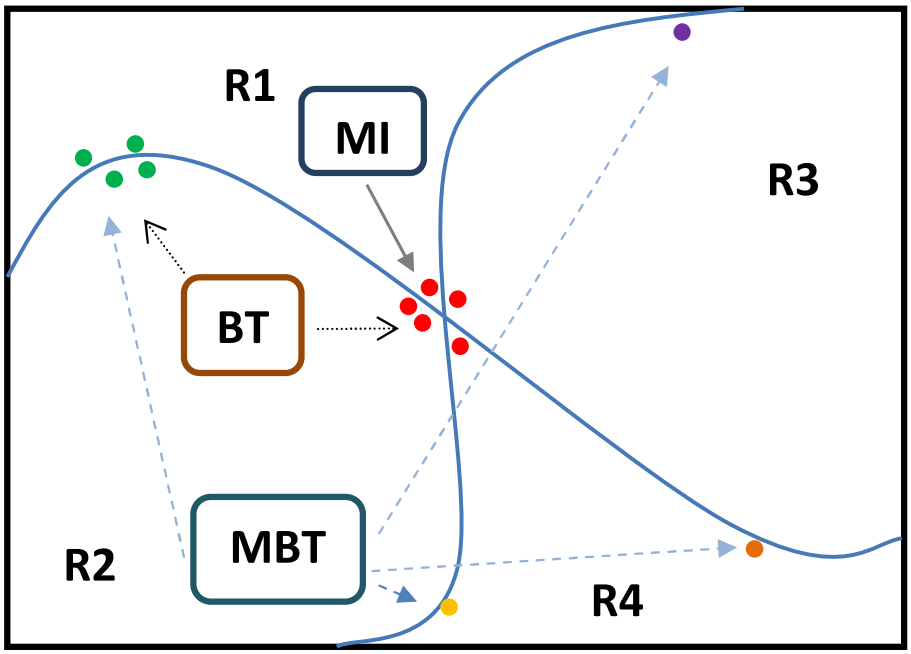


Figure 3.1: Graphical illustration of the MI, BT and MBT active learning approaches using a toy example.

product. Therefore, Eq. (3.13) turns to:

$$\mathrm{I}(\boldsymbol{\omega}; y_i|\mathbf{x}_i) \quad = \quad (1/2)\log\left(1+\prod_{k=1}^{K} p_{i,k}\ \mathbf{x}_i^T\mathbf{H}^{-1}\mathbf{x}_i\right). \tag{3.15}$$

According to Eq. (3.15), the function in Eq. (3.12) is maximized for $p_{i,k} \approx 1/K$, *i.e.*, for samples near the boundaries among classes and corresponding to probability vectors $\mathbf{p}_i$ with maximum entropy. This situation is graphically illustrated in Figure 3.1, in which a toy example with four simulated regions is used for demonstration purposes. As shown by Figure 3.1, the MI focuses on the most complex area (boundary between the four regions).

3.4.2 BT active learning

The BT active learning algorithm [95] was proposed to achieve diversity in the sampling, thus alleviating the bias in the MI-based sampling. The decision criterion is:

$$\widehat{\mathbf{x}}_i^{\mathrm{BT}} = \arg\min_{\mathbf{x}_i,\ i\in\mathcal{S}_U} \left\{ \max_{k\in\mathcal{L}} p(y_i=k|\mathbf{x}_i,\ \widehat{\boldsymbol{\omega}}) - \max_{k\in\mathcal{L}\setminus\{k^+\}} p(y_i=k|\mathbf{x}_i,\ \widehat{\boldsymbol{\omega}}) \right\}, \tag{3.16}$$

where $k^+ = \arg\max_{k\in\mathcal{L}} p(y_i=k|\mathbf{x}_i,\ \widehat{\boldsymbol{\omega}})$ is the most probable class for sample $\mathbf{x}_i$.

Other than the MI-based criterion, which focuses on the most complex regions (i.e, regions with the largest number of boundaries), the BT criterion focuses on the boundary region between two classes, with the goal of obtaining more diversity in the composition of the training set. In

spite of the better performance generally expected from the BT criterion with respect to the MI-based one, it may still produce biased sampling, namely, when there are many samples located close to a boundary. This can be seen in Figure 3.1, which illustrates how the BT criterion generally focuses on the boundaries comprising many samples, possibly disregarding boundaries with fewer samples but which may be crucial for the learning procedure needed to train discriminative classifiers. In the following subsection, we propose a new modified scheme (called MBT) which promotes even more diversity in the sampling process.

3.4.3 MBT active learning

For a given $\widehat{\boldsymbol{\omega}}$ and $s \in \mathcal{L}$, let $\mathcal{S}_{U_s} \subset \mathcal{S}_U$ be the set of pixels such that $p(y_i = s|\mathbf{x}_i,\ \widehat{\boldsymbol{\omega}}) \geq p(y_i = k|\mathbf{x}_i,\ \widehat{\boldsymbol{\omega}})$, for $i \in \mathcal{S}_{U_s}$ and $k \neq s$. Then, the MBT criterion simply works as follows:

$$
\begin{aligned}
&\textbf{do} \\
&\quad s = \text{next class} \\
&\quad \text{select}\ \ \mathcal{S}_{U_s} \\
&\quad \widehat{\mathbf{x}}_i^{\text{MBT}} = \arg \max_{\mathbf{x}_i,\, i \in \mathcal{S}_{U_s},\, k \in \mathcal{L}\setminus\{s\}} p(y_i = k|\mathbf{x}_i,\ \widehat{\boldsymbol{\omega}}), \\
&\textbf{while}\ \text{stop rule}
\end{aligned}
\tag{3.17}
$$

where the "next class" is chosen by scanning the index set $\mathcal{L}$ in a cyclic fashion. We highlight the following two characteristics of the MBT criterion in Eq. (3.17), both intended to promote diversity in the selection process as compared with the BT criterion:

- By cyclically selecting subsets of $\mathcal{S}_U$ containing the pixels with the same MAP label, it is assured that the MBT criterion does not get trapped in any class.
- The step $\max_{k \in \mathcal{L}\setminus\{s\}} p(y_i = k|\mathbf{x}_i,\ \widehat{\boldsymbol{\omega}})$ tends to select new samples away from complex areas. As shown by Figure 3.1, the main advantage of the proposed MBT with regards to other active learning approaches such as MI or BT is that the former method takes into account all the class boundaries which are crucial to the learning procedure when conducting the sampling, whereas MI mainly focuses on the most complex area and BT may get trapped in a single boundary.

After having presented the three sampling methods considered in this work: MI, BT and MBT, it is now important to emphasize that Eqs. (3.12), (3.16) and (3.17) assume that only one sample is labeled at each iteration. However, in practice we consider $u > 1$, *i.e.*, we label more than one sample per iteration. Let $\mathcal{D}_u \equiv \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_u, y_u)\}$ be the new labeled set. For the MBT sampling, we adopt a two-step scheme. First, $\text{round}(u/K) + 1$ new samples per

class are selected according to Eq. (3.17), where function round(·) simply rounds toward the nearest integer value. Second, we run Eq. (3.16) to select the u most informative samples for the recently obtained set. For binary classification problems, the MI, BT and MBT strategies can be considered equivalent since they lead to exactly the same new labeling for any u. However, for multi-class problems the three considered strategies may lead to different labelings. In turn, when u is very small the performance of BT and MBT becomes similar.

To conclude this section, Algorithm 3.2 shows the pseudo-code of the LORSAL algorithm using active learning (called LORSAL-AL), where $\beta \geq 0$ is the augmented Lagrangian LORSAL parameter (see Appendix). Finally, the supervised segmentation algorithm with active learning (called LORSAL-MLL-AL) is shown in Algorithm 3.3.

Algorithm 3.2 LORSAL using active learning (LORSAL-AL)

Require: $\widehat{\boldsymbol{\omega}}$, $\mathcal{D}_L$, $\mathcal{S}_U$, u, λ, β

1: **repeat**
2: $\quad \mathcal{D}_u := \text{AL}(\widehat{\boldsymbol{\omega}},\ \mathcal{S}_U)$ (function AL(·) is one of the sampling methods: RS, MI, BT and MBT.)
3: $\quad \mathcal{D}_L := \mathcal{D}_L + \mathcal{D}_u$
4: $\quad \mathcal{S}_U := \mathcal{S}_U - \{1, \ldots, u\}$
5: $\quad \widehat{\boldsymbol{\omega}} := \text{LORSAL}(\mathcal{D}_L, \lambda, \beta)$
6: **until** some stopping criterion is met

Algorithm 3.3 Supervised segmentation algorithm using active learning (LORSAL-AL-MLL)

Require: $\widehat{\boldsymbol{\omega}}$, $\mathcal{S}_U$, $\mathcal{D}_L$, u, λ, β

1: **repeat**
2: $\quad \mathcal{D}_u := \text{AL}(\widehat{\boldsymbol{\omega}},\ \mathcal{S}_U)$
3: $\quad \mathcal{D}_L := \mathcal{D}_L + \mathcal{D}_u$
4: $\quad \mathcal{S}_U := \mathcal{S}_U - \{1, \ldots, u\}$
5: $\quad \widehat{\boldsymbol{\omega}} := \text{LORSAL}(\mathcal{D}_L, \lambda, \beta)$
6: **until** some stopping criterion is met
7: $\widehat{\mathbf{y}} := \alpha\text{-Expansion}(\widehat{\mathbf{P}}, \mu)$

3.5 Experimental results

In this section, we evaluate the performance of the proposed algorithm using both simulated and real hyperspectral data sets. The main objective of the experimental validation with simulated data sets is the assessment and characterization of the algorithm in a fully controlled environment, whereas the main objective of the experimental validation with real data sets is to compare the performance of the proposed method with that reported for state-of-the-art competitors in the literature.

It should be noted that, in all of our experiments, we apply the Gaussian RBF kernel

to a normalized version of the input hyperspectral data. Alternative experiments have been conducted with other kernels, such as the polynomial one, obtaining very similar results. The scale parameter is set to a fixed value $\rho = 0.6$, as we have empirically proved that this setting leads to good characterization results. Another reason is that we have not observed significant improvements for small variations of ρ. In the following, we assume that $\mathcal{D}_{L_i}$ denotes the initial labeled set, which is a subset of the available training set, and that L_i denotes the number of samples (recall that L denotes the total number of labeled samples). In practice, we assume that the initial training samples for each class are uniformly distributed. Concerning the smaller classes, if the total labeled samples of class k in the ground truth image, say L_k, is smaller than L/K, we take $L_k/2$ as the initial number of labeled samples. In this case, larger classes have more samples. In all cases, the reported figures of overall accuracy (OA) are obtained by averaging the results obtained after conducting 10 independent Monte Carlo runs with respect to $\mathcal{D}_{L_i}$.

The remainder of the section is organized as follows. Section 3.5.1 reports experiments with simulated data, with Subsection 3.5.1.I conducting an evaluation of the LORSAL algorithm, Subsection 3.5.1.II evaluating the impact of the spatial prior, and Subsection 3.5.1.III evaluating the impact of the active learning approaches. Section 3.5.2 evaluates the performance of the proposed algorithm using four real hyperspectral scenes collected by the Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS), operated by NASA Jet Propulsion Laboratory, and by the Reflective Optics Imaging Spectrometer System (ROSIS), operated by the German Aerospace Agency (DLR).

3.5.1 Experiments with simulated data

In our simulated data experiments, we generate images of labels denoted by $\mathbf{y} \in \mathcal{L}^n$, sampled from a 128×128 MLL distribution with $\mu = 2$. The feature vectors are simulated according to:

$$\mathbf{x}_i = \mathbf{m}_{y_i} + \mathbf{n}_i, \quad i \in \mathcal{S}, \;\; y_i \in \mathcal{L}, \tag{3.18}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the spectral vector observed at pixel i, $\mathbf{m}_{y_i}$ denotes a set of K known vectors, and $\mathbf{n}_i$ denotes zero-mean Gaussian noise with covariance $\sigma^2\mathbf{I}$, *i.e.*, $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. In Subsections 3.5.1.I and 3.5.1.II we will not consider the active learning procedure (*i.e.*, $L = L_i$) because our focus in these two subsections will be on analyzing the competitiveness of the LORSAL algorithm and on evaluating the role of the spatial prior independently of the active learning mechanism, respectively. In both cases, the training set $\mathcal{D}_L$ is a subset of the ground-truth image, whereas the remaining samples are considered as the test set. Finally, Subsection 3.5.1.III analyzes the impact of including the active learning mechanism in the proposed method.

We would like to state that, in these experiments, the initial labeled set $\mathcal{D}_{L_i}$ is randomly selected from the ground-truth image, whereas the remaining samples are considered as the validation set. At each iteration of the active sampling procedure, the new set $\mathcal{D}_u$ is actively selected from the test set. This is a sub-optimal procedure for the evaluation of the accuracies. However, in these experiments, the maximum training set used is made up of 80 samples, which represents only 0.49% of the whole image. According to this, we believe that the active learning process would not be harmful to the evaluation of the accuracy in our proposed setting. Therefore, we do not separate the training and test sets, which also guarantees that the test set remains as large as possible. In the real image experiments, we completely separate the training and test sets.

I. Evaluation of the LORSAL algorithm

In this subsection, we generate the simulated hyperspectral data according to the model in Eq. (3.18), where spectral vectors $\mathbf{m}_i$, with $i = 1, \ldots, K$, were selected (randomly) from the U.S. Geological Survey (USGS) digital spectral library with $d = 224$, $K = 10$, $L = 1000$, and $\sigma = 1$.

In our first experiment, we illustrate the computational efficiency of the LORSAL algorithm. Figure 3.2 represents the log-posterior $\ell(\boldsymbol{\omega}) - \lambda\|\boldsymbol{\omega}\|_1$ as a function of the computation time for LORSAL, FSMLR, and SMLR algorithms (implemented in Matlab). As it can be seen in Figure 3.2, LORSAL is by far the fastest algorithm. For a similar log-posterior, the LORSAL algorithm took about 2 seconds in a desktop PC with Intel Core 2 Duo CPU at 2.40 GHz and 4 GB of RAM memory, while the FSMLR and SMLR algorithms took, respectively, around 48 and 880 seconds in the same computing environment.

As already mentioned, the regularization parameter λ in Eq. (3.7) controls the sparseness of the regressors, which is essential to the generalization capacity. However, an inappropriate value of λ may lead to overfitting or underfitting scenarios. In practice, we estimate λ by using cross-validation sampling [77] over the initial training set. Nevertheless, in our second experiment we conduct an analysis of the impact of λ on the achieved performance. Let $\xi = 100 \times \frac{n_{\omega_0}}{n_\omega}\%$, where n_ω and n_{ω_0} denote the number of components and zeros in $\boldsymbol{\omega}$, respectively. Figure 3.3 shows the OA and ξ as a function of λ, for $10^{-2} \leq \lambda \leq 30$. The impact of λ on the sparsity of ω is clear. The higher values of OA are obtained for $\lambda \in [2, 10]$ corresponding to levels of sparsity $\xi \in [50, 60]\%$.

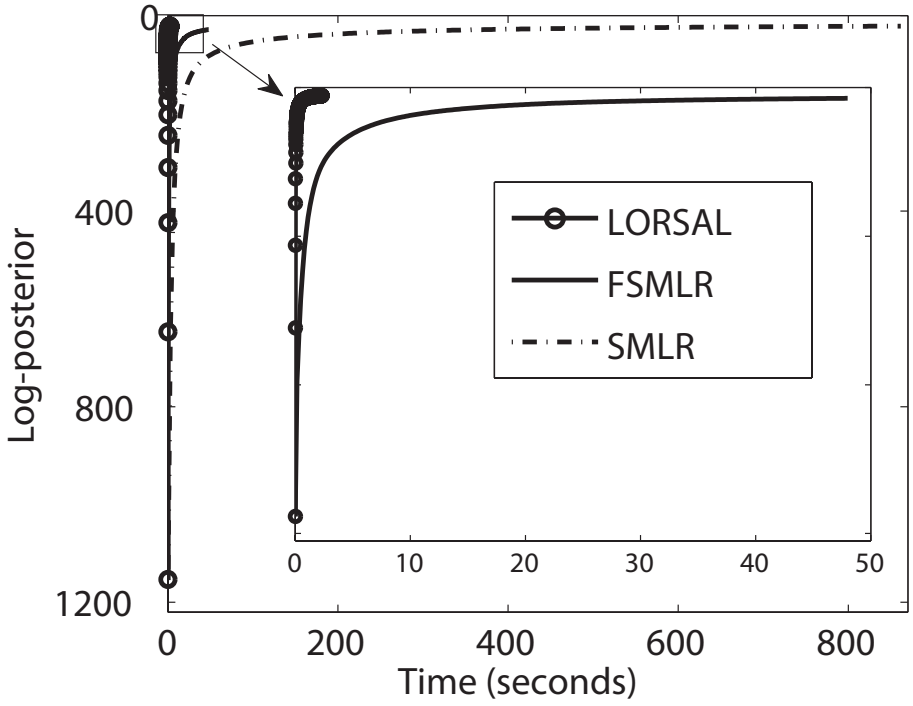


Figure 3.2: Evaluation of the log-posterior in Eq. (3.5) as a function of the computing time (measured in a desktop PC with Intel Core 2 Duo CPU at 2.40 GHz and 4 GB of RAM memory) for LORSAL, FSMLR, and SMLR algorithms.

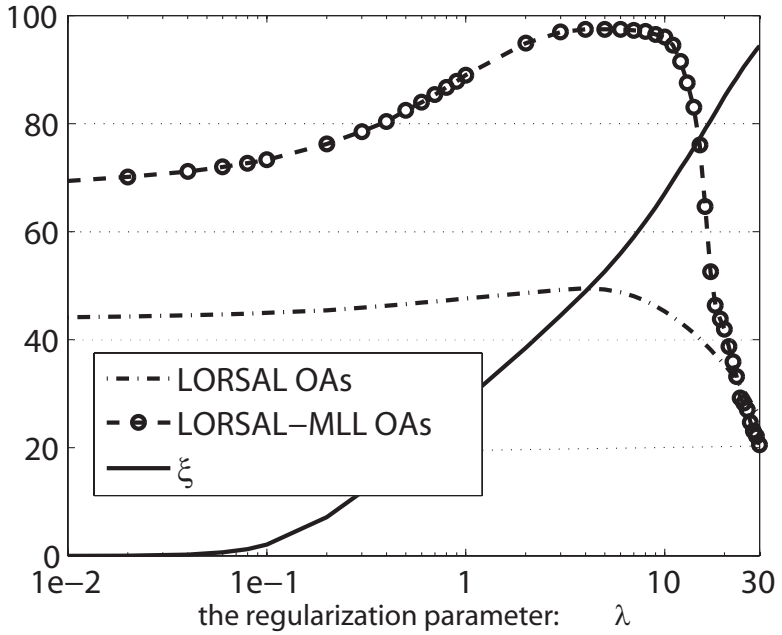
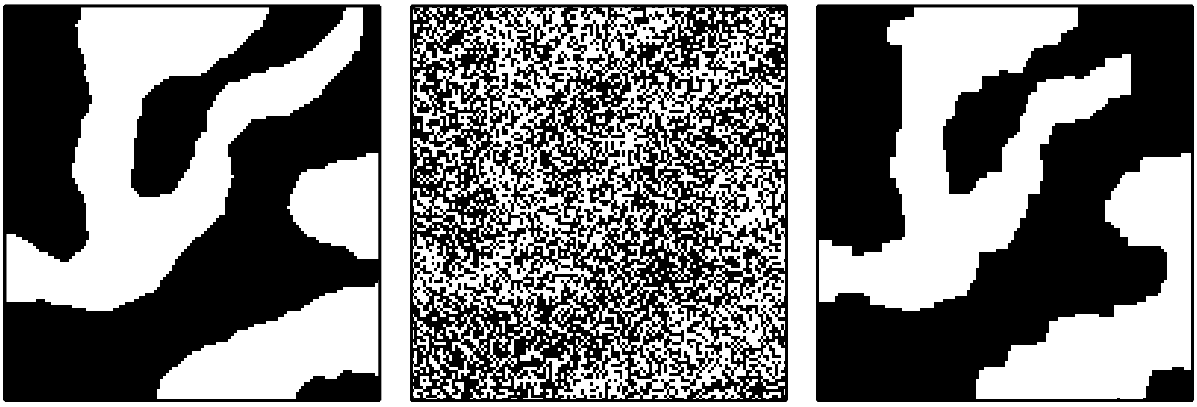


Figure 3.3: Evaluation of the impact of the regularization parameter, λ, on the overall accuracy, OA, and on the level of sparsity, ξ.



(a) Ground-truth (b) Classification map (c) Segmentation map

Figure 3.4: Classification and segmentation results obtained with the proposed algorithm. The simulated data set was generated according to Eq. (3.18) with $d = 500$ and $\sigma = 1.5$, $\mu = 2$. (a) Simulated binary map; (b) Classification map produced by the LORSAL algorithm using $L = 100$ labeled samples without active learning (OA=60.13%, with $\mathrm{OA}_{opt} = 71.91\%$, see text); (c) Same as (b) but using the MLL spatial prior (OA=92.48%).

