Semisupervised Hyperspectral Image Segmentation Using Multinomial Logistic Regression With Active Learning

Jun Li, José M. Bioucas-Dias, *Member, IEEE*, and Antonio Plaza, *Senior Member, IEEE*

***Abstract*—This paper presents a new semisupervised segmentation algorithm, suited to high-dimensional data, of which remotely sensed hyperspectral image data sets are an example. The algorithm implements two main steps: 1) semisupervised learning of the posterior class distributions followed by 2) segmentation, which infers an image of class labels from a posterior distribution built on the learned class distributions and on a Markov random field. The posterior class distributions are modeled using multinomial logistic regression, where the regressors are learned 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 multilevel logistic model, which enforces segmentation results in which neighboring labels belong to the same class. The maximum *a posteriori* 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/Infrared Imaging Spectrometer system of the National Aeronautics and Space Administration Jet Propulsion Laboratory over the regions of Indian Pines, IN, and Salinas Valley, CA, reveal that the proposed approach can provide classification accuracies that 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 learned class densities are considered, confirming the importance of jointly considering spatial and spectral information in hyperspectral image segmentation.**

***Index Terms*—Active learning, hyperspectral image classification, Markov random field (MRF), multilevel logistic (MLL) model, multinomial logistic regression (MLR), semisupervised learning.**

I. Introduction

In RECENT years, several important research efforts have been devoted to remotely sensed hyperspectral image segmentation and classification [1]. Hyperspectral image classification and segmentation are related problems. In order to define the problems in mathematical terms, let $\mathcal{S} \equiv \{1, \ldots, n\}$ denote a set of integers indexing the n pixels of a hyperspectral image. Similarly, let $\mathcal{L} \equiv \{1, \ldots, K\}$ be a set of K class labels, and let $\mathbf{x} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ denote an image in which the pixels are d-dimensional feature vectors. Finally, let $\mathbf{y} \equiv (y_1, \ldots, y_n) \in \mathcal{L}^n$ 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.[1] Nevertheless, in this paper, 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 data sets 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 [2].

In the past, both discriminative and generative models were 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 [3]–[5]. 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 [5].

Support vector machines (SVMs) [6] and multinomial logistic regression (MLR) [7] rely, respectively, on discriminant functions and posterior class distributions, based on which many state-of-the-art classification 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 [2], [8]–[10]. In turn, MLR-based techniques have the advantage

Manuscript received November 29, 2009; revised March 30, 2010. Date of publication August 30, 2010; date of current version October 27, 2010. This work was supported in part by the Marie Curie training Grant MEST-CT-2005-021175 from the European Commission, by the MRTN-CT-2006-035927, and by the AYA2008-05965-C04-02 Projects.

J. Li and J. M. Bioucas-Dias are with Instituto de Telecomunicações and Instituto Superior Técnico (IST), Technical University of Lisbon, 1049-001 Lisbon, Portugal (e-mail: jun@lx.it.pt; bioucas@lx.it.pt).

A. Plaza is with the Department of Technology of Computers and Communications, University of Extremadura, 10071 Caceres, Spain (e-mail: aplaza@unex.es).

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

Digital Object Identifier 10.1109/TGRS.2010.2060550

[1] 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 $R_i \cap R_j = \emptyset, i \neq j$.

of being able to model the posterior class distributions in a Bayesian framework, thus supplying (in addition to the boundaries between the classes) a degree of plausibility for such classes. Effective sparse MLR methods are available [11]. These ideas have been recently applied to hyperspectral image classification and segmentation, obtaining promising results [12].

In order to improve the accuracies obtained by SVMs or MLR-based techniques, some efforts have been directed toward the integration of spatial (contextual) information with spectral information in hyperspectral data interpretation [2], [9], [12]. 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 semisupervised learning in which classification techniques are trained with both labeled and unlabeled training samples [13], [14]. This trend has been successfully adopted in remote sensing studies [2], [15]–[18]. Most semisupervised 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 toward regions of low data density [14], 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 the variables defined on the vertices.

In this paper, we introduce a new semisupervised learning algorithm that exploits both the spatial contextual information and the spectral information in the interpretation of remotely sensed hyperspectral data. The algorithm implements two main steps: 1) semisupervised learning of the posterior class distributions, implemented by an efficient version of semisupervised learning algorithm in [13], followed by 2) segmentation, which infers an image of class labels from a posterior distribution built on the learned class distributions and on a multilevel logistic (MLL) prior on the image of labels. The posterior class distributions are modeled using MLR, where the regressors are learned using both labeled and (through a graph-based technique) unlabeled training samples. For step 1), 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 [13]. The spatial contextual information is modeled by means of an MLL prior. The final output of the algorithm is based on a maximum *a posteriori* (MAP) segmentation process which is computed via a very efficient min-cut-based integer optimization technique. The remainder of this paper is organized as follows. Section II formulates the problem and describes the proposed approach. Section III 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 IV gives details about the MLL prior. Section V 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 VI reports performance results for the proposed algorithm on synthetic and real hyperspectral data sets 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 Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over the regions of Indian Pines, IN, and Salinas Valley, CA. 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 VII concludes with some remarks and hints at plausible future research avenues.

II. Problem Formulation and Proposed Approach

With the notation introduced in Section I 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 as labelings, and partitions of $\mathcal{S}$, which we term as segmentations, is one-to-one. We will thus refer interchangeably to labelings and segmentations.