II. Impact of the spatial prior

In this experiment, we analyze the impact of the spatial prior on the segmentation accuracy in a binary problem, *i.e.*, with $K = 2$. The feature vector is set to $\mathbf{m}_i = \xi_i \phi$, where $\|\phi\| = 1$ and $\xi_i = \pm 1$. An image of class labels $\mathbf{y}$ generated according to the MLL prior in Eq. (3.18) is shown in Figure 3.4(a), where the labels $y_i = 1, 2$ correspond to $\xi_i = -1, +1$, respectively. In this problem, the theoretical OA, given by $\mathrm{OA}_{opt} \equiv 100\,(1 - \mathrm{P}_e)\%$ and corresponding to the minimal probability of error [49] is:

$$\mathrm{P}_e = \frac{1}{2}\,\mathrm{erfc}\left(\frac{1+\lambda_0}{\sqrt{2}\,\sigma}\right) p_0 + \frac{1}{2}\,\mathrm{erfc}\left(\frac{1-\lambda_0}{\sqrt{2}\,\sigma}\right) p_1, \tag{3.19}$$

where erfc is the complementary error function, $\lambda_0 = (\sigma^2/2)\ln(p_0/p_1)$ and p_0 and p_1 are the *a priori* class label probabilities. Usually, model parameters are estimated by cross-validation. However, in this work we concluded empirically that $\mu \in [2, 6]$ yields almost optimal results. In order to reduce computational efficiency, we have not applied cross-validation to derive the optimal value of this parameter. The aforementioned observation is illustrated in Figure 3.5 where we studied he impact of the spatial prior. Here, Figure 3.5(a) illustrates the OA results as a function of μ. For the considered problem, with $2 \leq \mu \leq 6$, the LORSAL-ALL algorithm obtained good segmentation results. It should be noted that 10 independent Monte Carlo runs were conducted in these experiments and we report only the mean scores obtained. The following conclusions may be drawn from Figure 3.5:

- The best overall results are obtained by the proposed segmentation algorithm (in all cases,

the classification accuracies and the values of OA_{opt} are higher). This confirms our introspection that the inclusion of a spatial prior can significantly improve the classification results provided by using only spectral information, even for very noisy scenarios [see Figure 3.5(b)].

- The classification OA approaches the optimal value OA_{opt} as the number of labeled samples increases [see Figure 3.5(c)]. However, the number of labeled samples needs to be relatively high in order to obtain classification accuracies which are close to optimal.

- For a fixed number of training samples, the classification accuracy of our proposed method decreases as the number of bands increases [see Figure 3.5(d)]. This is not surprising in light of the Huges phenomenon. On the contrary, after including the spatial prior our supervised segmentation algorithm performs very well even with small training sets and a large number of bands.

To give a broad picture of the good performance of the proposed algorithm, we finally illustrate the LORSAL classification and LORSAL-MLL segmentation maps in Figs. 3.4(b) and (c) for a problem with $\sigma = 1.5$ and $d = 500$ using $L = 100$ and $\mu = 2$. Clearly, the inclusion of the spatial prior yields, as expected, much better results.

III. Impact of the active learning approach

In this subsection we analyze the impact of the considered sampling strategies on our proposed approach. To do so, a new simulated hyperspectral data set is generated according to the model in Eq. (3.18), with $K = 4$, $\sigma = 0.8$, and vectors $\mathbf{m}_{y_i}$ obtained from the USGS library with $d = 224$. Figure 3.6 reports the learning results over 100 independent Monte Carlo runs, where we consider three different experiments: (a) OA results as a function of L by using $L_i = u = L/2$; (b) OA results as a function of L_i by using $L = 60$ and $u = L - L_i$; and (c) OA results as a function of u by using $L = 60$ and $L_i = 20$ (5 samples per class). Several conclusions can be obtained from the results reported in Figure 3.6:

- First of all, the active learning procedure improves the segmentation results as expected. In general, the MBT strategy achieves the best performance.

- Second, as already discussed in Section 3.4, with a small u both MBT and BT lead to very similar results.

- Furthermore, the results obtained by the MI sampling are highly dependent on the size of u. For a small size of u (such as $u < L_i$) good results are obtained, *e.g.* see Figure 3.6(c).

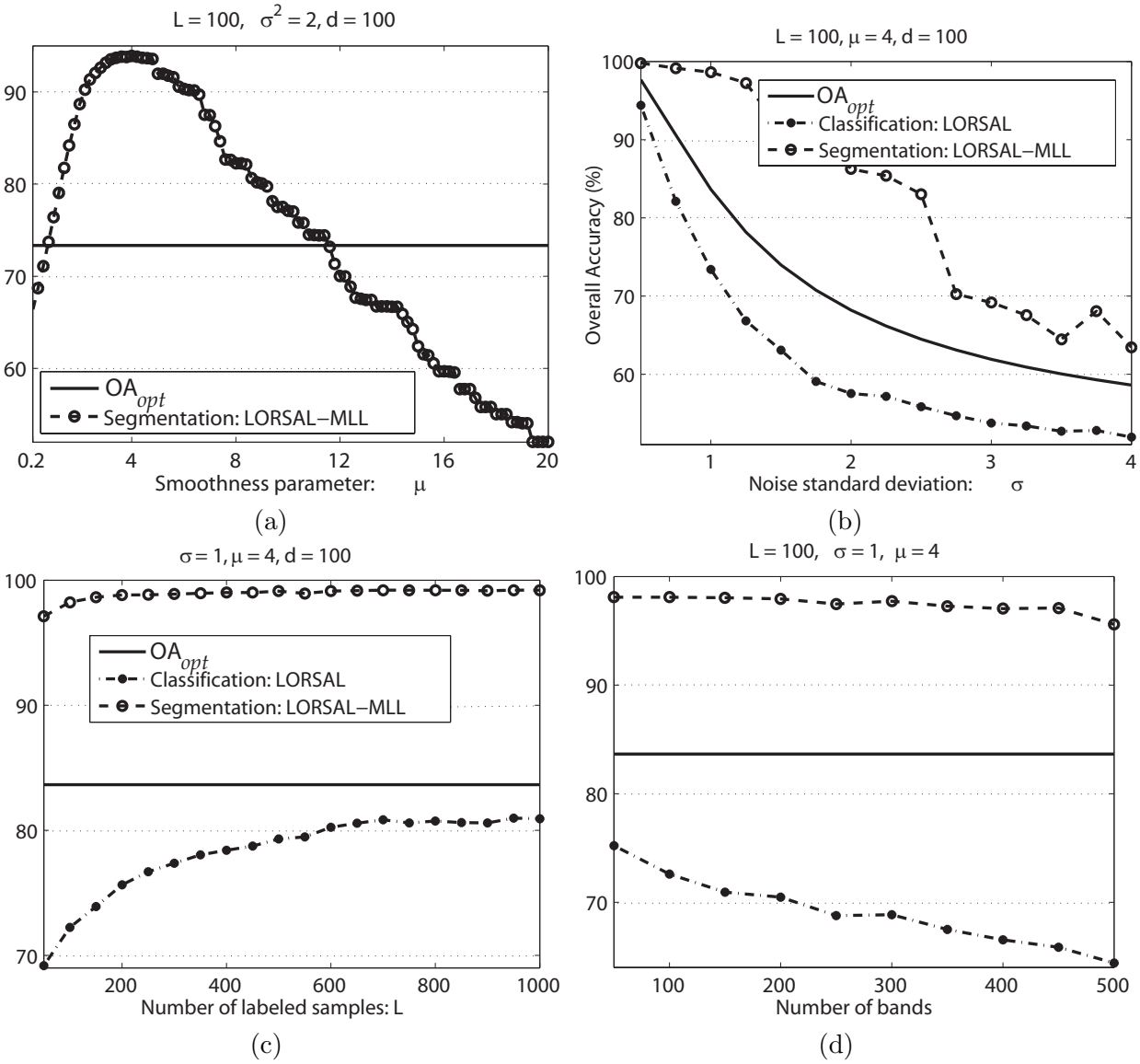


Figure 3.5: OA results obtained by the proposed algorithm: (a) As a function of the spatial prior parameter μ. (b) As a function of the noise standard deviation σ. (c) As a function of the number of labeled samples L. (d) As a function of the number of bands d.

However, for a large value of u, the MI sampling leads to results which are even worse than random selection. This is because the MI sampling focuses on the most complex area. Thus, with a large value of u the new predictions are concentrated in a most complex area which leads to poor generalization ability of the regressors.

- Finally, the improvements in performance due to active learning are less relevant as the size of the training set increases, *e.g.* see Figure 3.6(a). This is expected, since the uncertainty in the determination of classifier boundaries decreases as the training set size increases.

3.5.2 Experiments with real data sets

In this section, four real hyperspectral data sets are used to evaluate our algorithm. The first one is the well-known AVIRIS Indian Pines scene, collected over Northwestern Indiana in June 1992 [85]. The scene is available online[1] and contains 145×145 pixels and 224 spectral bands between 0.4 and 2.5 microns. A total of 20 spectral bands were removed prior to experiments due to noise and water absorption in those channels. The ground-truth image displayed in Figure 3.7(a), contains 16 mutually exclusive classes, 7 of which were discarded for their small size which resulted in insufficient training samples. The remaining 9 classes were used to randomly generate a set of 4757 training samples, with the remaining samples (4588) used for testing purposes.

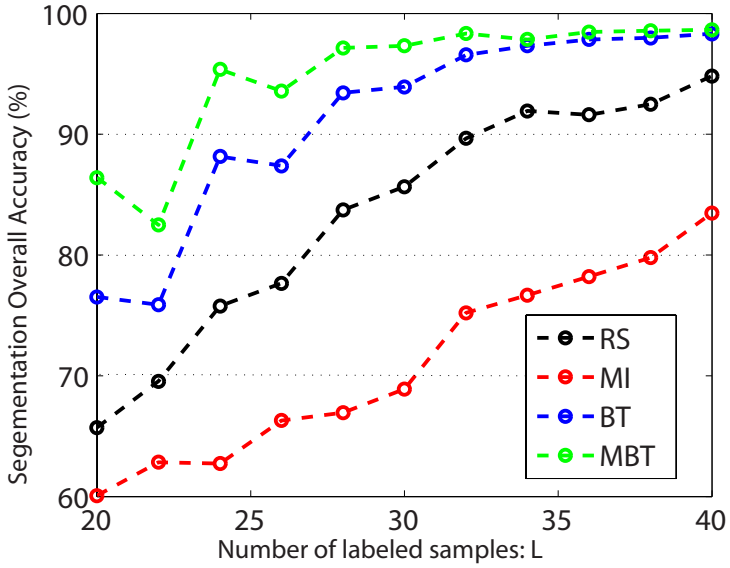
In addition to the AVIRIS Indian Pines scene, we have also used three ROSIS hyperspectral data sets collected over the town of Pavia, Italy. The data sets consist of 115 spectral bands between 0.4 and 1.0 microns. Three different subsets of the full data set are considered in our experiments:

- Subset #1, with 492×1096 pixels in size, collected over Pavia city center. The noisy bands were removed yielding a dataset with 102 spectral bands. The ground truth image contains 9 ground-truth classes, 5536 training samples, and 103539 test samples.

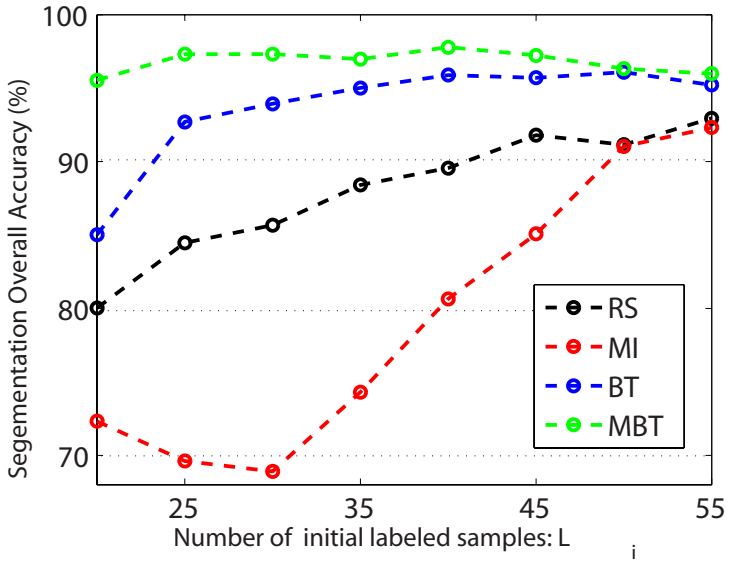
- Subset #2, with size of 610×340 pixels, centered at the University of Pavia in Italy. The noisy bands were removed yielding 103 spectral bands. The ground truth image in Figure 3.8(a), contains 9 ground-truth classes, 3921 training samples, and 42776 test samples.

- Subset #3 includes a dense residential area, with 715×1096 pixels. The ground-truth image contains 9 ground-truth classes, 7456 training samples and 148152 test samples.

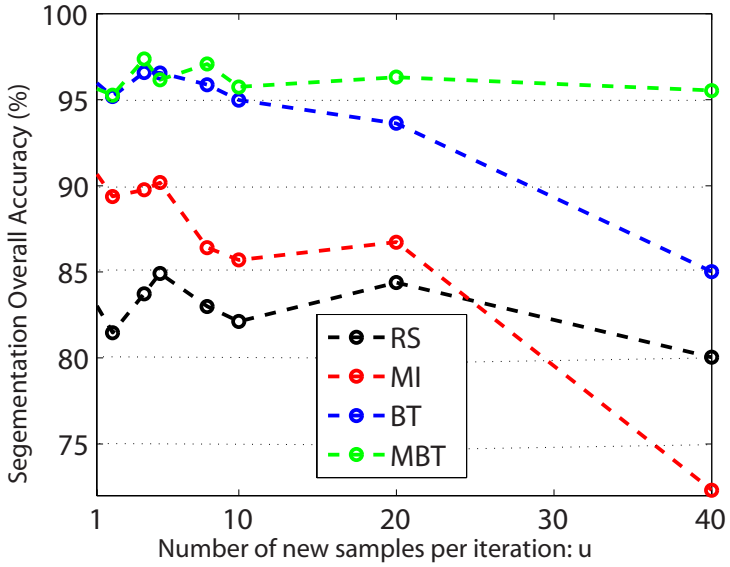
[1] https://engineering.purdue.edu/~biehl/MultiSpec/



(a) OA results as a function of L with $L_i = u = L/2$



(b) OA results as a function of L_i with $L = 60, u = L - L_i$



(c) OA results as a function of u with $L = 60, L_i = 20$

Figure 3.6: Segmentation results obtained by using active learning approaches.

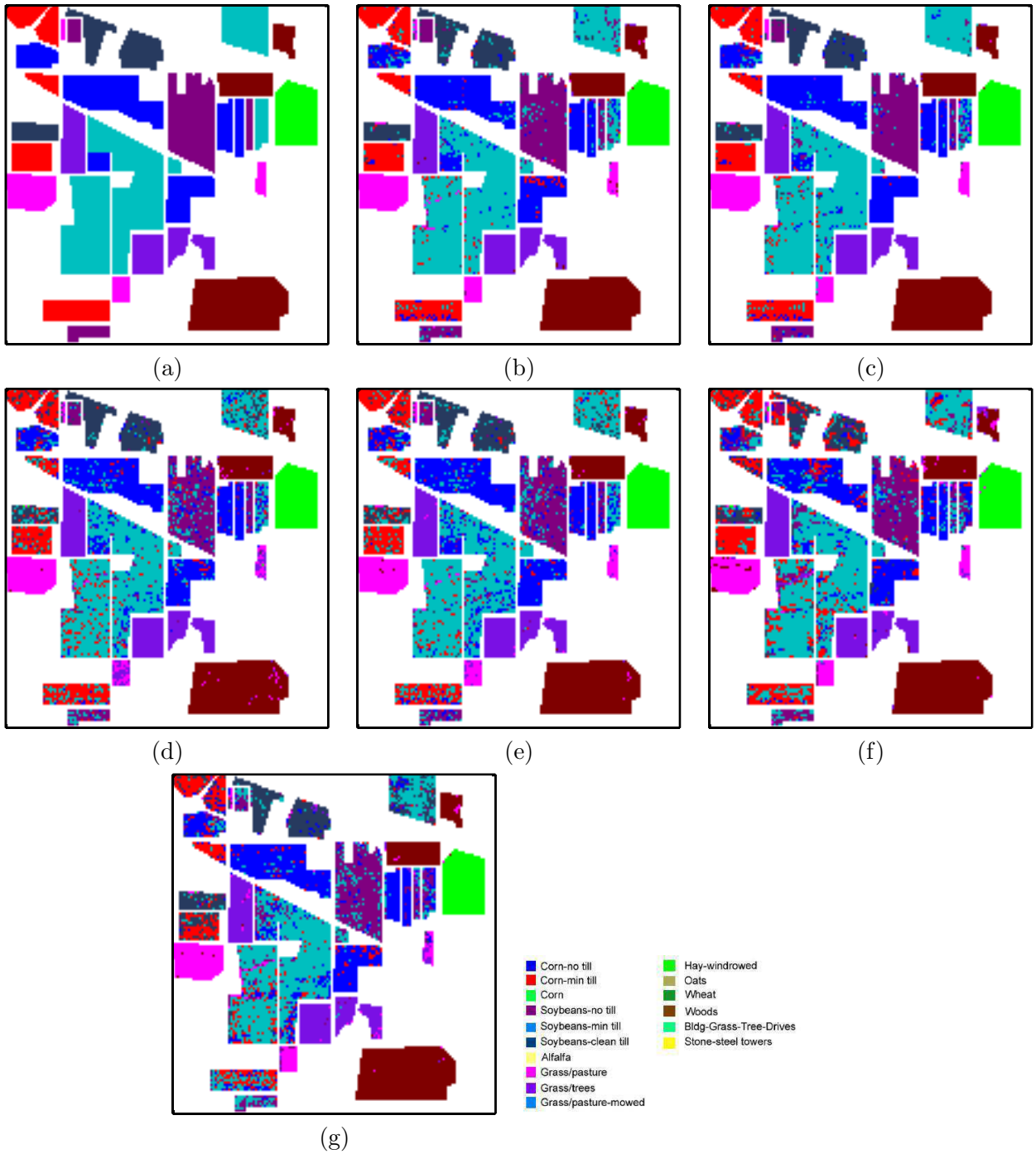


Figure 3.7: Classification maps by using $L = 475$, $L_i = 235$, $u = 60$. (a) Ground truth. (b) LORSAL-AL (RS), OA = 84.24%. (c) LORSAL-AL (MBT), OA = 86.38%. (d) LDA-AL (RS), OA = 69.35%. (e) LDA-AL (MBT), OA=70.83%. (f) SVM (RS), OA = 80.43%. (g) PCA+SVM (RS), OA=76.32%.