The goal of both image classification and image 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[2] $p(\mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$, where $p(\mathbf{x}|\mathbf{y})$ is the likelihood function (i.e., the probability of the feature image $\mathbf{x}$ given the labeling $\mathbf{y}$) and $p(\mathbf{y})$ is the prior on the labeling $\mathbf{y}$. Assuming conditional independence 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

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

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

$$\widehat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{L}^n} \left\{ \sum_{i=1}^{n} \left(\log p(y_i|\mathbf{x}_i) - \log p(y_i) \right) + \log p(\mathbf{y}) \right\}. \tag{2}$$

In this 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 [19], [20]. 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 [13], following a

[2] 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.

semisupervised 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 (MRF), which encourages neighboring pixels to have the same label. The MAP labeling/segmentation $\widehat{\mathbf{y}}$ is computed via the α-expansion algorithm [21], a min-cut-based tool to efficiently solve integer optimization problems. All these issues are detailed in the next section.

III. Estimation of the Logistic Regressors

The MLR model is formally given by [7]

$$p(y_i = k|\mathbf{x}_i, \boldsymbol{\omega}) = \frac{\exp\left(\boldsymbol{\omega}^{(k)}\mathbf{h}(\mathbf{x}_i)\right)}{\sum_{k=1}^{K}\exp\left(\boldsymbol{\omega}^{(k)}\mathbf{h}(\mathbf{x}_i)\right)} \tag{3}$$

where $\mathbf{h}(\mathbf{x}) \equiv [h_1(\mathbf{x}), \ldots, h_l(\mathbf{x})]^T$ is a vector of l fixed functions of the input, often termed as features; $\boldsymbol{\omega}^{(k)}$ is the set of logistic regressors for class k, and $\boldsymbol{\omega} \equiv [\boldsymbol{\omega}^{(1)^T}, \ldots, \boldsymbol{\omega}^{(K-1)^T}]^T$. Given the fact that the density (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., $\mathbf{h}(\mathbf{x}_i) = [1, x_{i,1}, \ldots, x_{i,d}]^T$, where $x_{i,j}$ is the jth component of $\mathbf{x}_i$, or nonlinear. Kernels [6], i.e., $\mathbf{h}(\mathbf{x}_i) = [1, K_{\mathbf{x},\mathbf{x}_1}, \ldots, K_{\mathbf{x},\mathbf{x}_l}]^T$, 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 [8]. From now on, d denotes the dimension of $\mathbf{h}(\mathbf{x})$.

In this problem, learning the class densities amounts to estimating the logistic regressors $\boldsymbol{\omega}$. Since we are assuming a semisupervised scenario, this estimation is based on a small set of labeled samples $\mathcal{D}_L \equiv \{(y_1, \mathbf{x}_1), \ldots, (y_L, \mathbf{x}_L)\}$ and a larger set of unlabeled samples $\mathcal{X}_U \equiv \{\mathbf{x}_{L+1}, \ldots, \mathbf{x}_{L+U}\}$. Given that our approach is Bayesian, we need to build the posterior density

$$p(\boldsymbol{\omega}|\mathcal{Y}_L, \mathcal{X}_L, \mathcal{X}_U) \propto p(\mathcal{Y}_L|\mathcal{X}_L, \mathcal{X}_U, \boldsymbol{\omega})p(\boldsymbol{\omega}|\mathcal{X}_L, \mathcal{X}_U) \tag{4}$$

$$= p(\mathcal{Y}_L|\mathcal{X}_L, \boldsymbol{\omega})p(\boldsymbol{\omega}|\mathcal{X}_{L+U}) \tag{5}$$

where $\mathcal{Y}_L \equiv \{y_1, \ldots, y_L\}$ denotes the set of labels in $\mathcal{D}_L$, $\mathcal{X}_L \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_L\}$ 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 (5).

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

$$\widehat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \{l(\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}|\mathcal{X}_{L+U})\} \tag{6}$$

where

$$l(\boldsymbol{\omega}) \equiv \log p(\mathcal{Y}_L|\mathcal{X}_L, \boldsymbol{\omega}) \equiv \log \prod_{i=1}^{L} p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$$

$$\equiv \sum_{i=1}^{L}\left(\mathbf{x}_i^T\boldsymbol{\omega}^{(y_i)} - \log\sum_{j=1}^{K}\exp\left(\mathbf{x}_i^T\boldsymbol{\omega}^{(j)}\right)\right) \tag{7}$$

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 [13], we adopt the Gaussian prior

$$p(\boldsymbol{\omega}|\boldsymbol{\Gamma}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\Gamma}\boldsymbol{\omega}\right\} \tag{8}$$

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 is written 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

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

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

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

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

With the previous definitions, we have

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

The quadratic term $\tau\|\boldsymbol{\omega}^k\|^2$ acts as a 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 rule of matrix $\mathbf{A}$, the value of τ should be much smaller than the largest eigenvalue of $\mathbf{A}$. In order to interpret the rule of the quadratic terms $\boldsymbol{\omega}^{(k)^T}\mathbf{A}\boldsymbol{\omega}^{(k)}$, let $\mathcal{V} \equiv \{1, \ldots, U + L\}$ 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

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

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 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 paper

$$\beta_{ij} = e^{-\|\mathbf{h}(\mathbf{x}_i)-\mathbf{h}(\mathbf{x}_j)\|^2}. \tag{9}$$

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 variance of Gaussian densities [22]. More precisely, we assume they are independent and that

$$\lambda_i \sim \mathrm{Gam}(\alpha, \beta), \qquad i = 1, \ldots, K-1 \tag{10}$$

where $\mathrm{Gam}(\alpha, \beta)$ 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 noninformative for scale parameters, is obtained by setting to zero the shape and the inverse scale parameters of a Gamma density.

A. Computing the MAP Estimate of the Regressors

To compute the MAP estimate of $\boldsymbol{\omega}$, we use an expectation-maximization (EM) algorithm [23] 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, the 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

$$Q(\boldsymbol{\omega}|\boldsymbol{\omega}_t) \equiv E\left[\log p(\boldsymbol{\omega}, \boldsymbol{\lambda}|\mathcal{D})|\boldsymbol{\omega}_t\right] \tag{11}$$

M-step

$$\boldsymbol{\omega}_{t+1} \in \arg\max_{\boldsymbol{\omega}} Q(\boldsymbol{\omega}|\boldsymbol{\omega}_t). \tag{12}$$

In (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 nondecreasing and, under mild assumptions, converges to local optima of the density $p(\boldsymbol{\omega}|\mathcal{D})$.

B. E-Step

From (5) and (8), we have

$$p(\boldsymbol{\omega}, \boldsymbol{\lambda}|\mathcal{D}) = p(\mathcal{Y}_L|\mathcal{X}_L, \boldsymbol{\omega})p\left(\boldsymbol{\omega}|\boldsymbol{\Gamma}(\boldsymbol{\lambda})\right)p(\boldsymbol{\lambda})c^{te} \tag{13}$$

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

$$\begin{aligned} Q(\boldsymbol{\omega}|\boldsymbol{\omega}_t) &= E\left[\log p(\mathcal{Y}_L|\mathcal{X}_L, \boldsymbol{\omega}) - (1/2)\boldsymbol{\omega}^T\boldsymbol{\Gamma}(\boldsymbol{\lambda})\boldsymbol{\omega} + C|\boldsymbol{\omega}_t\right] \\ &= \log p(\mathcal{Y}_L|\mathcal{X}_L, \boldsymbol{\omega}) - (1/2)\boldsymbol{\omega}^T E\left[\boldsymbol{\Gamma}(\boldsymbol{\lambda})|\boldsymbol{\omega}_t\right]\boldsymbol{\omega} + C' \\ &= l(\boldsymbol{\omega}) - (1/2)\boldsymbol{\omega}^T\boldsymbol{\Upsilon}(\boldsymbol{\omega}_t)\boldsymbol{\omega} + C' \end{aligned} \tag{14}$$

where $l(\boldsymbol{\omega})$ is the log-likelihood function given by (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 hyperpriors, the expectations $E[\lambda_i|\boldsymbol{\omega}]$ have well-known closed forms [22]. For the present setting, we have

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

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

C. M-Step

Given the matrix $\boldsymbol{\Upsilon}(\widehat{\boldsymbol{\omega}})$, the M-step amounts to maximizing the objective function (14), which is a logistic regression problem with a quadratic regularizer. Hereinafter, we adopt the generalized EM (GEM) [23] 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 to still generate a nondecreasing 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 [24], 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 $K-1$ and

$$\mathbf{p}_i \equiv \left[p(y=1|\mathbf{x}_i, \boldsymbol{\omega}), p(y=2|\mathbf{x}_i, \boldsymbol{\omega}), \ldots, p(y=K-1|\mathbf{x}_i, \boldsymbol{\omega})\right]^T. \tag{15}$$

Let us define the nonpositive definite matrix as

$$\mathbf{B} \equiv -\frac{1}{2}\left[\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{K-1}\right] \otimes \sum_{i=1}^{L}\mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)^T \tag{16}$$

where $\mathbf{1}$ denotes a column vector of 1s and $\mathbf{1}^T$ is the transpose of such column vector. The quadratic cost function is defined as

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

Let $\mathbf{H}(\boldsymbol{\omega})$ be the Hessian of $l(\boldsymbol{\omega})$. Matrix $\mathbf{H} - \mathbf{B}$ is semipositive definite [7], 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 (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 that $p(\boldsymbol{\omega}_t|\mathcal{D}) \leq p(\boldsymbol{\omega}_{t+1}|\mathcal{D})$, for $t = 1, 2, \ldots$.

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

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

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 [13]. Here, the complexity of a complete scanning of all elements of $\boldsymbol{\omega}$ is $O(Kd(L+d))$, which is 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 pseudocode for the GEM algorithm to compute the MAP estimate of $\boldsymbol{\omega}$ is shown in Algorithm 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 $\boldsymbol{\omega}^{(k)}$.

Algorithm 1 GEM algorithm to estimate the MLR regressors $\boldsymbol{\omega}$

```
Input: $\boldsymbol{\omega}_0$, $\mathcal{D}_L$, $\mathcal{X}_U$, $\alpha$, $\beta$, $\tau$, GEMiters, BSGiters
Define: $u_{k,l} \equiv [\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathbf{T}}/(K-1)]_{k,l}$
  $\mathbf{R} \equiv \sum_{i=1}^{L} \mathbf{h}(\mathbf{x_i})\mathbf{h}(\mathbf{x_i})^T$, $\mathbf{X} \equiv [\mathbf{h}(\mathbf{x_1}), \ldots, \mathbf{h}(\mathbf{x}_{L+U})]$
  $\mathcal{B} := \mathcal{B}(\mathbf{X})$ (* build the graph weights according to (9) *)
  $\Delta := \Delta(\mathcal{B})$ (*$\Delta$ 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
        $\boldsymbol{\omega}_{(i)}^{(k)} = \text{solution}\{\mathbf{C}_{k,k}\boldsymbol{\omega}^{(k)} = \mathbf{z}^{(k)} -$
        $\sum_{l=1,l\neq k}^{K-1} \mathbf{C}_{k,l}\boldsymbol{\omega}_i^{(l)}\}$
        end for
      end for
  end while
```

IV. Multilevel 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 paper, we integrate the contextual information with spectral information by using an isotropic MLL prior to modeling 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 to the same class. The MLL prior is a generalization of the Ising model [25] and has been widely used in image segmentation problems [26].