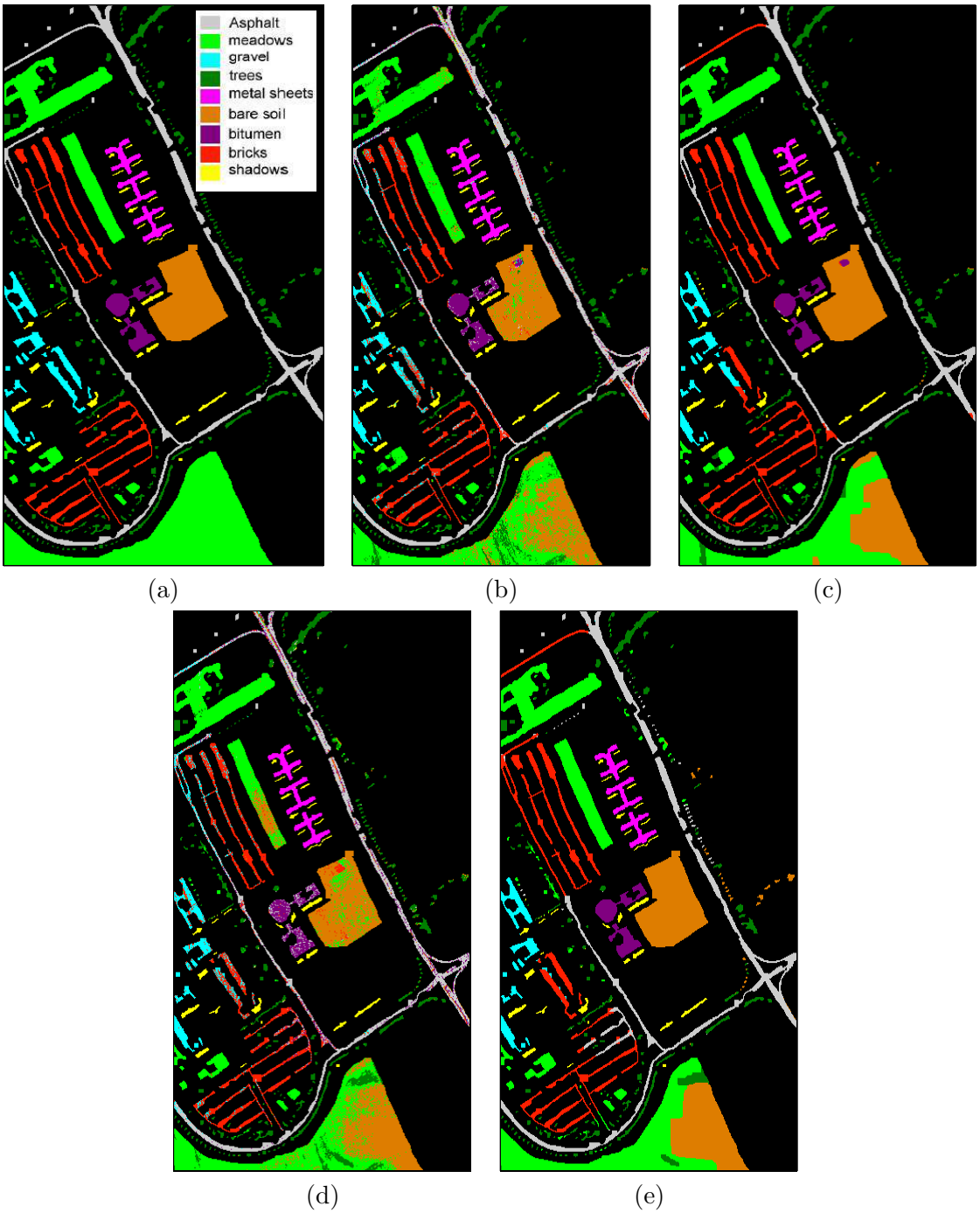


Figure 3.8: Classification and segmentation maps obtained for the ROSIS subset #2 by using the whole training set (L=3921). (a) Ground truth. (b) LORSAL, OA=80.24%. (c) LORSAL-MLL, OA=86.72%. (d) LDA, OA=73.45%. (e) LDA-MLL, OA=80.67%.

Table 3.1: Algorithms tested with each considered hyperspectral data set, where classification algorithms only use the spectral information and segmentation algorithms integrate both spectral and spatial information. The number of features extracted prior to classification are given in the parentheses.

Algorithm		Feature extraction	Indian Pines	Subset #1	Subset #2	Subset #3
Classification	LORSAL-AL	No	Yes	Yes	Yes	Yes
	LDA-AL	HySime	Yes (12)	Yes (5)	Yes (7)	No
	SVM	No	Yes	Yes	No	No
	PCA+SVM	PCA	Yes (31)	Yes (30)	No	No
Segmentation	LORSAL-AL-MLL	No	No	No	Yes	Yes
	LDA-AL-MLL	HySime	No	No	Yes (7)	No

Table 3.2: Parameter settings in our experiments with real hyperspectral data sets. For Subset #1, we only run classification experiments therefore no μ is used.

Dataset	Indian Pines	Subset #1	Subset #2	Subset #3
λ	0.001	0.001	0.001	0.001
μ	4	-	2	1

In our experiments, we compare our proposed approach with LDA [13] and SVMs [28], using feature extraction based on PCA [49] and hyperspectral signal identification by minimum error (HySime) [11]. This is because LDA requires that the number of labeled samples be larger than the dimensionality of the input features. In the case of SVM, we use PCA for feature extraction, as it is common practice in other studies; whereas in the case of LDA, we use HySime as different feature extraction strategy which efficiently estimates the subspace. In summary, Table 3.1 shows the different classification and segmentation algorithms considered in our real data experiments, where LDA-AL and LDA-AL-MLL integrate the standard LDA classifier and MLL spatial prior with the proposed active learning approaches. We would also like to emphasize that, in the real image experiments, no cross-validation is performed. Table 3.2 shows the parameter used for each data set. Although these parameter settings may be sub-optimal, we have experimentally tested that they lead to good results for each classifier as it will be shown in experiments. Finally, it is also worth noting that, in all experiments, all considered algorithms use exactly the same training sets when there is no active sampling strategy applied. Also, they all share the same initial training sets when active sampling is considered.

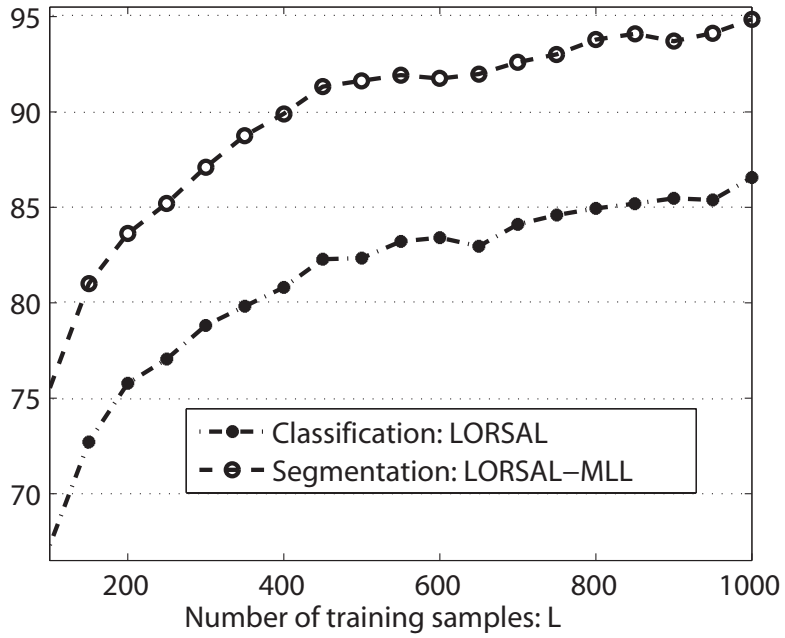


Figure 3.9: OA results as a function of the number of labeled samples for the AVIRIS Indian Pines data set.

Experiment 1: AVIRIS Indian Pines Data Set

Our first experiment with the AVIRIS Indian Pines data set is intended to illustrate the contribution of the spatial prior. For this purpose, Figure 3.9 plots the obtained OA results as a function of the number of labeled samples after 10 Monte Carlo runs (without active sampling). Here, the training samples are randomly selected from the original training set. From the results reported in Figure 3.9 we can observe that, by including the spatial prior, the LORSAL-MLL algorithm greatly improves the classification results obtained by the LORSAL algorithm which only uses the spectral information.

In a second experiment, we evaluate the performance of the proposed MLR-based classification algorithms by using training sets made up of 5% (237 samples), 10% (475 samples) and 25% (1189 samples) of the original training data. Table 3.3 shows the classification results obtained after 10 Monte Carlo runs, along with those provided by SVMs and LDA. From Table 3.3, it can be observed that the proposed MLR-based algorithms obtain good results when compared to other methods. As expected, the proposed active learning procedure improves the learning results. For illustrative purposes, the effectiveness of the proposed method with the AVIRIS Indian Pines scene is further illustrated in Figure 3.7 in which classification maps obtained are displayed along with their associated OA scores.

Table 3.3: OA [%] and κ statistic (in the parentheses) obtained with the proposed algorithm (using different sampling schemes) as a function of the number of labeled samples for the AVIRIS Indian Pines data set. For comparative purposes, results with LDA and SVMs (with and without PCA-based feature extraction) are also included.

Training set			LORSAL-AL				LDA-AL				SVMs	PCA+SVM
L	L_i	u	RS	MI	BT	MBT	RS	MI	BT	MBT	RS	RS
237	117	30	80.65 (0.77)	81.56 (0.78)	82.60 (0.80)	82.80 (0.79)	64.88 (0.59)	66.34 (0.61)	66.14 (0.60)	66.22 (0.59)	74.42 (0.70)	71.30 (0.67)
475	235	60	84.56 (0.82)	87.28 (0.85)	87.54 (0.85)	87.35 (0.84)	69.63 (0.65)	71.97 (0.67)	71.68 (0.67)	70.65 (0.64)	80.06 (0.77)	78.36 (0.74)
1189	597	148	88.45 (0.87)	91.31 (0.90)	91.37 (0.90)	90.56 (0.89)	73.29 (0.69)	75.43 (0.71)	76.05 (0.72)	76.01 (0.69)	86.96 (0.85)	84.62 (0.81)

Table 3.4: OA [%] and κ statistic (in the parentheses) for the ROSIS subset #1, where $L^{(k)}$ denotes the number of labeled samples per class.

	$L^{(k)}$	10	20	40	60	80	100
Training set	L_i	45	90	180	270	360	450
per class	u	9	18	36	54	72	90
LORSAL-AL	RS	95.13 (0.92)	96.29 (0.94)	96.91 (0.95)	97.07 (0.95)	97.37 (0.95)	97.49 (0.96)
	MBT	96.14 (0.93)	96.74 (0.94)	97.34 (0.95)	97.67 (0.96)	97.87 (0.96)	97.95 (0.96)
LDA-AL	RS	93.55 (0.89)	95.59 (0.92)	96.20 (0.93)	96.35 (0.94)	96.33 (0.94)	96.29 (0.94)
	MBT	95.10 (0.92)	96.34 (0.94)	96.76 (0.94)	97.02 (0.95)	96.97 (0.95)	97.03 (0.95)
SVM	RS	93.34 (0.89)	94.45 (0.91)	94.68 (0.91)	94.93 (0.91)	95.35 (0.92)	96.19 (0.94)
PCA+SVM	RS	85.57 (0.76)	91.20 (0.85)	94.79 (0.91)	95.68 (0.93)	96.30 (0.94)	96.37 (0.94)

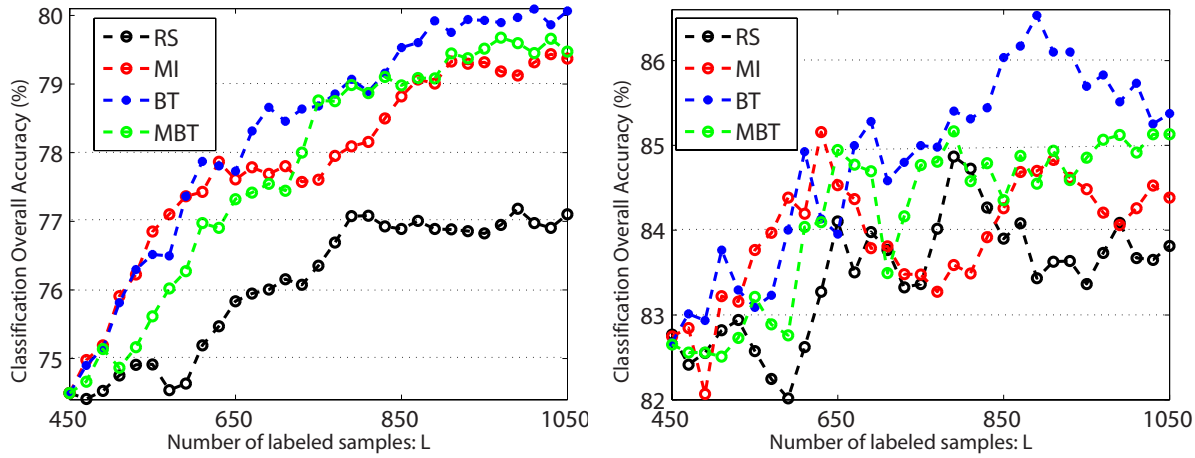
Experiment 2: ROSIS Pavia Data Sets

In this section, the three considered subsets of the ROSIS Pavia data are used to evaluate the proposed approach. The first experiment uses the ROSIS Pavia Data subset #1. In this experiment, we use small training sets, *i.e.*, $L^{(k)} = \{10, 20, 40, 60, 80, 100\}$ samples per class. Concerning the active learning approach, we focus on the MBT method as it provides the flexibility of selecting a given number of new samples per class at each iteration. Table 3.4 summarizes the results obtained after 10 Monte Carlo runs by the considered classification algorithms in comparison with the same standard methods used for reference in the previous subsection. We emphasize the good classification performance achieved by the proposed LORSAL and LORSAL-AL algorithms. Moreover, Table 3.4 reveals that the MBT sampling procedure further improves the OA results and the κ statistic.

In our second experiment, we use subset #2 of the Pavia ROSIS data to evaluate the proposed segmentation algorithm. Table 3.5 illustrates the OA results obtained after 10 Monte Carlo runs, by using the entire training set. Notice the good performances achieved by the proposed

Table 3.5: OA [%] and κ statistic (in the parentheses) obtained for the ROSIS Pavia subset #2.

L	LORSAL	LORSAL-MLL	LDA	LDA-MLL
3921	80.24 (0.76)	86.72 (0.82)	73.45 (0.67)	80.67 (0.76)



(a) LORSAL-AL results.

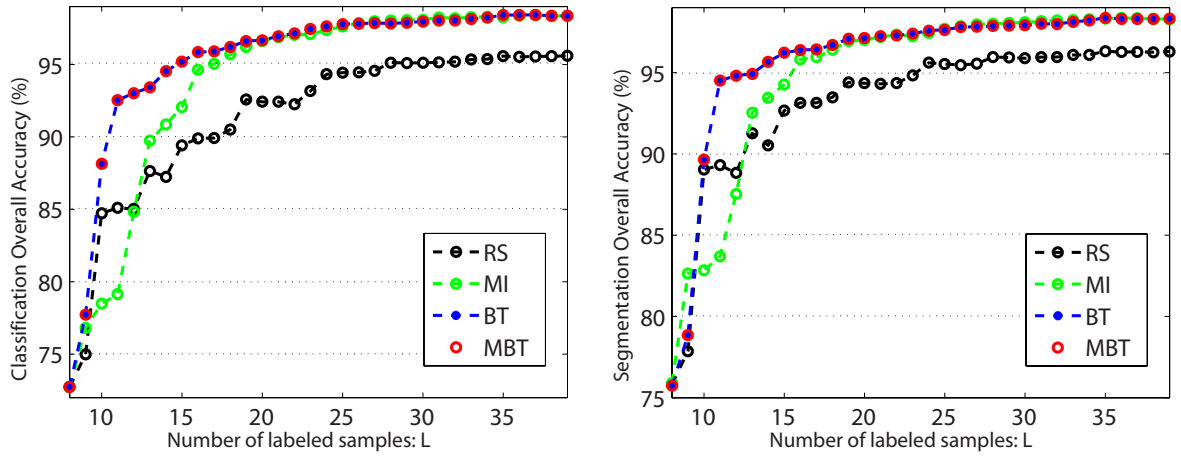
(b) LORSAL-AL-MLL results

Figure 3.10: OA [%] results as a function of the number of labeled samples for the ROSIS subset #2.

LORSAL and LORSAL-MLL algorithms (see Table 3.5), where the segmentation result obtained by the LORSAL-MLL algorithm is comparable to that reported in previous work for an SVM classifier using extended morphological profiles as input features in [107]. Although a more exhaustive comparison between these approaches should be conducted using the same training and test sets, we believe that the fact that our method provides comparable results to a highly consolidated technique that integrates the spatial and the spectral information is remarkable.

Furthermore, we also evaluate the sensitivity of the proposed AL-based approaches to the size of the considered training set by using subsets of the original training set. Figure 3.10 shows the OA results as a function of L, with $L_i = 450$ and $u = 20$. From Figure 3.10, it can be observed that the LORSAL-AL and LORSAL-AL-MLL algorithms achieve significant improvements as compared with the standard RS strategy. Finally, it is also worth noting that the integration of spatial and spectral information significantly improves the classification results obtained using spectral information only.

In our final experiment, we consider subset #3 of the Pavia ROSIS data to evaluate the proposed LORSAL-AL and LORSAL-AL-MLL algorithms by using $L_i = 8$ (only 1 sample per class) and $u = 1$. In this experiment, we do not consider the LDA-AL and LDA-AL-MLL algorithms because the LDA model requires a number of training samples which is larger than



(a) LORSAL-AL results. (b) LORSAL-AL-MLL results

Figure 3.11: OA results as a function of the number of labeled samples for the ROSIS subset #3.

the dimensionality of the feature space. Figure 3.11 illustrates the OA results (as a function of L) in this challenging scenario. The good performance achieved by the proposed LORSAL-AL and LORSAL-AL-MLL algorithms in this analysis scenario is remarkable where, as expected, the BT and MBT methods lead to similar estimates for the considered problem. Furthermore, the contribution of the spatial prior is less relevant as the value of L increases. As shown by Figure 3.11, the AL further improves the learning results and, eventually, MI, BT and MBT converge to very similar OA results. For illustrative purposes, Figure 3.8 displays the classification and segmentation maps obtained by the considered algorithm configurations (in comparison with other methods) using the ROSIS Pavia University data set.

3.6 Conclusions

In this paper, we have developed a new (supervised) Bayesian segmentation approach aimed at addressing ill-posed hyperspectral classification and segmentation problems. The proposed algorithm models the posterior class probability distributions using the concept of multinomial logistic regression (MLR), where the MLR regressors are learnt by the logistic regression via splitting and augmented Lagrangian (LORSAL) algorithm. The algorithm adopts a multi-level logistic (MLL) prior to model the spatial information present the class label images. The maximum *a posteriori* (MAP) segmentation is efficiently computed by the α-Expansion graph-cut based algorithm. The resulting segmentation algorithm (LORSAL-MLL) greatly improves the overall accuracies with respect to the classification results just based on the learnt class distribution. Another contribution of this work is the incorporation of active learning strategies

in order to cope with training sets containing a very limited number of samples. Three different sampling approaches, namely: a mutual information (MI)-based criterion, a breaking ties (BT) strategy, and a newly developed method called modified breaking ties (MBT) are integrated in the developed classification (LORSAL) and segmentation (LORSAL-MLL) methods, resulting in two new methods with active learning respectively called LORSAL-AL and LORSAL-MLL-AL. The effectiveness of the proposed algorithms is illustrated in this work using both simulated and real hyperspectral datasets. A comparison with state-of-the-art methods indicates that the proposed approaches yield comparable or superior performance using fewer labeled samples. Moreover, our experimental results reveal that the proposed MBT approach leads to an unbiased sampling as opposed to the MI and BT strategies. Further work will be directed towards testing the proposed approach in other different analysis scenarios dominated by the limited availability of training samples.

Appendix

The problem described in Eq. (3.5) is equivalent to:

$$\begin{aligned}(\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\nu}}) \quad &= \quad \arg\min_{\boldsymbol{\omega},\, \boldsymbol{\nu}} \; -\ell(\boldsymbol{\omega}) + \lambda\|\boldsymbol{\nu}\|_1 \\ &\qquad \text{subject to:} \quad \boldsymbol{\omega} = \boldsymbol{\nu}.\end{aligned} \tag{3.20}$$

By applying the alternating direction method of multipliers (ADMM) [51] (see also [1] and references therein) to solve the problem in Eq. (3.20), we get the iterative Algorithm 3.4. In this algorithm, $\beta \geq 0$ sets the augmented Lagrangian weight. Under mild conditions, the sequence $\widehat{\boldsymbol{\omega}}^t$, for $t = 0, 1, 2 \ldots$ converges to a minimizer of Eq. (3.20), for any $\beta \geq 0$ [51].

Algorithm 3.4 Logistic regression via variable splitting and augmented Lagrangian (LORSAL)

Require: $\boldsymbol{\omega}^{(0)}$, $\boldsymbol{\nu}^{(0)}$, $\mathbf{b}^{(0)}$, λ, β

1: t := 0
2: **repeat**
3: $\quad \widehat{\boldsymbol{\omega}}^{(t+1)} \in \arg\min_{\boldsymbol{\omega}} \; -\ell(\boldsymbol{\omega}) + \frac{\beta}{2}\|\boldsymbol{\omega} - \boldsymbol{\nu}^{(t)} - \mathbf{b}^{(t)}\|^2$ (3.21)
4: $\quad \widehat{\boldsymbol{\nu}}^{(t+1)} \in \arg\min_{\boldsymbol{\nu}} \; \lambda\|\boldsymbol{\nu}\|_1 + \frac{\beta}{2}\|\boldsymbol{\omega}^{(t+1)} - \boldsymbol{\nu} - \mathbf{b}^{(t)}\|^2$ (3.22)
5: $\quad \mathbf{b}^{(t+1)} := \mathbf{b}^{(t)} - \boldsymbol{\omega}^{(t+1)} + \boldsymbol{\nu}^{(t+1)}$
6: $\quad t := t + 1$
7: **until** some stopping criterion is met

It should be noted that the solution of the optimization problem in Eq. (3.21) (line 3 of Algorithm 3.4) is still a difficult problem because $\ell(\boldsymbol{\omega})$, although strictly convex and smooth, is non-quadratic and often very large. We tackle this difficulty by replacing $\ell(\boldsymbol{\omega})$ with a quadratic

lower bound given by [16]:

$$\ell(\boldsymbol{\omega}) \geq \ell(\boldsymbol{\omega}^{(t)}) + (\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^T \mathbf{g}(\boldsymbol{\omega}^{(t)}) + \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^T \mathbf{B}(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)}), \tag{3.23}$$

where $\mathbf{B} \equiv -(1/2)[\mathbf{I} - \mathbf{1}\mathbf{1}^T/K] \otimes \sum_{i=1}^{L} \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)^T$ (symbol $\mathbf{1}$ denotes a vector column of ones) and $\mathbf{g}(\boldsymbol{\omega}^{(t)})$ is the gradient of ℓ at $\boldsymbol{\omega}^{(t)}$. Since the system matrix involved in the optimization of Eq. (3.23), with $\ell(\boldsymbol{\omega})$ replaced with the quadratic bound given in Eq. (3.23) is fixed, its inverse can be pre-computed, provided that γ –the dimension of $\mathbf{h}(\mathbf{x}_i)$– is below, say, a few thousands. Under mild conditions, the convergence of Algorithm 3.4 with the aforementioned modification still holds [1, 51].