According to the Hammersley–Clifford theorem [27], the density associated with an MRF is a Gibbs's distribution [25]. Therefore, the prior model for segmentation has the following structure:

$$p(\mathbf{y}) = \frac{1}{Z} e^{\left(-\sum\limits_{c \in \mathcal{C}} V_c(\mathbf{y})\right)} \tag{17}$$

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[3] $\mathcal{C}$ over the image, and

$$-V_c(\mathbf{y}) = \begin{cases} \upsilon_{y_i}, & \text{if } |c| = 1 \text{ (single clique)} \\ \mu_c, & \text{if } |c| > 1 \text{ and } \forall_{i,j \in c} y_i = y_j \\ -\mu_c, & \text{if } |c| > 1 \text{ and } \exists_{i,j \in c} y_i \neq y_j \end{cases} \tag{18}$$

where μ_c is a nonnegative constant.

The potential function in (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. For example, the model generates texture-like regions if μ_c depends on c and bloblike regions, otherwise [28]. By taking $e^{\upsilon_{y_i}} \propto p(y_i)$ and $\mu_c = (1/2)\mu > 0$, (17) can be rewritten as

$$p(\mathbf{y}) = \frac{1}{Z} e^{\sum\limits_{i \in \mathcal{S}} \upsilon_{y_i} + \mu \sum\limits_{(i,j) \in \mathcal{C}} \delta(y_i - y_j)} \tag{19}$$

where $\delta(y)$ is the unit impulse function.[4] This choice gives no preference to any direction concerning υ_{y_i}. A straightforward computation of $p(y_i)$, i.e., the marginal of $p(\mathbf{y})$ with respect to i, leads to $p(y_i) \propto e^{\upsilon_{y_i}}$. Thus, in order to retain the compatibility between the prior and the marginal, we take $\upsilon_{y_i} = \log p(y_i) + c^{te}$, where c^{te} is a constant term. 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 the 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 one; 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 $i, j \in \mathcal{S}$.

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

[4] That is, $\delta(0) = 1$, and $\delta(y) = 0$, for $y \neq 0$.

V. Computing the MAP Estimate via Graph Cuts

Based on the posterior class densities $p(y_i|\mathbf{x}_i)$ and on the MLL prior $p(\mathbf{y})$ and according to (2), the MAP segmentation is finally given by

$$\begin{aligned}\widehat{\mathbf{y}} &= \arg\min_{\mathbf{y}\in\mathcal{L}^n} \sum_{i\in\mathcal{S}} -\left(\log p(y_i|\widehat{\boldsymbol{\omega}}) - \log p(y_i)\right) \\ &\quad - \left(\sum_{i\in\mathcal{S}} \log p(y_i) + \mu \sum_{i,j\in C} \delta(y_i - y_j)\right) \\ &= \arg\min_{\mathbf{y}\in\mathcal{L}^n} \sum_{i\in\mathcal{S}} -\log p(y_i|\widehat{\boldsymbol{\omega}}) - \mu \sum_{i,j\in C} \delta(y_i - y_j) \qquad (20)\end{aligned}$$

where $p(y_i|\widehat{\boldsymbol{\omega}}) \equiv p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$, computed at $\widehat{\boldsymbol{\omega}}$. Minimization of (20) is a combinatorial optimization problem involving unary and pairwise interaction terms. The exact solution for $K = 2$ was introduced by mapping the problem into the computation of a min-cut on a suitable graph [29]. This line of attack was reintroduced in the beginning of this century and has been intensely researched since then (see, e.g., [21] and [30]–[32]). As a result of this research, the number of 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 approaches to integer optimization is the so-called submodularity 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 [21] can be applied. It yields very good approximations to the MAP segmentation problem and is efficient from a computational point of view, its practical computational complexity being $O(n)$.

A. Semisupervised Algorithm

Let $\mathcal{X}_{\overline{L+U}} \equiv \{\mathbf{x}_{U+1}, \ldots, \mathbf{x}_n\}$ denote the unlabeled set in $\mathbf{x}$. The pseudocode for the proposed semisupervised segmentation algorithm with discriminative class learning MLL prior is shown in Algorithm 2.

Algorithm 2 Semisupervised segmentation algorithm

```
Input: $\mathcal{D}_L, \mathcal{X}_U, \mathcal{X}_{L+U}, \mathcal{X}_{\overline{L+U}}$, GEMiters, BSGiters, $\alpha, \beta, \tau$, m
1:  while stopping criterion is not satisfied do
2:      $\widehat{\boldsymbol{\omega}} := \mathrm{GEM}(\mathcal{D}_L, \mathcal{X}_U, \alpha, \beta, \tau, \mathrm{GEMiters}, \mathrm{BSGiters})$
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 (15) for all feature
        vectors in $\mathcal{X}_{\overline{L+U}}$*)
5:      $\mathcal{X}_{\mathrm{new}} := \varphi(\widehat{\mathbf{P}}, m)$
6:      (*$\varphi(\widehat{\mathbf{P}}, m)$ selects $m$ unlabeled samples from $\mathcal{X}_{\overline{L+U}}$. See
        explanation *)
7:      $\mathcal{X}_{L+U} := \mathcal{X}_{L+U} + \mathcal{X}_{\mathrm{new}}$
8:      $\mathcal{X}_{\overline{L+U}} := \mathcal{X}_{\overline{L+U}} - \mathcal{X}_{\mathrm{new}}$
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$)$
```

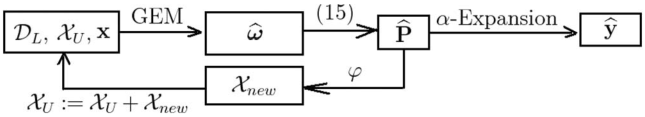

$\mathcal{D}_L, \mathcal{X}_U, \mathbf{x}$
GEM
$\widehat{\boldsymbol{\omega}}$
(15)
$\widehat{\mathbf{P}}$
α-Expansion
$\widehat{\mathbf{y}}$
φ
$\mathcal{X}_{new}$
$\mathcal{X}_U := \mathcal{X}_U + \mathcal{X}_{new}$

Fig. 1. Block scheme of Algorithm 2.