On the other hand, the solution of the optimization problem in Eq. (3.22) (line 4 of Algorithm 3.4) is simply the soft-threshold rule [45] given by $\widehat{\boldsymbol{\nu}}^{(t+1)} = \max\{\mathbf{0}, \mathrm{abs}(\mathbf{u})\}\mathrm{signal}(\mathbf{u})$, where $\mathbf{u} \equiv (\boldsymbol{\omega}^{(t+1)} - \mathbf{b}^{(t)}) - \lambda/\beta$ and the involved functions are to be understood component-wise. As a final note, we reiterate that the complexity of each iteration of the LORSAL algorithm is $O(\gamma^2 K)$, which is must faster than $O((\gamma K)^3)$ for the SMLR algorithm [80], and $O(\gamma^3 K)$ for the FSMLR algorithm [18].

Chapter 4

Spectral-Spatial Hyperspectral Image Segmentation Using Subspace Multinomial Logistic Regression and Markov Random Fields

Abstract – This paper introduces a new spectral-spatial supervised Bayesian segmentation algorithm for highly mixed hyperspectral images which exploits the contributions of both of the spectral and spatial information. The posterior probability distributions are learnt by using a multinomial logistic regression model (MLR), which uses the subspace method to circumvent the mixture of spectral signatures, thus to exploit the wealth of the spectral information. The contexture spatial information is modeled by a Markov random field (MRF) multi-level logistic (MLL) Markov-Gibbs prior. Finally, the maximum a posteriori segmentation (MAP) is efficiently computed by the α-Expansion min-cut based integer optimization algorithm. State-of-the-art performance of the proposed approach is illustrated using both simulated and real hyperspectral data sets in a number of experimental comparisons with recently introduced hyperspectral image classification methods.

Index Terms – Hyperspectral, Subspace Method, Segmentation, MRF

4.1 Introduction

Supervised classification (and segmentation) of high dimensional datasets such as remotely sensed hyperspectral images is a difficult endeavor [85]. Obstacles, such as the Hughes phenomenon [68], appear as the data dimensionality increases. This is because the number of training samples used for the learning stage of the classifier is generally very limited compared to the number of available spectral bands. In order to overcome this difficulty, several feature selection [123] and extraction [114] methods have been combined with machine learning tech-

niques able to perform accurately in the presence of limited training sets, including support vector machines (SVMs) [28, 121] or multinomial logistic regression (MLR) classifiers [16, 91].

Due to sensor design considerations, the wealth of spectral information in hyperspectral data is often not complemented by extremely fine spatial resolution. This leads to the presence of mixed pixels, which represent a challenge for accurate hyperspectral image classification [107]. In order to address this issue, subspace projection methods [136] have been shown to be a powerful class of statistical pattern classification algorithms [104]. These methods can handle the high dimensionality of an input data set by bringing it to the right subspace without losing the original information that allows for the separation of classes. In some cases, these methods can also reduce noise and, subsequently, they can reduce the impact of mixed pixels in the classification process. This is because noise can lead to confusion between spectrally similar classes resulting from a predominance of mixed pixels, as it is indeed the case in some of the most widely used data sets in the hyperspectral image classification community such as the famous Indian Pines image [85], collected by NASA Jet Propulsion Laboratory's Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS) [62]. In this scene, pixels in different classes exhibit spectrally similar signatures due to the early growth cycle of the agricultural features, which barely cover the soil in a proportion of 5% or less. Since the spatial resolution in this case is 20 meters per pixel, the scene is dominated by mixed pixels made up of different agricultural features and soil. However, the reference ground-truth widely used in the hyperspectral classification community associates image pixels with hard, mutually exclusive class labels. In this context, subspace projection methods can provide competitive advantages by separating classes which are very similar in spectral sense, thus addressing the limitations due to highly mixed pixels.

The idea of applying subspace projection methods to improve classification relies on the basic assumption that the samples within each class can mostly lie in a lower dimensional subspace. Thus, each class may be represented by a subspace spanned by a set of basis vectors, while the classification criterion for a new input sample would be the distance from the class subspace [57, 102, 134]. Recently, several subspace projection methods have been specifically designed for improving hyperspectral data characterization [3, 31, 123, 137], obtaining successful results. A more recent trend towards increasing classification accuracies in hyperspectral image analysis is to make combined use of spectral and spatial-contextual information [19, 53, 90, 91, 107, 129]. In some of these works, Markov random fields (MRF) have obtained great success in characterizing spatial information in hyperspectral data sets. MRFs exploit the continuity, in probability sense, of neighboring labels. The basic assumption is that, in a hyperspectral image, it is very likely that two neighboring pixels will have the class same label. Despite the fact that spatial information

plays a very important role in hyperspectral image classification and segmentation, to the best of our knowledge no previous work has proposed a technique that simultaneously combines the advantages that can be provided by subspace projection-based classifiers (in spectral sense) and spatial-contextual information.

In this work, we propose a new Bayesian approach to hyperspectral image segmentation which combines spectral and spatial information. The algorithm implements two main steps: (i) Learning, where the posterior probability distributions are modeled by an MLR combined with a subspace projection method, and (ii) Segmentation, which infers an image of class labels from a posterior distribution built on the learnt subspace classifier, and on a multi-level logistic (MLL) prior on the image of labels. The final output of the algorithm is based on a maximum a posteriori (MAP) segmentation process which is computed via an efficient min-cut based integer optimization technique. The main novelty of our proposed work is the integration of a subspace projection method with the MLR and further combined with spatial-contextual information, which provides a better characterization of the hyperspectral image content in both the spectral and the spatial domains. As will be shown by our experimental results, the accuracies achieved by our approach are competitive or superior to those provided by many other state-of-the-art supervised classifiers for hyperspectral image analysis.

The remainder of the paper is organized as follows. Section 4.2 formulates the problem. Section 4.3 describes the proposed approach. Section 4.4 reports segmentation results based on simulated and real hyperspectral datasets in comparison with other state-of-the-art supervised classifiers. Finally, Section 4.5 concludes with some remarks and hints at plausible future research lines.

4.2 Problem formulation

Before describing our proposed approach, let us first define some of the notations that will be used throughout the paper:

$\mathcal{S} \equiv \{1, \ldots, n\}$	Set of integers indexing the n pixels of an image;
$\mathcal{K} \equiv \{1, \ldots, K\}$	Set of K classes;
$\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$	Image in which the pixels are d-dimensional vectors;
$\mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{L}^n$	Image of labels;
$\mathcal{D}_{l^{(k)}}^{(k)} \equiv \{(y_1, \mathbf{x}_1), \ldots, (y_{l^{(k)}}, \mathbf{x}_{l^{(k)}})\}$	Set of labeled samples for class k with size $l^{(k)}$;
$\mathbf{x}_{l^{(k)}}^{(k)} \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_{l^{(k)}}\}$	Set of feature vectors in $\mathcal{D}_{l^{(k)}}^{(k)}$;
$\mathcal{D}_l \equiv \{\mathcal{D}_{l^{(1)}}^{(1)}, \ldots, \mathcal{D}_{l^{(K)}}^{(K)}\}$	Set of labeled samples with size $l = \sum_{k=1}^{K} l^{(k)}$.

With the above definitions in place, the goal of classification is to assign a label $y_i \in \mathcal{K}$ to each pixel vector $\mathbf{x}_i$, with $i \in \mathcal{S}$. This process results in an image of class labels $\mathbf{y}$, and we call this assignment a *labeling*. In turn, the goal of segmentation is to partition the set $\mathcal{S}$ such that the pixels in each subset $\mathcal{S}_k$, with $\mathcal{S} = \cup_k \mathcal{S}_k$, share some common property, *e.g.* they represent the same type of land cover. Notice that, given a labeling $\mathbf{y}$, the collection $\mathcal{S}_k = \{i \in \mathcal{S} \,|\, y_i = k\}$ for $k \in \mathcal{K}$ is a partition of $\mathcal{S}$. On the other hand, given the segmentation $\mathcal{S}_k$ for $k \in \mathcal{K}$, the image $\{y_i \,|\, y_i = k, \text{ if } i \in \mathcal{S}_k, \ i \in \mathcal{S}\}$ is a labeling. As a result, there is a one-to-one relationship between labelings and segmentations. Without loss of generality, in this paper we use the term *classification* when the spatial information in the original scene is not used in the labeling process. Similarly, we use the term *segmentation* when the spatial information in the original scene is used for such labeling.

In a Bayesian framework, the labeling process is usually conducted by maximizing the posterior distribution as follows:

$$p(\mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y})p(\mathbf{y}), \tag{4.1}$$

where $p(\mathbf{x}|\mathbf{y})$ is the likelihood function (*i.e.*, the probability of the feature image given the labels) and $p(\mathbf{y})$ is the prior over the image of labels. Assuming conditional independency of the features given the labels, *i.e*, $p(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{i=n} p(\mathbf{x}_i|y_i)$, then the posterior may be written as a function of $\mathbf{y}$ as follows:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \frac{1}{p(\mathbf{x})} p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \\ &= \frac{1}{p(\mathbf{x})} \prod_{i=1}^{i=n} p(\mathbf{x}_i|y_i)p(\mathbf{y}) \\ &= \alpha(\mathbf{x}) \prod_{i=1}^{i=n} \frac{p(y_i|\mathbf{x}_i)}{p(y_i)} p(\mathbf{y}), \end{aligned} \tag{4.2}$$

where $\alpha(\mathbf{x}) \equiv \prod_{i=1}^{i=n} p(\mathbf{x}_i)/p(\mathbf{x})$ is a factor not depending on $\mathbf{y}$. In the proposed approach, we assume the classes are equally likely, *i.e.*, $p(y_i = k) = 1/K$ for any $k \in \mathcal{K}$. However, any other distribution can be accommodated, as long as the marginal of $p(\mathbf{y})$ is compatible with the assumed distribution. Therefore, the maximum a posteriori (MAP) segmentation is given by:

$$\widehat{\mathbf{y}} = \arg\max_{y \in \mathcal{K}^n} \left\{ \sum_{i=1}^{n} (\log p(y_i|\mathbf{x}_i) + \log p(\mathbf{y}) \right\}. \tag{4.3}$$

Following the Bayesian framework described above, we have developed a new algorithm which naturally integrates the spectral and the spatial information contained in the original hyperspectral image data. In our proposed algorithm, the spectral information is represented by class densities $p(y_i|\mathbf{x}_i)$, which are learnt by a subspace projection-based MLR algorithm.

On the other hand, the spatial prior $p(\mathbf{y})$ is given by an MRF-based MLL which encourages neighboring pixels to have the same label. The MAP segmentation $\widehat{\mathbf{y}}$ is computed via the α-Expansion algorithm [23], a min-cut based tool to efficiently solve integer optimization problems. Additional details are given in the following section.

4.3 Proposed approach

Under the linear mixture model assumption, for any $i \in \mathcal{S}$ we have:

$$\mathbf{x}_i = \mathbf{m}\boldsymbol{\gamma}_i + \mathbf{n}_i, \tag{4.4}$$

where $\mathbf{m} \equiv [\mathbf{m}^{(1)}, \ldots, \mathbf{m}^{(K)}]$ denotes a mixing matrix composed by the spectral endmembers, $\mathbf{n}_i$ denotes the noise, and $\boldsymbol{\gamma}_i = [\gamma_i^{(1)}, \ldots, \gamma_i^{(K)}]^T$ denotes the fractional abundances of the endmembers in the mixed pixel $\mathbf{x}_i$. Since the distributions $p(\mathbf{m})$ and $p(\boldsymbol{\gamma}_i)$ are unknown, it is very difficult to compute $p(\mathbf{x}_i|y_i = k)$ using a generative model. It happens, however, that the linear term $\mathbf{m}\boldsymbol{\gamma}_i$ in (4.4) lives in class dependent subspaces. This is consequence of the linearity of this term and of the fact the set of materials corresponding to any two different classes are very likely to be different. With this simple fact in mind, we may then write the observation mechanism for class k as

$$\mathbf{x}_i^{(k)} = \mathbf{U}^{(k)}\mathbf{z}_i^{(k)} + \mathbf{n}_i^{(k)}, \tag{4.5}$$

where $\mathbf{n}_i^{(k)}$ is the noise of class k and $\mathbf{U}^{(k)} = \{\mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_{r^{(k)}}^{(k)}\}$ is a set of $r^{(k)}$-dimensional orthonormal basis vectors for the subspace associated with class k, and $\mathbf{z}_i^{(k)}$ is, apart from the noise $\mathbf{n}_i^{(k)}$ the coordinates of $\mathbf{x}_i^{(k)}$ with respect to the basis $\mathbf{U}^{(k)}$.

We assume that the class independent random vectors $\mathbf{n}_i^{(k)}$ and $\mathbf{z}_i^{(k)}$ are Gaussian distributed with zero mean and diagonal covariance matrices, *i.e.*, $\mathbf{n}_i^{(k)} \sim \mathcal{N}(0, {\sigma^{(k)}}^2\mathbf{I})$, and $\mathbf{z}_i^{(k)} \sim \mathcal{N}(0, \alpha^{(k)}\mathbf{I})$. We are aware that these assumptions are very strong and that they rarely hold in real data. However, and shown below, they allow to preserve the subspace structure of our model and yield a robust discriminative model. Based on the above assumptions, we have the following generative model:

$$p(\mathbf{x}_i|y_i = k) \sim \mathcal{N}(0, \alpha^{(k)}\mathbf{U}^{(k)}\mathbf{U}^{(k)^T} + {\sigma^{(k)}}^2\mathbf{I}). \tag{4.6}$$

Under the present setup, the generative model in Eq. (4.6) can be computed as follows:

$$\begin{aligned} p(\mathbf{x}_i|y_i=k) &\propto \exp\left\{-\tfrac{1}{2}\mathbf{x}_i^T(\alpha^{(k)}\mathbf{U}^{(k)}\mathbf{U}^{(k)^T}+\sigma^{(k)^2}\mathbf{I})^{-1}\mathbf{x}_i\right\} \\ &= \exp\left\{-\tfrac{1}{2}\mathbf{x}_i^T\left(\tfrac{\mathbf{I}}{\sigma^{(k)^2}}-\tfrac{\mathbf{U}^{(k)}}{\sigma^{(k)^2}}\left(\tfrac{\mathbf{I}}{\alpha^{(k)}}+\tfrac{\mathbf{U}^{(k)^T}\mathbf{U}^{(k)}}{\sigma^{(k)^2}}\right)^{-1}\tfrac{\mathbf{U}^{(k)^T}}{\sigma^{(k)^2}}\right)\mathbf{x}_i\right\} \\ &= \exp\left\{-\tfrac{1}{2}\mathbf{x}_i^T\left(\tfrac{\mathbf{I}}{\sigma^{(k)^2}}-\tfrac{\alpha^{(k)}}{\alpha^{(k)}+\sigma^{(k)^2}}\mathbf{U}^{(k)}\mathbf{U}^{(k)^T}\right)\mathbf{x}_i\right\} \\ &= \exp\left\{-\tfrac{1}{2}\tfrac{\mathbf{x}_i^T\mathbf{x}}{\sigma^{(k)^2}}+\tfrac{1}{2}\tfrac{\alpha^{(k)}}{\alpha^{(k)}+\sigma^{(k)^2}}\|\mathbf{x}_i^T\mathbf{U}^{(k)}\|^2\right\} \end{aligned} \tag{4.7}$$

Let $\omega_1^{(k)} \equiv -\frac{1}{2\sigma^{(k)^2}}$, $\omega_2^{(k)} \equiv \frac{1}{2}\frac{\alpha^{(k)}}{\alpha^{(k)}+\sigma^{(k)^2}}$, $\boldsymbol{\omega}^{(k)} \equiv [\omega_1^{(k)}\ \omega_2^{(k)}]^T$, and $\boldsymbol{\omega} \equiv [\boldsymbol{\omega}^{(1)^T},\ldots,\boldsymbol{\omega}^{(K)^T}]^T$. With these definitions in mind, we can compute the posterior class density $p(y_i|\mathbf{x}_i)$ as follows:

$$\begin{aligned} p(y_i=k|\mathbf{x}_i,\,\boldsymbol{\omega}) &= \frac{p(\mathbf{x}_i|y_i=k,\,\boldsymbol{\omega})p(y_i=k)}{\sum_{k=1}^K p(\mathbf{x}_i|y_i=k,\,\boldsymbol{\omega})p(y_i=k)} \\ &= \frac{\exp(\boldsymbol{\omega}^{(k)^T}\boldsymbol{\phi}^{(k)}(\mathbf{x}_i))p(y_i=k)}{\sum_{k=1}^K \exp(\boldsymbol{\omega}^{(k)^T}\boldsymbol{\phi}^{(k)}(\mathbf{x}_i))p(y_i=k)}, \end{aligned} \tag{4.8}$$

where $\boldsymbol{\phi}^{(k)}(\mathbf{x}_i) = [\|\mathbf{x}_i\|^2,\ \|\mathbf{x}_i^T\mathbf{U}^{(k)}\|^2]^T$. Assuming equiprobable classes, *i.e.*, $p(y_i=k)=1/K$, the problem in Eq. (4.8) turns to

$$p(y_i=k|\mathbf{x}_i,\,\boldsymbol{\omega}) = \frac{\exp(\boldsymbol{\omega}^{(k)}\boldsymbol{\phi}^{(k)}(\mathbf{x}_i))}{\sum_{k=1}^K \exp(\boldsymbol{\omega}^{(k)}\boldsymbol{\phi}^{(k)}(\mathbf{x}_i))}, \tag{4.9}$$

which is exactly an MLR [16].

4.3.1 Learning the class independent subspace

Let $\mathbf{R}^{(k)} = \langle \mathbf{x}_{l^{(k)}}^{(k)}\mathbf{x}_{l^{(k)}}^{(k)^T}\rangle$ denote the sample correlation matrix associated with class k, computed from the training set. By computing the eigendecomposition of $\mathbf{R}^{(k)}$, we have

$$\mathbf{R}^{(k)} = \mathbf{E}^{(k)}\boldsymbol{\Lambda}^{(k)}\mathbf{E}^{(k)^T}, \tag{4.10}$$

where $\mathbf{E}^{(k)} = \{\mathbf{e}_1^{(k)},\ldots,\mathbf{e}_d^{(k)}\}$ is the eigenvector matrix and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1^{(k)},\ldots,\lambda_d^{(k)})$ is the eigenvalue matrix with decreasing magnitude *i.e.*, $\lambda_1^{(k)} \geq \cdots \geq \lambda_d^{(k)}$. Moreover, for $i \in \mathcal{S}$, vector $\mathbf{x}_i$ can be represented as a sum of two mutually orthogonal vectors $\mathbf{x}_i = \widehat{\mathbf{x}}_i + \tilde{\mathbf{x}}_i$, where $\widehat{\mathbf{x}}_i$ is the projection of vector $\mathbf{x}_i$ on the $r^{(k)}$-dimensional subspace spanned by the first $r^{(k)}$ eigenvalues, *i.e.*, $\lambda_1^{(k)},\ldots,\lambda_{r^{(k)}}^{(k)}$, and $\tilde{\mathbf{x}}_i$ is projection on the orthogonal subspace spanned by the remaining eigenvalues.

We take $\mathbf{U}_{r^{(k)}}^{(k)} = \{\mathbf{e}_1^{(k)},\ldots,\mathbf{e}_{r^{(k)}}^{(k)}\}$ as an estimate of the class independent, $r^{(k)}$-dimensional

subspace with $r^{(k)} < d$ and:

$$r^{(k)} = \min_{r^{(k)}} \{r^{(k)} : \sum_{i=1}^{r^{(k)}} \lambda_i^{(k)} \geq \sum_{i=1}^{d} \lambda_i^{(k)} \times \tau\}, \tag{4.11}$$

where $0 \leq \tau \leq 1$ is a threshold parameter controlling the loss of spectral information after projecting the data into the subspace.

4.3.2 Learning the MLR regressors

In order to cope with difficulties in learning the regression vector $\boldsymbol{\omega}$ associated with bad or ill conditioning of the underlying inverse problem, we adopt a quadratic prior on $\boldsymbol{\omega}$, so that:

$$p(\boldsymbol{\omega}) \propto e^{-\beta/2\|\boldsymbol{\omega}\|^2}, \tag{4.12}$$

where $\beta \geq 0$ is a regularization parameter controlling weight of the prior.

In the present problem, learning the class densities amounts to estimating the logistic regressors $\boldsymbol{\omega}$. Inspired by previous work [16, 19, 80, 90, 91], we can compute $\boldsymbol{\omega}$ by calculating the MAP estimate:

$$\widehat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \ell(\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}), \tag{4.13}$$

where $\ell(\boldsymbol{\omega})$ is the log-likelihood function given by:

$$\ell(\boldsymbol{\omega}) \equiv \log \prod_{i=1}^{l} p(y_i|\mathbf{x}_i, \boldsymbol{\omega}). \tag{4.14}$$

The optimization problem in Eq. (4.13) is convex, although the term $\ell(\boldsymbol{\omega})$ is non-quadratic. This term can be approximated by a quadratic lower bound given by [16]; for any $k \in \mathcal{K}$, we have:

$$\ell(\boldsymbol{\omega}^{(k)}) \geq \ell(\boldsymbol{\omega}_t^{(k)}) + (\boldsymbol{\omega}^{(k)} - \boldsymbol{\omega}_t^{(k)})^T \mathbf{g}(\boldsymbol{\omega}_t^{(k)}) + \frac{1}{2}(\boldsymbol{\omega}^{(k)} - \boldsymbol{\omega}_t^{(k)})^T \mathbf{B}^{(k)} (\boldsymbol{\omega}^{(k)} - \boldsymbol{\omega}_t^{(k)}), \tag{4.15}$$

with:

$$\mathbf{B}^{(k)} \equiv -(1/2)[\mathbf{I} - \mathbf{1}\mathbf{1}^T/(K+1)] \otimes \sum_{i=1}^{l} \boldsymbol{\phi}^{(k)}(\mathbf{x}_i) \boldsymbol{\phi}^{(k)}(\mathbf{x}_i)^T, \tag{4.16}$$

where $\mathbf{1}$ denotes a column vector of ones and $\mathbf{g}(\boldsymbol{\omega}_t^{(k)})$ is the gradient of $\ell(\cdot)$ at $\boldsymbol{\omega}_t^{(k)}$. Based on the lower bound (4.15), we implement a instance of the minorization maximization algorithm [69], which consists in replacing, in each iteration, the objective function $\ell(\boldsymbol{\omega})$ with the lower

bound (4.15) and then maximizing it. This procedure leads to

$$\widehat{\boldsymbol{\omega}}_{t+1}^{(k)} = \arg\max_{\boldsymbol{\omega}^{(k)}} \quad \boldsymbol{\omega}^{(k)^T}(\mathbf{g}(\widehat{\boldsymbol{\omega}}_t^{(k)}) - \mathbf{B}^{(k)}\widehat{\boldsymbol{\omega}}_t^{(k)}) + \frac{1}{2}\boldsymbol{\omega}^{(k)^T}(\mathbf{B}^{(k)} - \beta\mathbf{I})\boldsymbol{\omega}^{(k)}. \tag{4.17}$$

Now the optimization problem in Eq. (4.17) is quadratic and easy to solve, leading to the following update function:

$$\widehat{\boldsymbol{\omega}}_{t+1}^{(k)} = \tfrac{1}{2}(\mathbf{B}^{(k)} - \beta\mathbf{I})^{-1}(\mathbf{B}^{(k)}\widehat{\boldsymbol{\omega}}_t^{(k)} - \mathbf{g}(\widehat{\boldsymbol{\omega}}_t^{(k)})), \text{ for } k \in \mathcal{K}. \tag{4.18}$$

The system matrix in Eq. (4.18) is fixed, thus the term $(\mathbf{B}^{(k)} - \beta\mathbf{I})^{-1}$ can be pre-computed. With this in mind, it is now possible to perform an exact MAP-based MLR under a quadratic prior. The pseudo-code for the subspace projection-based MLR algorithm, referred to hereinafter as MLR*sub*, is shown in Algorithm 4.1. In the algorithm description, *iters* denotes the maximum number of iterations. The overall complexity of Algorithm 4.1 is dominated by the computation of the correlation matrix, which has complexity $O(ld^2)$ (recall that l is the number of labeled samples and d is the dimensionality of the feature vectors).

Algorithm 4.1 MLR*sub*

Input: $\boldsymbol{\omega}_0$, $\mathcal{D}_l$, β, τ, *iters*
Output: $\boldsymbol{\omega}$, $\mathbf{U} \equiv \{\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(k)}\}$
for $k = 1$ to K **do**
  $\mathbf{U}^{(k)} \equiv \mathcal{U}(\mathcal{X}_{l^{(k)}}^{(k)}, \tau)$ ($* \mathcal{U}$ computes the subspace according to Eq. (4.10) $*$)
  $\mathbf{B}^{(k)} \equiv \mathcal{B}(\mathbf{U}^{(k)}, \mathcal{D}_l)$ ($* \mathcal{B}$ computes the system matrix $\mathbf{B}$ according to Eq. (4.16) $*$)
end for
$t := 1$
while $t \leq$ *iters* **or** stopping criterion is not satisfied **do**
  for $k := 1$ to K **do**
    $\mathbf{g}(\boldsymbol{\omega}_{t-1}^{(k)}) \equiv \nabla\ell(\boldsymbol{\omega}_{t-1}^{(k)})$
    $\boldsymbol{\omega}_t^{(k)} = \text{solution}\,\{\mathbf{B}^{(k)}, \mathbf{g}(\boldsymbol{\omega}_{t-1}^{(k)}), \mathbf{U}^{(k)}, \beta\}$
  end for
end while

4.3.3 MRF-based MLL spatial prior

In order to improve the classification performance achieved by using the spectral information alone, in this work we integrate the contextual information with spectral information by using an isotropic MLL prior to model the image of class labels $\mathbf{y}$. This approach exploits the fact that, in segmenting real-world images, it is very likely that spatially neighboring pixels belong to the same class. This prior, which belongs to the MRF class, encourages piecewise smooth segmentations and promotes solutions in which adjacent pixels are likely to belong the same

class. The MLL prior constitutes a generalization of the Ising model [58] and has been widely used in image segmentation problems [92].

According to the Hammersly-Clifford theorem [10], the density associated with an MRF is a Gibbs's distribution [58]. Therefore, the prior model for segmentation has the following structure:

$$p(\mathbf{y}) = \frac{1}{Z}\, e^{\left(-\sum_{c\in\mathcal{C}} V_c(\mathbf{y})\right)}, \tag{4.19}$$

where Z is a normalizing constant for the density, the sum in the exponent is over the so-called prior potentials $V_c(\mathbf{y})$ for the set of cliques[1] $\mathcal{C}$ over the image, and:

$$-V_c(\mathbf{y}) = \begin{cases} \upsilon_{y_i}, & \text{if } |c| = 1 \text{ (single clique)} \\ \mu_c, & \text{if } |c| > 1 \text{ and } \forall_{i,\, j\in c} y_i = y_j \\ -\mu_c, & \text{if } |c| > 1 \text{ and } \exists_{i,\, j\in c} y_i \neq y_j \end{cases} \tag{4.20}$$

where μ_c is a non-negative constant. The potential function in Eq. (4.20) encourages neighbors to have the same label. The introduced MLL prior offers a great deal of flexibility by allowing variations of the set of cliques and the parameters υ_{y_i} and μ_c. For example, the model generates texture-like regions if μ_c depends on c, and blob-like regions otherwise [93]. In this work we take $\upsilon_{y_i} = c^{te}$ and $\mu_c = \frac{1}{2}\mu > 0$. Thus Eq. (4.19) can be rewritten as follows:

$$p(\mathbf{y}) = \frac{1}{Z}\, e^{\mu \sum_{(i,\, j)\in\mathcal{C}} \delta(y_i - y_j)}, \tag{4.21}$$

where $\delta(y)$ is the unit impulse function[2]. This choice gives no preference to any direction concerning. A straightforward computation of $p(y_i)$, *i.e.*, the marginal of $p(\mathbf{y})$ with respect to i, leads to $p(y_i)$ constant and thus equiprobable, therefore compatible with the assumption made in (4.2) and (4.8). Notice that the pairwise interaction terms $\delta(y_i - y_j)$ attach higher probability to equal neighboring labels than the other way around. In this way, the MLL prior promotes piecewise smooth segmentations, where parameter μ controls the level of smoothness.

4.3.4 MAP estimate via graph-cuts

Let us assume that the posterior class densities $p(y_i|\mathbf{x}_i)$ are estimated using Eq. (4.8). Let us also assume that the MLL prior $p(\mathbf{y})$ is estimated using Eq. (4.21). According to Eq. (4.3), the

[1] A clique is a single term or either a set of pixels that are neighbors of one another.
[2] *i.e.*, $\delta(0) = 1$ and $\delta(y) = 0$, for $y \neq 0$

MAP segmentation is finally given by:

$$\widehat{\mathbf{y}} \;=\; \arg\min_{\mathbf{y}\in\mathcal{K}^n} \left\{ \sum_{i\in\mathcal{S}} -\log p(y_i|\mathbf{x}_i,\, \widehat{\boldsymbol{\omega}}) - \mu \sum_{i\sim j} \delta(y_i - y_j) \right\}. \tag{4.22}$$

This is a combinatorial optimization problem involving unary and pairwise interaction terms, which is very difficult to compute. Several new algorithms such as graph-cuts [22, 23, 79], loopy belief propagation [141, 142], and tree-reweighted message passing [78] have been proposed in the literature in order to tackle this optimization problem. In this work, we resort to the α-Expansion graph-cut based algorithm [4, 23]. This method yields good approximations to the MAP segmentation and is quite efficient from a computational viewpoint, with computational complexity $O(n)$.

4.3.5 Supervised segmentation algorithm: MLR*sub*MLL

To conclude the description of our proposed method, Algorithm 4.2 provides a pseudo-code for our newly developed supervised segmentation algorithm based on a subspace MLR classifier with MRF-based MLL prior. This algorithm, called MLR*sub*MLL hereinafter, integrates all the different modules described in this section. Specifically, line 3 in Algorithm 4.2 learns the logistic regressors using MLR*sub*, which is applied to the full hyperspectral image. Here, the quadratic regularization parameter $\beta \geq 0$ is used to tackle ill-conditioned problems. Line 4 in Algorithm 4.2 computes the probabilities based on the outcome of MLR*sub*. Line 5 in Algorithm 4.2 efficiently computes the MAP segmentation by applying the α-Expansion graph-cut based algorithm, where the neighborhood parameter μ determines the strength of the spatial prior.

Algorithm 4.2 MLR*sub*MLL

1: **Input:** $\mathbf{x}$, $\mathcal{D}_l$, β, τ, μ
2: **Output:** $\widehat{\mathbf{y}}$
3: $\{\widehat{\boldsymbol{\omega}},\, \mathbf{U}\} = \text{MLR}sub\{\mathcal{D}_l,\, \beta,\, \tau\}$
4: $\widehat{\mathbf{P}} := \widehat{\mathbf{p}}(\mathbf{x},\, \widehat{\boldsymbol{\omega}},\, \mathbf{U})$ ($*$ $\widehat{\mathbf{P}}$ collects the probabilities in Eq. (4.9) $*$)
5: $\widehat{\mathbf{y}} := \alpha$-Expansion($\widehat{\mathbf{P}}$, μ, neighborhood)

The overall complexity of the proposed MLR*sub*MLL algorithm is dominated by the MLR*sub* algorithm inferring the regressors, which has computational complexity $O(ld^2)$, and also by the α-Expansion algorithm used to determine the MAP segmentation, which has practical complexity $O(n)$. In conclusion, if $ld^2 > n$ (*e.g.*, the problem is high dimensional, with a large number of training samples), then the overall complexity is dominated by the subspace-based learning step. Otherwise, if $ld^2 < n$ (*e.g.,* the problem is given by a large number of pixels), then the overall complexity is dominated by the α-Expansion algorithm.

4.4 Experimental results

This section uses both simulated and real hyperspectral data sets to illustrate the effectiveness of the proposed MLR*sub*MLL segmentation algorithm in different analysis scenarios. The main goal of using simulated data sets is to assess the performance of the algorithm in a fully controlled environment, whereas the main goal of using real data sets is to compare the algorithm with other state-of-the-art analysis techniques using widely used hyperspectral scenes. The remainder of this section is organized as follows. Subsection 4.4.1 first explains the parameter settings adopted in our experimental evaluation. Subsection 4.4.2 then evaluates the proposed MLR*sub*MLL algorithm by using simulated data sets, whereas Subsection 4.4.3 evaluates the proposed segmentation algorithm using real hyperspectral images.

4.4.1 Parameter settings

Before describing our results with simulated and real hyperspectral data sets, it is first important to discuss the parameter settings adopted in our experiments. In our tests we assume $l^{(k)} \simeq l/K$ for $k \in \mathcal{K}$. For small classes, if the total number of labeled samples per class k in the ground truth image, say $L^{(k)}$, is smaller than l/K, we take $l^{(k)} = L^{(k)}/2$. In this case, we use more labeled samples to represent large classes. It should be noted that, in all experiments, the labeled sets $\mathcal{D}_l$ are randomly selected from the available labeled samples, and the remaining samples are used for validation. Each value of overall accuracy (OA [%]) is obtained after conducting 10 Monte Carlo runs with respect to the labeled samples $\mathcal{D}_l$. The labeled samples for each Monte Carlo simulation are obtained by resampling the available labeled samples. Prior to the experiments, we infer the setting of the quadratic parameter β. In practice, β is relevant to the condition number of $\mathbf{B}^{(k)}$, for $k \in \mathcal{K}$. In this work, we set $\beta = e^{-10}$ for all experiments.

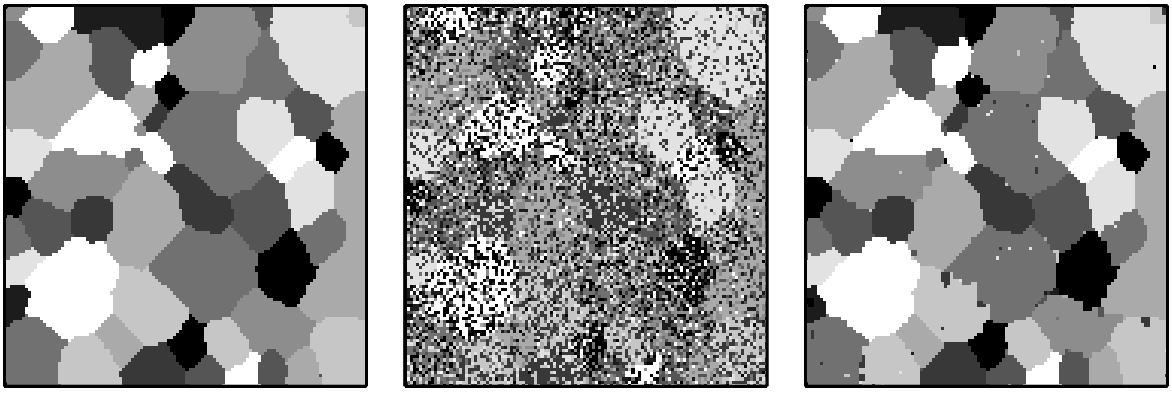
4.4.2 Experiments with simulated hyperspectral data

In our experiments we have generated a simulated hyperspectral scene as follows. First, we generate an image of features using a linear mixture model:

$$\mathbf{x}_i = \sum_{k=1}^{K} \mathbf{m}^{(k)} \gamma_i^{(k)} + \mathbf{n}_i, \tag{4.23}$$

with $K = 10$. Here, $\mathbf{m}^{(k)}$ for $k \in \mathcal{K}$ are spectral signatures obtained from the U.S. Geological Survey (USGS) digital spectral library[3], and $\mathbf{x}_i$ is a simulated mixed pixel. An MLL distribution with smoothness parameter $\mu = 2$ is used to generate the spatial information, and the total size

[3]The USGS library of spectral signatures is available online: http://speclab.cr.usgs.gov.



(a) Ground truth (b) Classification map (c) Segmentation map

Figure 4.1: Classification and segmentation maps obtained after applying the proposed method to a simulated hyperspectral scene with $\sigma = 0.8$ and $\gamma = 0.7$ by using $\tau = 0.9$, $l = 288$, $\mu = 2$. (a) Ground truth class labels. (b) Classification result (OA=49.09%). (c) Segmentation result (OA=94.34%).

of the simulated image is of 120×120 pixels. Zero-mean Gaussian noise with covariance $\sigma^2\mathbf{I}$, *i.e.*, $\mathbf{n}_i \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is finally added to our simple simulated hyperspectral scene. For illustrative purposes, the image of class labels $\mathbf{y}$ is shown in Figure 4.1(a). Assume that $\mathbf{x}_i$ has class label $y_i = k_k$, then we define $\gamma_i^{(k_k)}$ as the abundance of the objective class and $\gamma_i^{(k)}$ (for $k \in \mathcal{K}$ and $k \neq k_k$) as the abundance of the remaining signatures which contribute to the mixed pixel, where $\gamma_i^{(k)}$ are generated according to a simple uniform distribution in the proposed problem. In order to simplify notations, we take $\gamma_i^{(k_k)} = \gamma$ and $\sum_{k\in\mathcal{K},\, k\neq k_i} \gamma_i^{(k)} = 1 - \gamma$.

We have conducted five different experiments with the simulated hyperspectral image described above. These experiments have been carefully designed in order to analyze several relevant aspects of our proposed MLR*sub*MLL segmentation algorithm in a fully controlled environment:

1. In our first experiment, we evaluate the impact of the presence of mixed pixels on the segmentation output.

2. In our second experiment, we analyze the impact of the parameter τ (controlling the amout of spectral information retained after subspace projection) on the segmentation output.

3. In our third experiment, we evaluate the impact of the training set size on the segmentation output.

4. In our fourth experiment, we analyze the impact of the smoothness parameter μ on the segmentation output.

5. In our fifth experiment, we evaluate the impact of noise on the segmentation output.

In all these experiments we will use the optimal value of classification accuracy (OA_{opt}) as a reference to evaluate the goodness of our reported OA scores. Here, $\mathrm{OA}_{opt} \equiv 100\,(1 - \mathrm{P}_e)\%$, where P_e is defined as follows [49]:

$$\mathrm{P}_e \;=\; \sum_{k=1}^{K} \sum_{i \in \mathcal{K},\; i \neq k} p(y_j = k,\; j \in \mathcal{S}_k,\;\; \mathcal{K}_i), \tag{4.24}$$

where $\mathcal{K}_k$ denotes the k-th class. Eq. (4.24) is minimized when the regions $\mathcal{S}_k$ are chosen such that each $\mathbf{x}_j$ is assigned to the class for which $p(y_j = k,\; j \in \mathcal{S}_k,\;\; \mathcal{K}_i)$ is the smallest. For a multi-class problem, we use the following error bound as an alternative since Eq. (4.24) is difficult to compute:

$$\mathrm{erfc}\left(\frac{\mathrm{dist}_{\min}}{2\sigma}\right) \leq \mathrm{P}_e \leq \frac{K-1}{2}\mathrm{erfc}\left(\frac{\mathrm{dist}_{\min}}{2\sigma}\right), \tag{4.25}$$

where erfc denotes the complementary error function and $\mathrm{dist}_{\min}$ denotes the minimum distance between any point of mean vectors, *i.e.*, $\mathrm{dist}_{\min} = \min_{i \neq j} \|\mathbf{m}_i - \mathbf{m}_j\|$, for any $i, j \in \mathcal{K}$. This is the so-called union bound [57], which is widely used in multi-class problems. However, union bound is not a good measurement to present the difficulty because of the mixtures. Nevertheless, it is worthnothing that, as γ decreases, the difficulty increases, *i.e.*, OA_{opt} decreases. Thus, in this work we use the union bound, while $\gamma = 1$, to define the difficulty of our problem.