Lines 2, 10, and 11 of Algorithm 2 embody the core of our proposed algorithm. Specifically, line 2 implements the semisupervised learning of the MLR regressors through the GEM procedure described in Algorithm 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, Fig. 1 shows the most relevant components of the proposed segmentation algorithm in a flowchart.

B. Active Selection of Unlabeled Samples

Lines 3–8 in Algorithm 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 will hopefully lead to the best performance gains for the classifier. Contrary to active selection of labeled samples [33]–[35], the selection of 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 paper, we have considered two strategies for the selection criterion implemented by function φ shown in line 5 of Algorithm 2.

1) Random: In step 5, these m unlabeled samples are randomly selected from $\mathcal{X}_{\overline{L+U}}$.
2) 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 [13], [34]. However, this paper 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.

C. Overall Complexity

The complexity of Algorithm 2 is dominated by the semisupervised learning stage of the MLR regressors implemented through the GEM process in Algorithm 1, which has computational complexity $O(d^3(K-1))$ as described in Section III-A, and also by the α-expansion algorithm used to determine the MAP segmentation, which has practical complexity $O(n)$ as described in Section V. Since in most applications, $d^3(K-1) > n$, the overall complexity is dominated by that of the GEM process in Algorithm 1, which is used to learn the MLR regressors.

As already discussed, compared with the semisupervised algorithm presented in [13], the proposed semisupervised algorithm is $(K-1)^2$ faster. For a problem with 500 labeled pixels, 224 bands, and ten classes on a 2.31-GHz personal computer, with only the first 20 iterations, the proposed algorithm took 10.53 s, whereas the algorithm in [13] took 106.77 s.

VI. 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 VI-A reports the experiments with simulated data and contains the following experiments. In Section VI-A1, we conduct an evaluation of the impact of the spatial prior on the analysis of simulated data sets. Section VI-A2 performs an evaluation of the impact of incorporating unlabeled samples to the analysis. Finally, Section VI-A3 conducts an experimental evaluation of the increase in classification results after including the active-selection methodology. On the other hand, Section VI-B evaluates the performance of the proposed algorithm using two real hyperspectral scenes collected by AVIRIS over agricultural fields located at Indian Pines, IN [1], and the Valley of Salinas, CA [1]. 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 the data.[5] 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 ten Monte Carlo runs with respect to the labeled samples $\mathcal{D}_L$, except for the results over the AVIRIS Salinas data set, which are obtained after five Monte Carlo runs. The labeled samples for each Monte Carlo simulation are obtained by resampling a much larger set of labeled samples. Finally, 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).

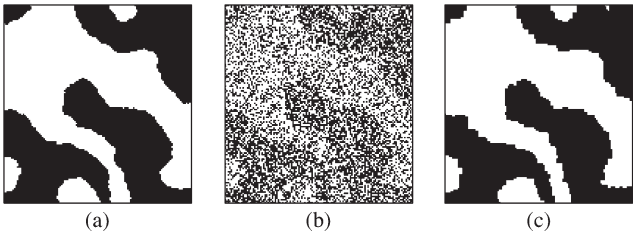


Fig. 2. Classification and segmentation results obtained after applying the proposed method on a simulated hyperspectral scene representing a binary classification problem. (a) Ground-truth class labels. (b) Classification result (OA $= 66.94\%$, with OA$_{\text{opt}} = 75.95\%$). (c) Segmentation result (OA $= 96.41\%$).

A. Experiments With Simulated Data

In this section, a simulated hyperspectral scene is used to evaluate the proposed semisupervised algorithm mainly to analyze 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

$$\mathbf{x}_{y_i} = \mathbf{m}_{y_i} + \mathbf{n}_{y_i}, \qquad i \in \mathcal{S}, \quad y_i \in \mathcal{L}^n \tag{21}$$

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., $\mathbf{n}_{y_i} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

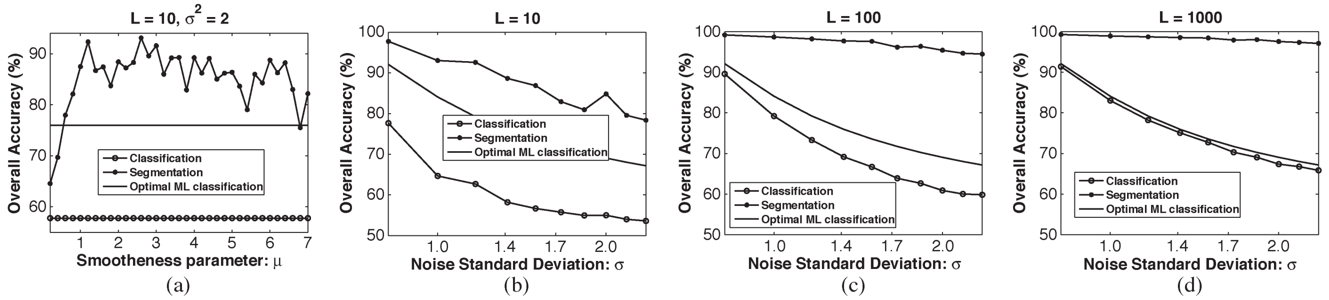
In Section VI-A1, we address a binary classification problem, i.e., $K = 2$, with $\mathbf{x}_{y_i} \in \mathbb{R}^{50}$, $\mathbf{m}_{y_i} = \xi_i\phi$, $\|\boldsymbol{\phi}\| = 1$, and $\xi_i = \pm 1$. The image of class labels $\mathbf{y}$ is shown in Fig. 2(a), where labels $y_i = 1, 2$ correspond to $\xi_i = -1, +1$, respectively. In this problem, the theoretical OA, given by $\text{OA}_{\text{opt}} \equiv 100(1 - P_e)\%$ and corresponding to the minimal probability of error [36] is

$$P_e = \frac{1}{2}\text{erfc}\left(\frac{1+\lambda_0}{\sqrt{2}\sigma}\right)p_0 + \frac{1}{2}\text{erfc}\left(\frac{1-\lambda_0}{\sqrt{2}\sigma}\right)p_1 \tag{22}$$

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

In Section VI-A2, 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

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



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

Fig. 3. (a) OA results as a function of the spatial-prior parameter μ with $L = 10$ and $\sigma^2 = 2$. (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.

Survey (USGS) digital spectral library.[6] For a multiclass classification problem, because the probability of error is difficult to compute, we use the error bound

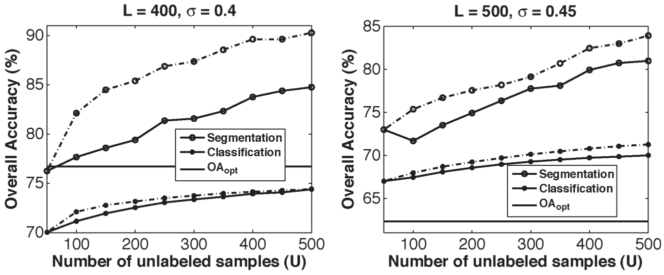
$$P_e \leq \frac{K-1}{2}\mathrm{erfc}\left(\frac{\mathrm{dist}_{\min}}{2\sigma}\right) \tag{23}$$

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 [5], which is widely used in multiclass classification problems [37], [38].

Finally, in Section VI-A3 we use the same experimental setting as in Section VI-A1 except for the number of spectral band, which is set to 200, i.e., $\mathbf{x}_{y_i} \in \mathbb{R}^{200}$.

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 independent of other considerations. Fig. 3(a) shows 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, Fig. 3(b)–(d) shows the OA results with 10, 100, and 1000 labeled samples per class, respectively, as a function of the noise standard deviation σ. As shown by the plots, it can be observed that the classification OA approaches the optimal value $\mathrm{OA}_{\mathrm{opt}}$ 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 that are close to optimal. In turn, it can also be seen in Fig. 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 $\mathrm{OA}_{\mathrm{opt}}$). 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 results provided by using only spectral information. For illustrative purposes, Fig. 2(b) and (c) shows 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 regard to the optimal classification that can be achieved $(\mathrm{OA}_{\mathrm{opt}} = 75.95\%)$ is clearly noticeable (about 20.46%), thus revealing the importance of including the spatial prior after classification.



(a) (b)

Fig. 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 dashed–dot lines represent random selection and maximum-entropy-based active selection, respectively.

Impact of Incorporating Unlabeled Samples: In this section, 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: 1) random and 2) maximum-entropy-based. The latter corresponds to selecting unlabeled samples close to the boundaries between regions in a feature space. Fig. 4 shows the OA results obtained for the proposed algorithm as a function of the number of unlabeled samples for two different analysis scenarios: 1) fixed number of labeled training samples, $L = 400$ (40 per class) and noise standard deviation $\sigma = 0.4$, and 2) fixed $L = 500$ (50 per class) and $\sigma = 0.45$. The theoretical OA, termed as $\mathrm{OA}_{\mathrm{opt}} \equiv 100(1 - P_e)\%$, where P_e denotes the union bound in this problem, is also plotted. After analyzing the results reported in Fig. 4, the following general observations can be made.

1) The inclusion of a spatial prior improves the classification OA.
2) The inclusion of unlabeled samples improves the segmentation OA by roughly 15% in Fig. 4(a) and by

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

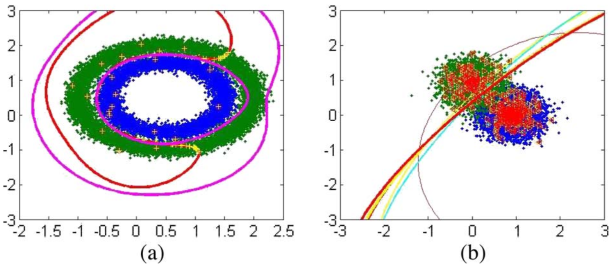


Fig. 5. Changes in the definition of the boundary by the proposed classifier in a binary classification problem as the number of unlabeled samples (selected using a maximum-entropy-based criterion) is increased.

approximately 10% in Fig. 4(b). This effect is observed for all considered numbers of unlabeled samples.

3) Finally, it is clear from Fig. 4 that the maximum-entropy-based active selection performs uniformly better than the random active selection in terms of OAs.