Experiment 1: Impact of the presence of mixed pixels

In this experiment we first consider a problem with $\sigma = 0.8$ by using $\tau = 0.9$, $\mu = 2$ and $\gamma \in [0.5 \;\; 1.0]$. In this context, the optimal value of classification accuracy is given by $\mathrm{OA}_{opt} \leq 71.04\%$ with $\gamma = 1$. It should be noted that the values of parameters τ and μ in our simulation are probably sub-optimal. However, we have decided to fix them to the specified values because we have experimentally observed that these settings lead to good performance in the considered analysis scenario. Figure 4.2 illustrates the obtained OA results as a function of γ (which determines the degree of spectral purity in the simulated pixels). In order to show the good capability of the proposed MLR*sub*MLL in the task of dealing with limited training sets, only 288 labeled samples (2% of the available samples, evenly distributed among classes) are used as the training set. Notice the good performance achieved by the proposed MLR*sub*MLL algorithm with the classes dominated by mixed pixels. In those classes, the segmentation results provided by MLR*sub*MLL significantly outperform the classification results obtained by the MLR*sub* using only the spectral information. For illustrative purposes, Figure 4.1 (b) and (c) shows the respective classification and segmentation maps obtained for the problem with $\sigma = 0.8$ and

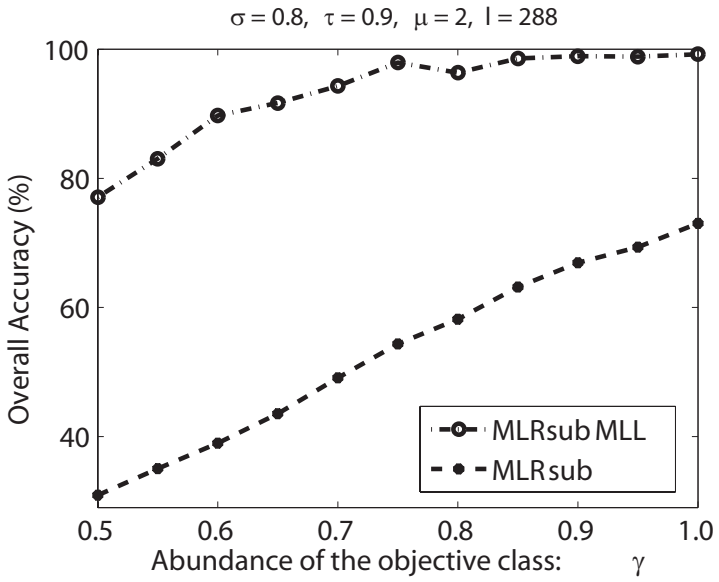


Figure 4.2: OA results as a function of the abundance of the objective class: γ, with $\tau = 0.9$, $\mu = 2$ and $l = 288$ for a problem with mixed pixels and $\sigma = 0.8$. Dash-dot lines with circles denote the segmentation results obtained by the MLR*sub*MLL algorithm; dashed lines with asterisks denote the classification results obtained by the MLR*sub* algorithm.

$\gamma = 0.7$, using $\tau = 0.9$, $\mu = 2$ and $l = 288$. Moreover, Figure 4.2 indicates that the performance of both MLR*sub*MLL and MLR*sub* increases as the abundance of the objective classes increase. This is expected, since the problem is easier to solve as the presence of mixed pixels is decreased. In the following experiments, we will consider $\gamma = 0.7$ which leads to a difficult segmentation problem as shown in Figure 4.2.

Experiment 2: Impact of parameter τ

In our second experiment, we analyze the impact of the threshold parameter τ intended to control the loss of spectral information after projecting the original hyperspectral data into a subspace. This parameter is directly related with the number of components retained after the projection, and with the amount of information comprised by the retained components. To address this issue, we analyze the performance of the proposed methods for different values of τ in a problem with $\sigma = 0.8$ ($\text{OA}_{opt} \leq 71.04\%$ with $\gamma = 1$) and $\gamma = 0.7$, by using $\mu = 2$. Figure 4.3 illustrates the OAs obtained by the proposed MLR*sub* and MLR*sub*MLL algorithms as a function of τ, where 288 labeled samples are again used as the (limited) training set. Notice the good performance achieved by the proposed MLR*sub*MLL segmentation algorithm, which yielded higher OA results than OA_{opt} in all cases. Furthermore, both classification and segmentation results increase as τ increases. This is reasonable since the amount of spectral information that is retained after the projection of the original data into the subspace is increased as τ increases. This also indicates

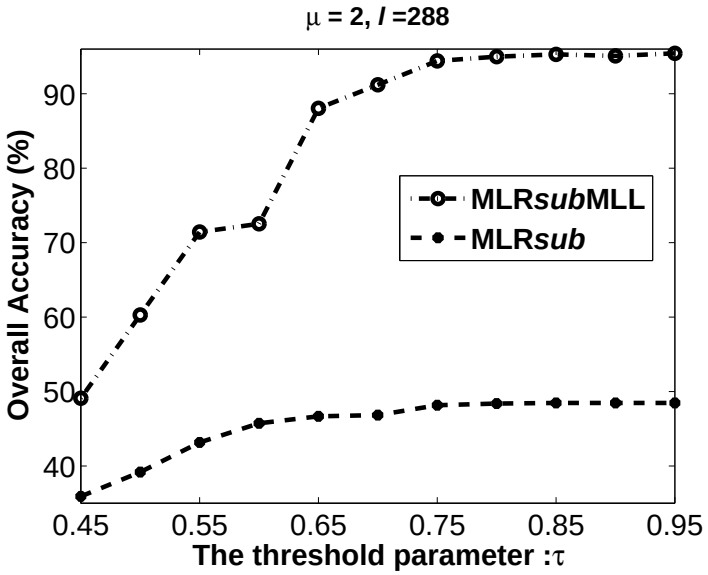


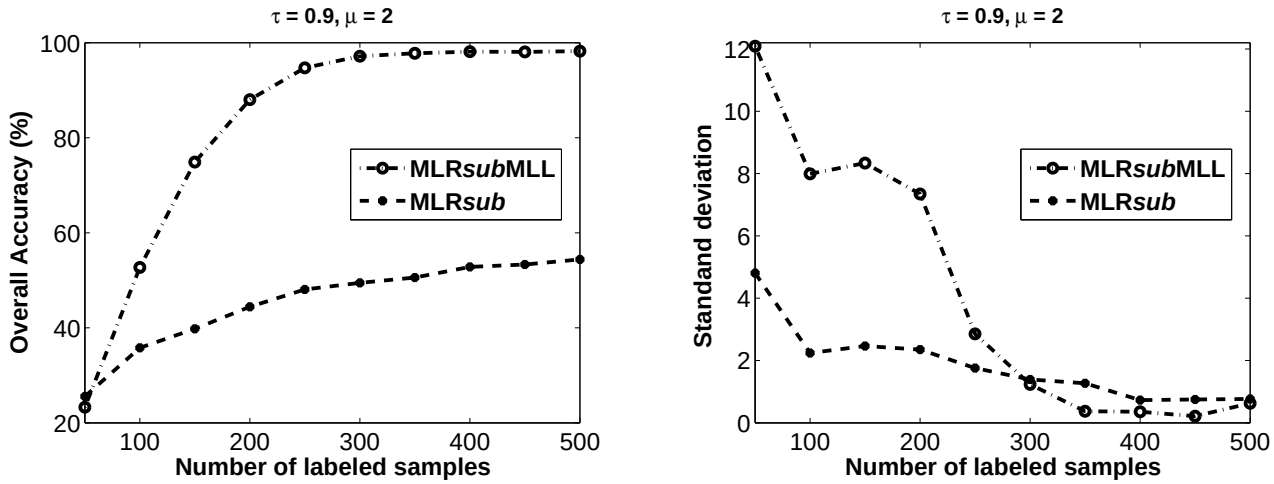
Figure 4.3: OA results (as a function of τ) with $\mu = 2$, for a problem with $\sigma = 0.8$ and $\gamma = 0.7$.

that the proposed methods can perform well in the presence of limited training sets, even after the dimensionality of the subspace is increased. The robustness of the proposed methods in the presence of very limited training sets is analyzed in more detail in the following experiment.

Experiment 3: Impact of the training set size

In our third simulated image experiment, we analyze the impact of the training set size on the segmentation performance. Figure 4.4(a) and (b) respectively report the OA and standard deviation (std) results obtained by our proposed methods as a function of the number of labeled samples (l) used in the training process, with $\tau = 0.9$ and $\mu = 2$. Again, these parameter settings may be sub-optimal but lead to very good results in our experiments. Notice the quality of the segmentation results obtained by our proposed MLR*sub*MLL algorithm, which shows high robustness even with very limited training set sizes. As the number of labeled samples increases, the OA increases and the standard deviation decreases. This is expected, since an increase of the number of labeled samples should decrease in the uncertainty when estimating the right subspace for each class.

On the other hand, we have experimentally observed that the OA and the standard deviation results respectively converge to very high and very low values for a certain number of labeled samples. In our particular case, the use of 350 labeled samples resulted in an OA of 97.76% with std = 0.37. This indicates that robust generalization can be achieved by the combination of MLR regressors and spatial-contextual information. From this experiment, we can conclude that our proposed algorithm converges to almost identical results once the classes are well-separated using



(a) (b)

Figure 4.4: Classification and segmentation results obtained for a problem with $\sigma = 0.8$ and $\gamma = 0.7$ by using $\tau = 0.9$, $\mu = 2$. (a) OA results as a function of the number of labeled samples. (b) Standard deviation (std) results as a function of the number of labeled samples.

a *sufficient* number of labeled training samples, where the term *sufficient* in our experiments means a low percentage of labeled samples. Despite the encouraging results obtained thus far with the conducted simulations, a more detailed investigation of two additional aspects: the relevance of the smoothness parameter μ on spatial characterization, and the overall performance of our proposed approaches in the presence of different noise levels, should be conducted. This will be done in the next two experiments performed with our simulated hyperspectral scene.

Experiment 4: Impact of parameter μ

In this experiment we conduct an evaluation of the impact of the smoothness parameter μ on the obtained segmentation results. In practice, we use the cross-validation sampling method [77] to estimate μ by using available training samples. Figure 4.5 plots the obtained OA results as a function of μ, with $\tau = 0.9$ and $l = 288$ (2% of the available samples per class, evenly distributed among classes). From Figure 4.5, we can conclude that the segmentation performance indeed depends on the setting of μ. However, even with a sub-optimal parameter setting $1.5 \leq \mu \leq 4$, the proposed MLR*sub*MLL algorithm leads to good segmentation results for the considered problem. This indicates that the algorithm is not very sensitive to the setting of parameter μ since all values of this parameter in a certain range of interest ultimately lead to high values of the OA for the considered problem.

It should be noted that, in all experiments conducted thus far, the noise standard deviation considered in the simulations was $\sigma = 0.8$ (a reasonable parameter setting according to our tests). However, a remaining aspect to be analyzed is the sensitivity of the proposed method to

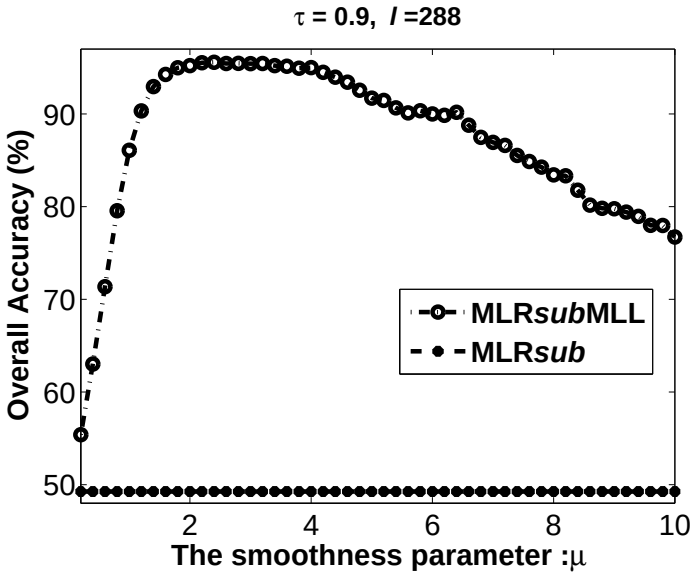


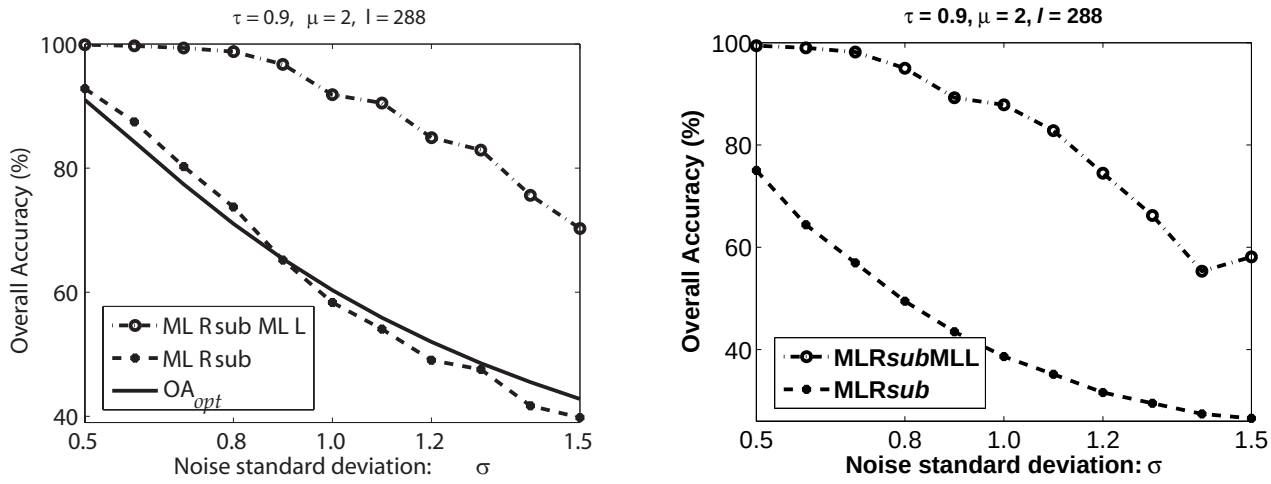
Figure 4.5: OA results as a function of the smoothness parameter μ for a problem with $\sigma = 0.8$ and $\gamma = 0.7$ with $\tau = 0.9$ and $l = 288$.

different noise levels.

Experiment 5: Impact of noise

In our last experiment with simulated data we evaluate the impact of noise on the proposed segmentation algorithm by using only $l = 288$ labeled samples (2% of the available samples per class, evenly distributed among classes) as in previous experiments. Figure 4.6 plots the OA results as a function of the noise standard deviation σ for two different problems: (a) $\gamma = 1$; and (b) $\gamma = 0.7$. As shown by Figure 4.6, the performance of the proposed MLR*sub*MLL algorithm decreases as σ increases, but the increase in the OAs obtained with regards to the MLR*sub* classification are always remarkable. From Figure 4.6, we can also conclude that the results achieved by our proposed segmentation algorithm are superior to the OA_{opt} result. Specifically, for the problem with $\sigma = 1.5$, the MLR*sub*MLL obtained a segmentation OA of 58.12% with $\gamma = 0.7$ [see Figure 4.6(b)], which is 15.34% higher than the optimal value ($\text{OA}_{opt} \leq 42.78\%$ with $\gamma = 1$) in Figure 4.6(a).

Summarizing, the experimental results conducted with simulated data sets indicate that the proposed MLR*sub*MLL algorithm achieves adequate performance in noisy environments and with limited training sets, exhibiting robustness for a wide range of parameter settings that simplify the choice of such parameters by the end-user. In other words, although the performance of the algorithm has been shown to be dependent on the setting of parameters τ and μ, sub-optimal settings of these parameters are easy to obtain and lead to good characterization results in different simulation environments. Although the experimental evaluation conducted with



(a) $\gamma = 1$ (b) $\gamma = 0.7$

Figure 4.6: OA results achieved (for two different values of γ) as a function of the noise standard deviation σ with $\tau = 0.9$, $\mu = 2$, and $l = 288$.

simulated data sets provided very encouraging results, further analyses with real hyperspectral scenes and comparisons to other state-of-the-art methods are highly desirable in order to fully substantiate the proposed method.

4.4.3 Experiments with real hyperspectral data

In order to evaluate the proposed MLR*sub*MLL algorithm in real analysis scenarios, we use two widely used hyperspectral data sets respectively collected by AVIRIS and the Reflective Optics System Spectrographic Imaging System (ROSIS) operated by the German Aerospace Agency (DLR). For the purpose of comparison, we use other state-of-the-art supervised classifiers such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic discriminant analysis (LogDA), and SVMs [54, 57, 116], which are well-established techniques in the machine learning community [6, 28, 48, 110]. For these methods, we project the original hyperspectral datasets into a subspace by using the hyperspectral signal identification by minimum error (HySime) method [11] which was observed to perform better than standard eigenvector calculation considered for the other tested methods, where the loss of spectral information after projecting the data into the subspace is also controlled by parameter τ as what we consider in our approach. Furthermore, in order to have a fair comparison with our segmentation method (which includes spatial-contextual information), in this work we have also expanded the considered discriminant analysis approaches (LDA, QDA and LogDA) with the MLL spatial prior to obtain segmentation methods (referred to hereinafter as LDAMLL, QDAMLL, and LogDAMLL) that can be compared with our proposed algorithm. In all experiments, we empirically set $\tau = 0.999$ and $\mu = 2$. Although sub-optimal, we have experimentally tested that these settings lead to good

characterization results with all the considered data sets, a fact that reveals that the proposed approach can perform accurately using a variety of hyperspectral scenes collected by different instruments.

Experiment 1: AVIRIS Indian Pines data set

In our first experiment, we use the well-known AVIRIS Indian Pines data set to analyze the performance of the proposed algorithm in comparison with other methods. The scene contains 145×145 pixels and 202 spectral bands. The ground truth data contains 16 mutually exclusive classes, and a total of 10366 labeled pixels. This image is a classical benchmark to validate the accuracy of hyperspectral image analysis algorithms and constitutes a challenging problem due to the significant presence of mixed pixels in all available classes, and also because of the unbalanced number of available labeled pixels per class.

In order to test the the performance of the proposed algorithms with limited training sets, a total size of $l = 1036$ (which represents 10% of the available labeled samples, evenly distributed among classes) is used for training purposes, where the remaining 90% of the samples were used for validation. Table 4.1 illustrates the OA, average accuracy (AA), kappa statistic coefficient (κ), and individual class accuracy [%] results achieved by the proposed algorithms after 10 Monte Carlo runs. By adopting an MLL spatial prior, the segmentation algorithms significantly improved the classification results obtained by the considered classification algorithms. For instance, the MLR*sub*MLL obtained an OA of 93.66%, 19.51% larger than that obtained by the MLL*sub* algorithm, whereas the QDAMLL obtained an OA of 90.02%, which is 10.18% higher than the result obtained by the QDA algorithm. It is remarkable that the MLR*sub*MLL algorithm did not provide the best classification results in our experiments (it only outperformed the classification results provided by LDA). However, the inclusion of the MLL prior improved more significantly the results obtained by MLR*sub* than those obtained by the discriminant analysis methods. A possible explanation for this result is due to the inclusion of the subspace projection method, which leads to reliable posterior probabilities for each class after reducing the negative effects caused by noise and mixed pixels.

For illustrative purposes, Figure 4.7 presents the the ground truth and some of the classification/segmentation results obtained by the different tested methods for the AVIRIS Indian Pines scene. For each method, we randomly selected one of the maps obtained after conducting 10 Monte Carlo runs. As shown by Figure 4.7, the SVM produced the best classification map while the MLR*sub*MLL produced the best segmentation map. An immediate issue resulting from experiments in Figure 4.7 is whether the use of spatial-contextual information could result

Table 4.1: Overall, average, individual class accuracies [%] and κ statistic obtained for the AVIRIS Indian Pines data set. The best results are highlighted in bold typeface.

Class	# samples	Classification algorithms					Segmentation algorithms			
		MLR*sub*	LDA	QDA	LogDA	SVM	MLR*sub*MLL	LDAMLL	QDAMLL	LogDAMLL
Alfalfa	54	66.94	82.34	55.91	83.50	94.48	94.62	95.38	67.94	**97.69**
Bldg-grass-tree-drives	1434	59.05	64.97	68.88	64.42	71.34	86.51	76.37	81.95	**96.12**
Corn	834	54.48	41.43	71.05	60.72	68.00	87.07	51.29	**87.27**	81.14
Corn-no till	234	73.68	68.76	88.65	77.04	85.23	91.72	78.16	92.42	**93.16**
Corn-min till	497	57.43	68.89	73.37	67.68	73.19	**86.27**	82.44	82.38	79.11
Grass/pasture	747	98.90	97.31	96.82	94.77	96.94	**99.52**	99.76	98.43	99.12
Grass/pasture-mowed	26	91.01	53.88	74.65	68.97	77.77	85.54	71.99	**87.40**	85.97
Grass/tree	489	71.89	87.00	19.56	76.67	85.89	90.00	**92.00**	23.67	90.00
Hay-windrowed	20	71.41	48.60	70.42	67.90	73.74	**97.88**	77.28	89.69	90.23
Oats	968	75.13	67.25	86.82	75.66	86.31	**93.16**	73.79	95.16	83.01
Soybeans-no till	2468	77.25	66.76	86.58	74.45	87.03	98.35	94.44	97.30	**99.12**
Soybeans-min till	614	90.47	70.35	89.86	89.22	92.71	96.42	69.44	86.49	**94.13**
Soybeans-clean till	212	93.85	95.12	93.23	90.33	94.27	**99.18**	98.75	96.44	98.36
Stone-steel towers	1294	91.64	84.99	91.64	95.37	97.42	**98.86**	85.58	98.57	**98.86**
Wheat	380	99.33	99.73	97.71	97.92	99.13	99.67	**99.87**	98.79	99.73
Woods	95	94.40	84.32	94.68	92.13	90.29	**98.43**	87.30	95.78	93.68
OA		74.15	65.22	79.84	75.42	80.56	**93.66**	79.41	90.02	89.23
AA		77.30	73.67	78.73	79.82	77.81	**93.95**	83.36	86.23	90.46
κ		70.30	60.61	77.02	72.00	85.86	**92.69**	76.41	88.56	87.66

in an increase in the SVM classification results. In order to analyze this issue in more detail, in the following experiment we will consider a recently developed SVM-based classifier which combines spatial and spectral information [53]. Further, we will also consider a segmentation method based on the watershed transform [129]. The results for these methods were only available to us in the framework of experiments previously conducted with the ROSIS University of Pavia data set [107, 129], and hence could not be included in the AVIRIS Indian Pines image experiments.