Impact of the Considered Active-Selection Approach: The main objective of this section is to provide an informal justification about why the proposed method for maximum-entropy-based active selection of unlabeled samples performs accurately in the experiments. Fig. 5, with 20 labeled samples (ten per class), shows the improvements in the definition of the separation boundaries established by our proposed classifier as the number of unlabeled samples increases, using a toy example. In Fig. 5(a) in which the noise standard deviation is set to $\sigma = 0.1$, the red circles denote the labeled samples. The red line is the classifier boundary defined with zero 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 has pushed the contour outward, 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 the unlabeled samples is 98.6%. A similar phenomenon is observed in Fig. 5(b) in which $\sigma = 0.3$ is considered. For illustrative purposes, Table I shows the OA results as a function of the number of unlabeled samples for the example shown in Fig. 5(b). Each column of Table I corresponds to a different type of color/thickness in Fig. 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.

B. 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 the National Aeronautics and Space Administration Jet Propulsion Laboratory.

1) The first data set used in experiments was collected over the Valley of Salinas, Southern California, in 1998. It contains 217 × 512 pixels and 224 spectral bands from 0.4 to 2.5 μm, with nominal spectral resolution of 10 nm. It was taken at low altitude with a pixel size of 3.7 m. The data include vegetables, bare soils, and vineyard fields. The upper leftmost part of Fig. 6 shows the entire scene (with overlaid ground-truth areas) and a subscene of the data set (called hereinafter as Salinas A), outlined by a red rectangle. The Salinas A subscene comprises 83 × 86 pixels and is known to represent a difficult classification scenario with highly mixed pixels [39], where the lettuce fields can be found for different weeks since being planted. The upper rightmost part of Fig. 6 shows the available ground-truth regions for the scene, and the bottom part of Fig. 6 shows some photographs taken in the field for the different agricultural fields at the time of data collection.

2) The second data set used in the experiments is the well-known AVIRIS Indian Pines scene, collected over Northwestern Indiana in June of 1992 [1]. 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 m by pixel. For illustrative purposes, Fig. 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,[7] a fact which has made this scene a widely used benchmark for testing the accuracy of hyperspectral data-classification and segmentation algorithms.

Experiments With the Full AVIRIS Salinas Data Set: Table II 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 II, the proposed algorithm obtains very good OAs with limited training samples. Specifically, with only 240 labeled pixels (15 per class), the OAs obtained are 93.87% $(U = 0)$, 94.70% $(U = L)$, and 95.13% $(U = 2 \times L)$, which are better than the best result reported in [10] 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 [10] 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 paper are only slightly lower than those reported in [39] using a multilayer perceptron (MLP) neural-network classifier trained with 2%

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

TABLE I
OA (IN PERCENT) AS A FUNCTION OF THE NUMBER OF UNLABELED SAMPLES IN THE TOY EXAMPLE SHOWN IN FIG. 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