Experiment 2: ROSIS University of Pavia data set

The second real hyperspectral data set that we have considered in experiments was acquired in 2001 by the ROSIS instrument, flown over the city of Pavia, Italy. The image scene, with size of 610×340 pixels, is centered at the University of Pavia. After removing 12 bands due to noise and water absorption, it comprises 103 spectral channels. Nine ground truth classes, with a total of 3921 training samples and 42776 test samples were considered in experiments. Two different tests were performed with this scene.

In our first test we used the entire training set available for this scene in order to train different classifiers. Table 4.2 reports the obtained values of OA, AA, κ and individual accuracies. In this comparison, we included the same set of classifiers used in the experiments with the AVIRIS Indian Pines image, along with two additional spectral-spatial classifiers: an SVM-based classifier trained with extended morphological profiles (designated in the table by EMP/SVM) [53], and a segmentation method based on the watershed transform [129]. The results reported in the table are respectively taken from [107] and [129], where exactly the same training and test sets

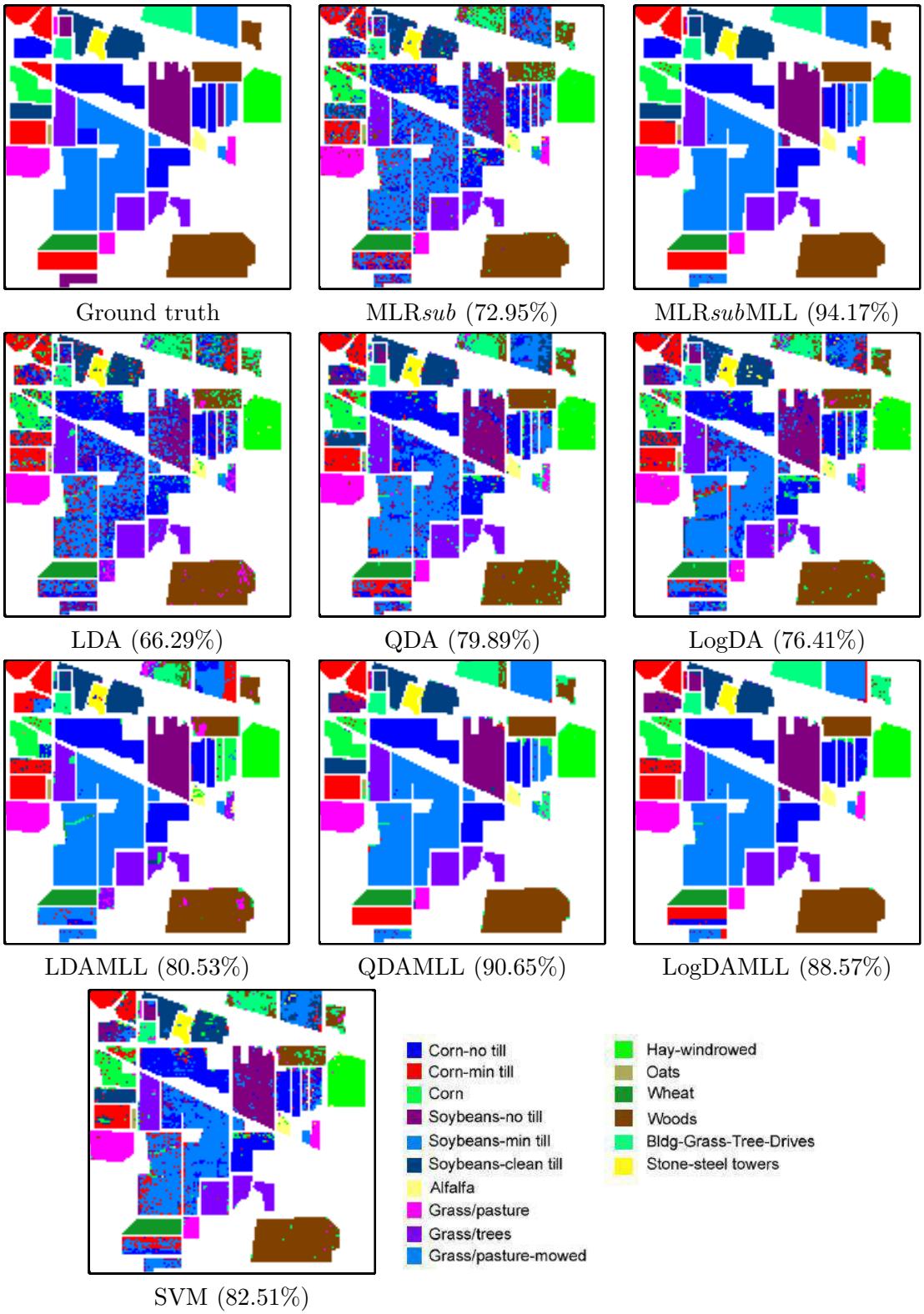


Figure 4.7: Classification/segmentation maps obtained by the different tested methods for the AVIRIS Indian Pines scene (overall accuracies are reported in the parentheses).

Table 4.2: Overall, average, individual class accuracies [%] and κ statistic obtained for the ROSIS University of Pavia data set. The best results are highlighted in bold typeface.

Class	# samples		Classification algorithms					Segmentation algorithms					
	Train	Test	MLR*sub*	LDA	QDA	LogDA	SVM†	MLR*sub*MLL	LDAMLL	QDAMLL	LogDAMLL	EMP/SVM†	[129]‡
Asphalt	548	6631	65.63	69.45	67.65	70.89	83.71	93.83	89.56	90.68	86.91	**95.36**	93.64
Bare soil	540	18649	64.80	46.59	73.49	75.06	92.25	98.43	45.93	98.27	**98.57**	63.72	97.35
Bitumen	392	2099	81.28	63.31	93.53	83.98	81.58	**99.32**	62.11	93.38	90.38	98.87	96.23
Bricks	524	3064	59.78	88.29	89.52	87.91	92.59	95.19	**99.08**	98.45	97.39	95.41	97.92
Gravel	265	1345	66.51	39.11	59.79	55.31	70.32	71.13	26.82	62.17	38.21	**87.61**	66.12
Meadows	532	5029	64.19	81.92	75.73	76.72	70.25	**94.80**	85.15	84.73	84.59	80.33	75.09
Metal sheets	375	1330	99.78	99.41	99.93	**100**	99.41	**100**	99.70	99.93	**100**	99.48	99.91
Shadows	514	3682	92.82	99.79	99.26	99.79	96.62	96.20	**100**	99.79	94.09	97.68	96.98
Trees	231	947	72.19	95.07	96.64	96.38	97.81	92.17	94.09	**99.80**	94.94	98.37	98.56
OA			67.08	75.59	77.95	78.41	80.99	**94.10**	80.27	89.48	87.04	85.22	85.42
AA			74.11	75.88	83.95	82.90	88.28	**93.45**	78.05	91.91	83.32	90.76	91.31
κ			58.53	68.16	71.93	72.47	76.16	**92.24**	73.90	86.46	87.23	80.86	81.30

Notes:
† Results are directly taken from [107], which used EMPs for spectral-spatial characterization prior to SVM-based classification.
‡ Results are directly taken from [129], which used a spectral-spatial classifier based on a pixel-wise SVM classifier with majority voting within the watershed regions to produce the final segmentation.

mentioned above were used to produce the reported results, thus allowing a fair inter-comparison of methods. By using the entire training set, the proposed MLR*sub*MLL algorithm obtained an OA of 94.10% in the considered analysis scenario. For illustrative purposes, Figure 4.8 presents the classification and segmentation maps achieved by some of the considered methods.

In our second test we analyze the sensitivity of the considered methods to different training sets made up of a limited number of samples. For this purpose, we constructed small training sets by randomly selecting 20, 30, 40, 60, 80, 100 labeled samples per class. Figure 4.9 illustrates the obtained OA results by the different methods as a function of the number of labeled samples per class. By using only 60 labeled samples per class ($l = 540$ samples, which represents around 14% of the entire training set), the proposed MLR*sub*MLL obtained an OA of 88.85%. This result is quite remarkable since, for instance, the OA obtained by the EMP/SVM algorithm by using the entire training set was slightly lower. When a spatial prior was adopted, the segmentation algorithms in Figure 4.9(b) always achieved significantly better results than their classification counterparts in Figure 4.9(a), thus indicating the importance of including spatial-contextual information. In this case, the SVM classifier in Figure 4.9(a) could not improve any of the segmentation methods in Figure 4.9(b). The figure also indicates that the segmentation performance of the proposed MLR*sub*MLL can significantly increase as the number of labeled samples increases. However, Figure 4.9(a) indicates that the classification OAs cannot increase so significantly as the size of the training set becomes larger. This is because more reliable estimates of the posterior probabilities can be obtained by the MAP segmentation algorithm as the number of labeled samples increases. As a result, the proposed segmentation method can perform well in the presence of limited training samples and also significantly increase its performance when additional training samples become available.

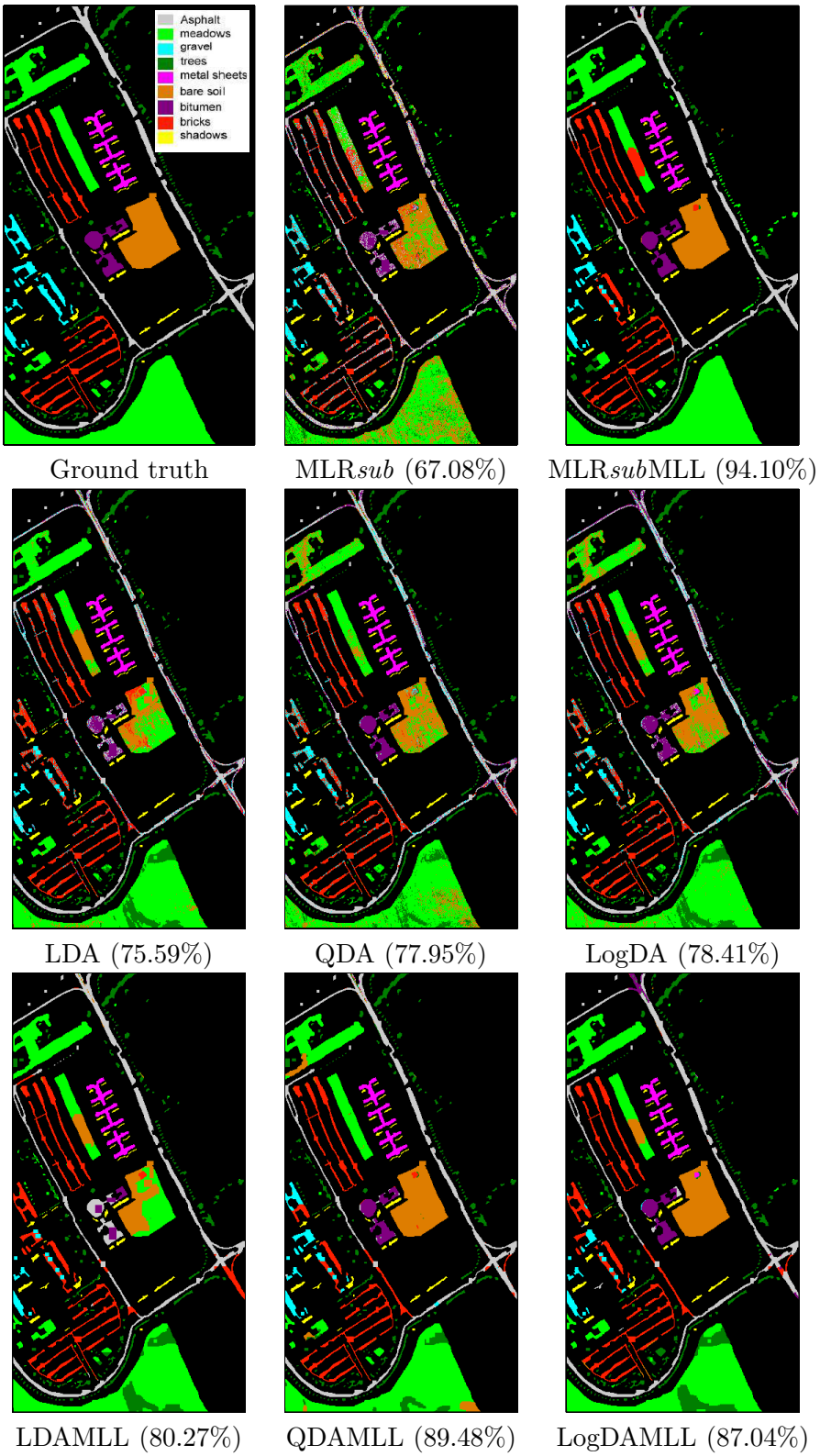
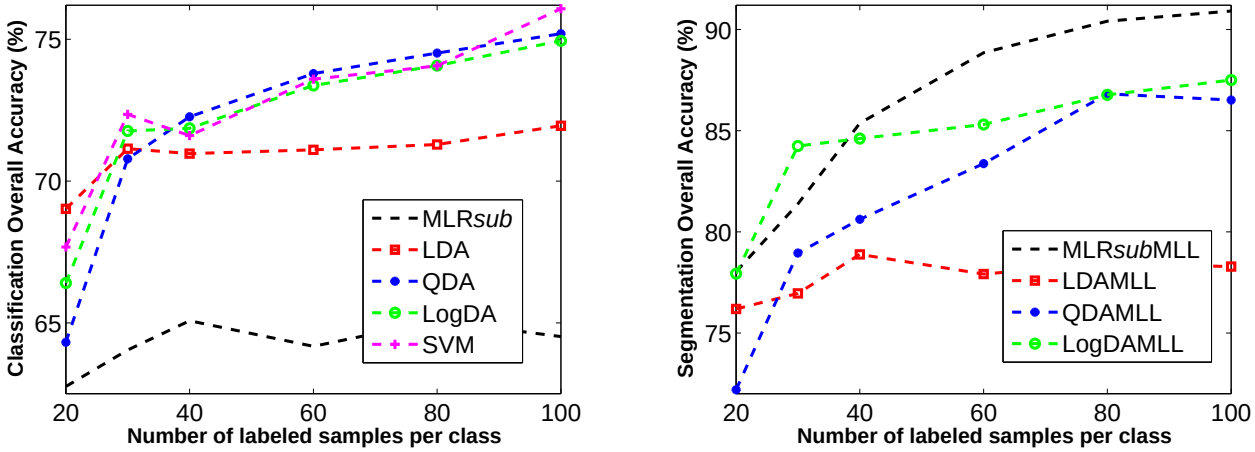


Figure 4.8: Classification/segmentation maps obtained by the different tested methods for the ROSIS University of Pavia scene (overall accuracies are reported in the parentheses)



(a) classification results

(b) segmentation results

Figure 4.9: OA results as a function of the number of labeled samples per class for the University of Pavia data set.

4.5 Conclusions

In this paper, we have developed a new spectral-spatial segmentation approach which combines multinomial logistic regression (MLR) with a subspace projection method to better characterize noise and mixed pixels. It includes contextual information using a multi-level logistic (MLL) Markov-Gibbs prior. By computing the maximum a posteriori (MAP) segmentation with an optimized α-expansion graph-cut based algorithm, the proposed segmentation method provides good accuracies when compared with other methods. It also exhibits robustness to different criteria, such as noise, presence of mixed pixels, and limited availability of training samples without the need for fine tuning of input parameters. Although our experimental results are competitive with those reported for other state-of-the-art spectral and spectral-spatial classification/segmentation methods, further work should be focused on conducting additional experiments with real hyperspectral scenes collected by other instruments, such as the new generation of spaceborne instruments that are currently under development. Given the similar spectral and spatial resolutions of these instruments with regards to the airborne systems adopted in our real experiments, we anticipate that the proposed robust segmentation techniques can also perform accurately with the new generation of satellite instruments.

Chapter 5

Conclusions and Future Work

This thesis presented new developments for the problems of remotely sensed hyperspectral image classification and segmentation, in which the ultimate goal is to accurately interpret the image data provided by remotely sensed hyperspectral imaging instruments in the context of Earth observation applications. Our proposed classification techniques exploit the rich spectral information available in this kind of data (typically, hundreds of spectral bands), while the developed segmentation techniques make combined use of both the spatial and the spectral information present in the data. Specifically, we have focused on the problem of supervised and semi-supervised hyperspectral image classification/segmentation, in which some training data is assumed to be available *a priori*, and particularly addressed some of the most relevant challenges that can be found in this context. These challenges can be summarized as follows:

- First and foremost, we have addressed the problems related with the imbalance between the high dimensionality of hyperspectral data in the spectral domain and the limited availability of training samples in real applications, which poses critical problems for supervised algorithms, most notably, in order to avoid the well-known Hughes effect. In order to address this challenge, we have adopted strategies based on semi-supervised learning and active sampling which allowed us to increase the training set without significant cost and effort.

- Second, we have used a particular class of discriminative classifiers based on the concept of multinomial logistic regression (MLR), which represent an innovation with regards to previous developments in the hyperspectral imaging literature. These discriminative classifiers are able to learn directly the posterior class distributions and deal with the high dimensionality of hyperspectral data in a very effective way. The structure of MLR classifiers is very open and flexible. Compared to other techniques, the MLR is based on computing posterior probabilities, which is a crucial step for Bayesian segmentation (based on the incorporation of spatial information).

- Third, in our developments we have taken advantage of the fact that, in addition to the very rich spectral information available in the hyperspectral data, hyperspectral images exhibit piecewise statistical continuity among neighboring pixels. In order to take advantage from this feature, our proposed techniques have been designed to exploit spatial information in conjunction with spectral information in order to partition an image into a set of homogeneous regions (in statistical sense).

- Finally, we have also developed innovative strategies to cope with one of the most important problems in hyperspectral image analysis: the presence of mixed pixels (with possibly many participating constituents at a sub-pixel level) due to limited spatial resolution, mixing phenomena happening at different scales, etc. To address this issue we resort to subspace-based techniques that can better discriminate land-cover classes in the presence of heavily mixed pixels.

After describing our general contributions, we describe next the specific contributions in the three main chapters of this thesis. In each case future research lines are identified.

- In chapter 2, we developed a new supervised Bayesian segmentation approach, namely LORSAL-AL-MLL, aimed at addressing ill-posed hyperspectral classification problems. LORSAL-AL-MLL models the posterior class probability distributions using the concept of multinomial logistic regression (MLR), where the MLR regressors are learned by the logistic regression via splitting and augmented Lagrangian (LORSAL) algorithm. The algorithm adopts a multi-level logistic (MLL) prior to model the spatial information present the class label images. The maximum *a posteriori* (MAP) segmentation is efficiently computed by the α-Expansion graph-cut based algorithm. Moreover, active learning based on maximizing the mutual information between the regressors and the class labels is considered which can effectively cope with training sets containing a very limited number of samples. The effectiveness of the proposed algorithm is illustrated using both simulated and real hyperspectral datasets. A comparison with state-of-the-art methods indicates that the proposed approach yields comparable or superior performances using fewer labeled samples. Further work should be conducted in order to test the proposed method with additional scenes and analysis scenarios.

- In chapter 3, we developed a new (supervised) Bayesian segmentation approach aimed at addressing ill-posed hyperspectral classification and segmentation problems. The proposed algorithm models the posterior class probability distributions using the concept of multinomial logistic regression (MLR), where the MLR regressors are learned by the logistic

regression via splitting and augmented Lagrangian (LORSAL) algorithm. The algorithm adopts a multi-level logistic (MLL) prior to model the spatial information present the class label images. The maximum a posteriori (MAP) segmentation is efficiently computed by the α-Expansion graph-cut based algorithm. The resulting segmentation algorithm (LORSAL-MLL) greatly improves the overall accuracies with respect to the classification results just based on the learned class distribution. Another contribution of this work is the incorporation of active learning strategies in order to cope with training sets containing a very limited number of samples. Three different sampling approaches, namely: a mutual information (MI)-based criterion, a breaking ties (BT) strategy, and a newly developed method called modified breaking ties (MBT) are integrated in the developed classification (LORSAL) and segmentation (LORSAL-MLL) methods, resulting in two new methods with active learning respectively called LORSAL-AL and LORSAL-MLL-AL. The effectiveness of the proposed algorithms is illustrated in this work using both simulated and real hyperspectral datasets. A comparison with state-of-the-art methods indicates that the proposed approaches yield comparable or superior performance using fewer labeled samples. Moreover, our experimental results reveal that the proposed MBT approach leads to an unbiased sampling as opposed to the MI and BT strategies. Further work will be directed towards testing the proposed approach in other different analysis scenarios dominated by the limited availability of training samples.

- In chapter 4, we developed a new spectral-spatial segmentation approach which combines multinomial logistic regression (MLR) with a subspace projection method to better characterize noise and mixed pixels. It includes contextual information using a multi-level logistic (MLL) Markov-Gibbs prior. By computing the maximum a posteriori (MAP) segmentation with an optimized α-expansion graph-cut based algorithm, the proposed segmentation method provides good accuracies when compared with other methods. It also exhibits robustness to different criteria, such as noise, presence of mixed pixels, and limited availability of training samples without the need for fine tuning of input parameters. Although our experimental results are competitive with those reported for other state-of-the-art spectral and spectral-spatial classification/segmentation methods, further work should be focused on conducting additional experiments with real hyperspectral scenes collected by other instruments, such as the new generation of spaceborne instruments that are currently under development. Given the similar spectral and spatial resolutions of these instruments with regards to the airborne systems adopted in our real experiments, we anticipate that the proposed robust segmentation techniques can also perform accu-

rately with the new generation of satellite instruments. Another important research line deserving future experimentation focuses on the fusion/aggregation of the results obtained by different classifiers, *i.e.*, by merging the results obtained by different methods using pixel-wise majority voting.

Finally, another future direction worth being investigated in all cases is the computational complexity of the developed methods. Although the proposed algorithms have been implemented in an efficient way by means of software optimizations, hardware optimizations related with parallel computing and efficient partitioning for exploitation of high performance computing architectures are also feasible. This is a highly relevant problem in the context of hyperspectral imaging, in which the dimensionality of the hyperspectral data is ever-increasing (instruments with thousands of spectral bands are currently under development) and the time constraints to process the data are more and more demanding in many application domains, in which near real-time performance of algorithm analysis is required in order to adequately exploit the data. With these issues in mind, a future research line that we are considering is related with the computationally efficient implementation of the proposed approaches in high performance computing architectures such as clusters of computers, or even more specialized hardware accelerators (susceptible of being used on-board the sensor platform) including digital signal processors (DSPs), field programmable gate arrays (FPGAs), or commodity graphic processing units (GPUs).

References

[1] M. V. Afonso, J. Bioucas-Dias, and M. Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE Transactions on Image Processing*, 19:2345–2356, 2010.

[2] Y. Altun, T. Hofmann, and A. J. Smola. Gaussian process classification for segmenting and annotating sequences. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[3] H. Bagan, Y. Yasuoka, T. Endo, X. Wang, and Z. Feng. Classification of airborne hyperspectral data based on the average learning subspace method. *IEEE Geoscience and Remote Sensing Letters*, 5:368–372, 2008.

[4] S. Bagon. Matlab wrapper for graph cut, December 2006.

[5] S. Baluja. Probabilistic modeling for face orientation discrimination: learning from labeled and unlabeled data. In *Proceedings of Neural Information Procesing systems*, pages 854–860, 1998.

[6] T. V. Bandos, L. Bruzzone, and G. Camps-Valls. Classification of hyperspectral images with regularized linear discriminant analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 47:862–873, 2009.

[7] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 43:480–491, 2005.

[8] J. A. Benediktsson, M. Pesaresi, and K. Amason. Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *IEEE Transactions on Geoscience and Remote Sensing*, 41(9):1940–1949, 2003.

[9] J. Bernardo and A. Smith. *B. Theory.* J. Wiley & Sons, Chichester, UK, 1994.

[10] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, 36:192–236, 1974.

[11] J. Bioucas-Dias and J. Nascimento. Hyperspectral subspace identification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(8):2435–2445, 2008.

[12] J. M. Bioucas-Dias and M. Figueiredo. Logistic regression via variable splitting and augmented Lagrangian tools. Technical report, Instituto Superior Técnico, TULisbon, 2009.

[13] C. M. Bishop. *Pattern recognition and machine learning (information science and statistics).* Springer, 1st edition, 2007.

[14] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the 18th International Conference on Machine Learning*, pages 19–26, 2001.

[15] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory*, pages 92–100, 1998.

[16] D. Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44:197–200, 1992.

[17] J. Bolton and P. Gader. Random set framework for context-based classification with hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3810–3821, 2009.

[18] J. Borges, J. Bioucas-Dias, and A. Marçal. Fast sparse multinomial regression applied to hyperspectral data. In *International Conference on Image Analysis and Recognition*, pages 700–709, 2006.

[19] J. Borges, J. Bioucas-Dias, and A. Marçal. Evaluation of Bayesian hyperspectral imaging segmentation with a discriminative class learning. In *IEEE International Geoscience and Remote sensing Symposium*, pages 3810–3813, 2007.

[20] J. S. Borges, J. M. Bioucas-Dias, and A. R. S. Marcal. Bayesian hyperspectral image segmentation with discriminative class learning. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1–14, 2011.

[21] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

[22] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy mminimization in vision. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[23] Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1222–1239, 2001.

[24] U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel. Efficient co-regularised least squares regression. In *Proceedings of the 23rd International Conference on Machine learning*, 2006.

[25] L. Bruzzone, M. Chi, and M. Marconcini. A novel transductive SVM for the semisupervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3363–3373, 2006.

[26] C. Campbell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *Proceedings of the 17th International Conference on Machine Learning*, pages 111–118, 2000.

[27] G. Camps-Valls, T. Bandos, and D. Zhou. Semi-supervised graph-based hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 45:3044–3054, 2007.

[28] G. Camps-Valls and L. Bruzzone. Kernel-based methods for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 43:1351–1362, 2005.

[29] G. Camps-Valls, L. Gómez-Chova, J. Calpe, E. Soria, J. D. Martín, L. Alonso, and J. Moreno. Robust support vector method for hyperspectral data classification and knowledge discovery. *IEEE Transactions on Geosceience and Remote Sensing*, 42:1530–1542, 2004.

[30] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Marí, Joan Vila-Francés, and Javier Calpe-Maravilla. Composite kernels for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 3:93–97, 2006.

[31] C.-I Chang. Orthogonal subspace projection (OSP) revisited: a comprehensive study and analysis. *IEEE transactions on geoscience and remote sensing*, 43(3):502–518, 2005.

[32] O. Chapelle, M. Chi, and A. Zien. A continuation method for semi-supervised SVMs. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 185–192, 2006.

[33] J. Chen, C. Wang, and R. Wang. Using stacked generalization to combine SVMs in magnitude and shape feature spaces for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2193–2205, 2009.

[34] M. Chi and L. Bruzzone. A semi-labeled-sample driven bagging technique for ill-posed classification problems. *IEEE Geoscience and Remote Sensing Letters*, 2(1):69–73, 2005.

[35] M. Chi and L. Bruzzone. An ensemble-driven k-NN approach to ill-posed classification problems. *Pattern Recognition Letters*, 27:301–307, 2006.

[36] M. Chi and L. Bruzzone. Semi-supervised classification of hyperspectral images by SVMs optimized in the primal. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6):1870–1880, 2007.

[37] M. Chi, R. Feng, and L. Bruzzone. Classification of hyperspectral remote sensing data with primal support vector machines. *Advances in Space Research*, 41(11):1793–1799, 2008.

[38] M. Chi, Q. Qian, and J. A. Benediktsson. Cluster-based ensemble classification for hyperspectral remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 6(4):762–766, 2009.

[39] M. L. Clark, D. A. Roberts, and D. B. Clark. Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote Sensing of Environment*, 96:375–398, 2005.

[40] L. Copa, D. Tuia, M. Volpi, and M. Kaneski. Unbiased query-by-bagging active learning for VHR image classification. In *SPIE Europe Remote Sensing*, 2010.

[41] I. Dagan and S. P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 150–157, 1995.

[42] F. Dell'Acqua, P. Gamba, A. Ferrari, J. A. Palmason, J. A. Benediktsson, and K. Arnason. Exploiting spectral and spatial information in hyperspectral urban data with high resolution. *IEEE Geoscience and Remote Sensing Letters*, 1(4):322–326, 2004.

[43] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[44] W. Di and M. M. Crawford. Locally consistend graph regularization based active learning for hyperspectral image classification. In *2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, pages 1–4, 2010.

[45] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41:613–627, 2002.

[46] Q. Du and C.-I Chang. A linear constrained distance-based discriminant analysis for hyperspectral image classification. *Pattern Recognition*, 34:361–373, 2001.

[47] Q. Du and C.-I Chang. Real-time constrained linear discriminant analysis to target detection and classification in hyperspectral imagery. *Pattern Recognition*, 36:1–12, 2003.

[48] Qian Du and N. H. Younan. On the performance improvement for linear discriminant analysis-based hyperspectral image classification. In *2008 IAPR Workshop on Pattern Recognition in Remote Sensing*, pages 1–4, 2009.

[49] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[50] M. M. Dundar and D. A. Landgrebe. A cost-effective semisupervised classifier approach with kernels. *IEEE Transactions on Geosceience and Remote Sensing*, 42:264–270, 2004.

[51] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

[52] A. A. Farag, R. M. Mohamed, and A. El-Baz. A unified framework for MAP estimation in remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 43(7):1617–1634, 2005.

[53] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3804–3814, 2008.

[54] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[55] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.

[56] A. Fujino, N. Ueda, and K. Saito. A hybrid generative/discriminative approach to semi-supervised classifer degin. In *Proceedings of the 20th national coference on Artificial intelligence*, volume 2, pages 764–769, 2005.

[57] J. Galambos and I. Simonelli. *Bonferroni-type inequalities with applications*. New York, NY: Springer, 1996.

[58] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

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

[60] L. Gomez-Chova, G. Camps-Valls, J. Munoz-Mari, and J. Calpe. Semisupervised image classification with Laplacian support vector machines. *IEEE Geoscience and Remote Sensing Letters*, 5(3):336–340, 2008.

[61] P. Gong, R. Pu, and B. Yu. Conifer species recognition: An exploratory analysis of *in situ* hyperspectral data. *Remote Sensing of Environment*, 62:189–200, 1997.

[62] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis, M. R. Olah, and O. Williams. Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sensing of Environment*, 65:227–248, 1998.

[63] D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriory estimation for binary images. *Jounal of Royal Statistics Society B*, 51(2):271–279, 1989.

[64] J. A. Gualtieri, S. R. Chettri ad R. F.Cromp, and L. F. Johnson. Support vector machine classifiers as applied to AVIRIS data. In *Presented at 1999 Airborne Geoscience Workshop*, 1999.

[65] J. A. Gualtieri and R. F. Cromp. Support vector machines for hyperspectral remote sensing classification. In *Proceeding of SPIE 27th Advances in Computer-Assisted Recognition Workshop*, pages 221–232, 1998.

[66] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23:73–102, 1995.

[67] C. Huang, L. S. Davis, and J. R. G. Townshend. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23:725–749, 2002.

[68] G. F. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory, IT*, 14(1):55–63, 1968.

[69] D. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58:30–37, 2004.

[70] Q. Jackson and D. A. Landgrebe. Adaptive Bayesian contextual classification based on Markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 40(11):2454–2463, 2002.

[71] H. Jeffreys. An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 186 (1007)*, pages 453–461, 1946.

[72] L. O. Jimenez, J. L. Rivera-Medina, E. Rodriguez-Diaz, E. Arzuaga-Cruz, and M. Ramirez-Velez. Integration of spatial and spectral information by means of unsupervised extraction and classification for homogenous objects applied to multispectral and hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4):844–851, 2005.

[73] J. Jin, B. Wang, and L. Zhang. A novel approach based on fisher discriminat null space for decompositon of mixed pixels in hyperspectral imagery. *IEEE Geosceience and Remote Sensing Letters*, 7:699–703, 200.

[74] Z. Jin, J.-Y. Yang, Z.-S. Hu, and Z. Lou. Face recognition based on the uncorrelated discriminant transformation. *Pattern Recognition*, 34:1405–1416, 2001.

[75] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, pages 200–209, 1999.

[76] G. Jun and J. Ghosh. An efficient active learning algorithm with knowledge transfer for hyperspectral data analysis. In *IEEE International Geoscience and Remote Sensing Symposium*, volume 1, pages I–52 – I–55, 2008.

[77] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint conference on artificial intelligence*, pages 1137–1143, 1995.

[78] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1568–1583, 2006.

[79] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

[80] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–968, 2005.

[81] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo. On semi-supervised classification. In *Proceeding of the 18th Annual Conference on Neural Information Processing Systems*, 2004.

[82] S. Kumar. *Models for Learning Spatial Interactions in Natural Images for Context-Based Classification*. PhD thesis, The Robotics Institute, School of Computer Science, Carnegie Mellon University, 2005.

[83] J. Lafferty. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *18th International Conference on Machine Learning*, pages 282–289, 2001.

[84] D. A. Landgrebe. Hyperspectral image data analysis as a high dimensional signal processing problems. *IEEE Signal Processing Magazine*, 19:17–28, 2002.

[85] D. A. Landgrebe. *Signal theory methods in multispectral remote sensing*. John Wiley, 2003.

[86] K. Lange, D. Hunter, and I. Yang. Optimizing transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9:1–59, 2000.

[87] J. Li, J. Bioucas-Dias, and A. Plaza. Semi-supervised hyperspectral classification. In *1st IEEE GRSS Workshop on Hyperspectral Image and Signal Processing*, pages 1–4, 2009.

[88] J. Li, J. Bioucas-Dias, and A. Plaza. Semi-supervised hyperspectral classification and segmentation with discriminative learning. In *SPIE Europe Remote Sensing*, volume 7477, 2009.

[89] J. Li, J. Bioucas-Dias, and A. Plaza. Semi-supervised hyperspectral image classification based on a Markov random field and sparse multinomial logistic regression. In *IEEE International Geoscience and Remote sensing Symposium*, volume 3, pages III–817 – III–820, 2009.

[90] J. Li, J. Bioucas-Dias, and A. Plaza. Hyperspectral image segmentation using a new Bayesian approach with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, accepted, 2010.

[91] J. Li, J. Bioucas-Dias, and A. Plaza. Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):4085–4098, 2010.

[92] S. Z. Li. *Markov random field modeling in computer vision*. Springer-Verlag, London, UK, 1995.

[93] S. Z. Li. *Markov random field modeling in image analysis.* Springer-Verlag New York, Inc., second edition, 2001.

[94] A. Lobo. Image segmentation and discriminant analysis for the identification of land cover units in ecology. *IEEE Transactions on Geosceience and Remote Sensing*, 35:1136–1145, 1997.

[95] T. Luo, K. Kramer, D. B. Goldgof, S. Samson, A. Remsen, T. Hopkins, and D. Cohn. Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research*, 6:589–613, 2005.

[96] D. Mackay. Information-based objective functions for active data selection. *Neural Computation*, 4:590–604, 1992.

[97] G. J. Mclachlan. *Discriminant Analysis and Statistical Pattern Recognition (Wiley Series in Probability and Statistics).* Wiley-Interscience, 2004.

[98] F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42:1178–1790, 2004.

[99] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. J. Smola, and K. Müller. Invariant feature extraction and classification in kernel spaces. In *Proceeding of Neural Information Processing Systems*, pages 526–532, 1999.

[100] T. M. Mitchell. The role of unlabeled data in supervised learning. In *Proceedings of the 6th International Colloquium on Cognitive Science*, 1999.

[101] P. Mitra, B. U. Shankar, and S. K. Pal. Segmentation of multispectral remote sensing images using active support vector machines. *Pattern Recognition Letters*, 25(9):1067 – 1074, 2004.

[102] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Proceeding of 16th Annual Conference on Neural Information Processing Systems*, 2002.

[103] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the 9th International Conference on Information and Knowledge Management*, pages 86–93, 2000.

[104] E. Oja. *Subspace methods of pattern recognition.* Research Studies Press Ltd., Lechworth, England, 1983.

[105] 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 *IEEE International Geoscience and Remote Sensing Symposium*, volume 1, page 4 pp., 2005.

[106] M. Pesaresi and J. A. Benediktsson. A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 39(2):309–320, 2001.

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

[108] A. Plaza, P. Martinez, J. Plaza, and R. Perez. Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations. *IEEE Transactions on Geoscience and Remote Sensing*, 43:466–479, 2005.

[109] J. Plaza, A. Plaza, and C. Barra. Multi-channel morphological profiles for classification of hyperspectral images using support vector machines. *Sensors*, 9(1):196–218, 2009.

[110] J. Poulsen and A. French. *Discriminant function analysis*.

[111] S. Prasad and L. M. Bruce. Information fusion in kernel induced spaces for robust subpixel hyperspectral ATR. *IEEE Geosceience and Remote Sensing Letters*, 6:572–576, 2009.

[112] S. Prasad, L. M. Bruce, and H. Kalluri. A robust multi-classifier decision fusion framework for hyperspectral, multi-temporal classification. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 3048–3051, 2008.

[113] S. Rajan, J. Ghosh, and M. M. Crawford. An active learning approach to hyperspectral data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 46:1231–1242, 2008.

[114] J. A. Richards and X. Jia. *Remote sensing digital image analysis: an introduction*. Springer, 4th edition, 2005.

[115] E. Riloff, J. Wiebe, and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th conference on Natural language learning*, pages 25–32, 2003.

[116] B. D. Ripley. *Pattern classification and neural networks*. Cambridge, 1966.

[117] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-supervised self-training of object detection models. In *7th IEEE Workshop on Applications of Computer Vision*, 2005.

[118] Y. D. Rubinstein and T. Hastie. Discriminative vs. informative learning. In *ACM KDD*, volume AAAI Press, pages 49–53, 1997.

[119] V. R. Sa. Learning classification with unlabeled data, 1993.

[120] G. Schohn and D. Cohn. Less is more: active learning with support vector machines. In *Proceedings of the 17th International Conference on Machine Learning*, pages 839–846, 2000.

[121] B. Scholkopf and A. Smola. *Learning with kernels-support vector machines, regularization, optimization and beyond*. MIT Press Series, Cambridge, MA, 2002.

[122] M. R. Schwaller. A geobotanical investigation based on linear discriminant and profile analyses of airborne thematic mapper simulator data. *Remote Sensing of Environment*, 23:23–34, 1987.

[123] S. B. Serpico and G. Moser. Extraction of spectral channels from hyperspectral images for classification purposes. *IEEE transactions on geoscience and remote sensing*, 45(2):484–495, 2007.

[124] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 287–294, 1992.

[125] G. Shaw and D. Manolakis. Signal processing for hyperspectral image exploitation. *IEEE Signal Processing Magazine*, 19:12–16, 2002.

[126] V. Sindhwani and P. Niyogi. A co-regularized approach to semi-supervised learning with multiple views. In *the International Conference on Machine Learning Workshop on Learning with Multiple Views*, 2005.

[127] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot. Spectral spatial classification of hyperspectral imagery based on partitional clustering techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 47(8):2973–2987, 2009.

[128] Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton. Multiple spectral spatial classification approach for hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):4122–4132, 2010.

[129] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson. Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition*, 43:2367–2379, 2010.

[130] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson. SVM and MRF-based method for accurate classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 7:640–736, 2010.

[131] M. E. Tipping and A. Smola. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[132] D. Tuia and G. Camps-Valls. Semi-supervised hyperspectral image classification with cluster kernels. *IEEE Geoscience and Remote Sensing Letters*, 6:224–228, 2009.

[133] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232, 2009.

[134] V. Vapnik. *Statistical learning theory.* John Wiley, New York, 1998.

[135] S. Velasco-Forero and V. Manian. Improving hyperspectral image classification using spatial preprocessing. *IEEE Geoscience and Remote Sensing Letters*, 6:297–301, 2009.

[136] S. Watanabe, P. F. Lambert, C. A. Kulikowski, J. L. Buxton, and R. Walker. Evaluation and selection of variables in pattern recognition. In *Computer and Information Sciences II*, 1967.

[137] J.-M. Yang, B.-C. Kuo, P.-T. Yu, and C.-H. Chuang. A dynamic subspace method for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48:2840–2853, 2010.

[138] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, pages 189–196, 1995.

[139] J. Ye, R. Janardan, V. Cherkassky, T. Xiong, J. Bi, and C. Kambhamettu. Efficient model selection for regularized linear discriminant analysis. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 532–539, 2006.

[140] J. Ye and B. Yu. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6:483–502, 2005.

[141] J. Yedidia, W. Freeman, and Y.Weiss. Understanding belief propagation and and its generalizations. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 239–269, 2003.

[142] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51:2282–2312, 2004.

[143] Y. Zhong, L. Zhang, B. Huang, and L. Pingxiang. An unsupervised artificial immune classifier for multi/hyperspectral remote sensing imagery. *IEEE Transations on Geoscience and Remote Sensing*, 44(2):420–431, 2006.

[144] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[145] X. Zhu. *Semi-Supervised Learning with Graphs*. PhD thesis, Carnegie Mellon University, 2005.