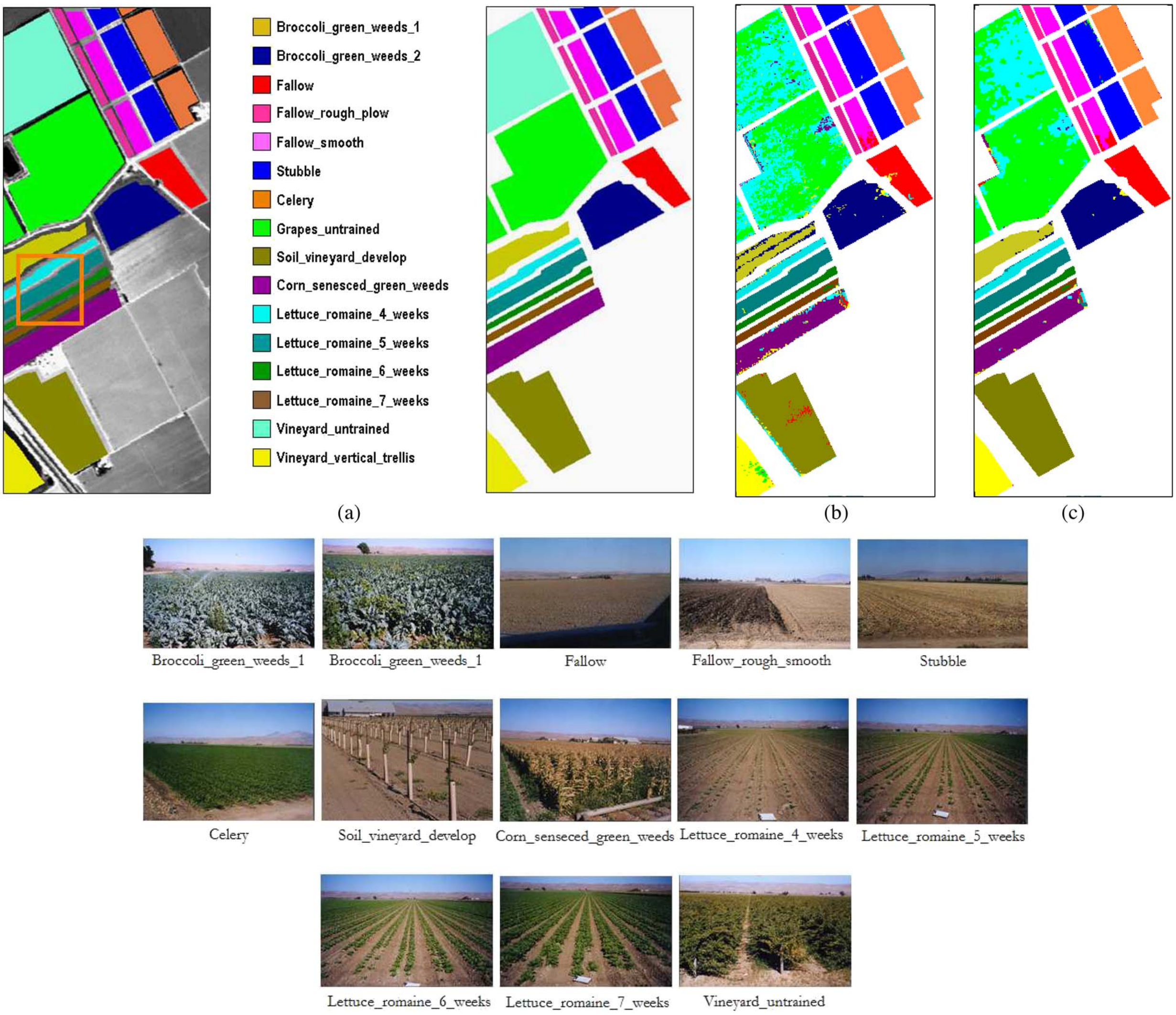


Fig. 6. AVIRIS Salinas data set along with the classification maps by using $L = 128$ and $U = 256$. (a) Right side: Original image at 488-nm wavelength with the red rectangle indicating a subscene called Salinas A. (a) Left side: Ground-truth map containing 16 mutually exclusive land-cover classes. (b) Classification map (OA $= 82.55\%$). (c) Segmentation map (OA $= 91.14\%$). (Bottom) Photographs taken at the site during data collection.

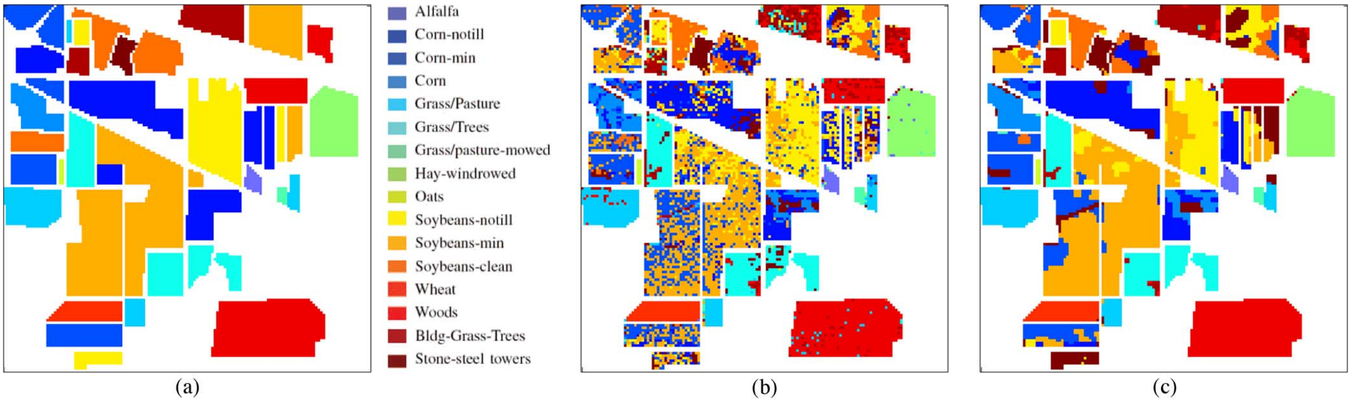


Fig. 7. AVIRIS Indian Pines scene along with the classification and segmentation maps by using $L = 160$ and $U = 288$. (a) Ground-truth map containing 16 mutually exclusive land-cover classes. (b) Classification map (OA $= 62.98\%$). (c) Segmentation map (OA $= 74.98\%$).

TABLE II
CLASSIFICATION (IN PARENTHESES) AND SEGMENTATION OAs (IN PERCENT) 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 AFTER FIVE 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)

TABLE III
SEGMENTATION OAs (IN PERCENT) ACHIEVED AFTER APPLYING THE PROPOSED ALGORITHM TO THE AVIRIS SALINAS A SUBSCENE 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 AFTER TEN 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

of the available ground-truth pixels and with multidimensional morphological feature extraction prior to classification (the maximum OA reported in [39] 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 II that the inclusion of a spatial prior significantly improves the results obtained by using the spectral information only (approximately, on the order of 6% increase in OA). Furthermore, the inclusion of unlabeled samples in the proposed approach increases the OA by approximately 1% or 2% with regard to the case in which only labeled samples are used. The aforementioned results confirm our introspection (already reported in the simulated data experiments) that the proposed approach can greatly benefit from the inclusion of spatial prior and unlabeled samples in order to increase the already good classification accuracies obtained using the spectral information only. Fig. 6(b) and (c) shows the classification and segmentation maps. Effective results can be seen in these maps.

Experiments With the AVIRIS Salinas A Subscene: In this experiment, we use a subscene of the Salinas data set, which comprises 83 × 86 pixels and six classes. As mentioned earlier, this subscene is known to represent a challenging classification scenario due to the similarity of the different lettuce classes comprised by the subscene, which 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-m pixels of the scene. Table III reports the segmentation (with spatial prior) scores achieved for the proposed method with the AVIRIS Salinas A subscene in which the accuracy results are displayed for different numbers of labeled samples (ranging from three to ten 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 III. As shown by Table III, the proposed algorithm achieved a segmentation OA of up to 99.28% for $U = 4 \times L$ and only five 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 [10] and [39] for the classes included in the AVIRIS Salinas A subscene using an SVM-based classifier and an MLP-based classifier with multidimensional morphological feature extraction, respectively.

Experiments With the AVIRIS Indian Pines Data Set: Table IV 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 the range from $U = 0$ (no unlabeled samples) to $U = 32 \times k$, with $k = 0, 1, \ldots, 9$. As with the previous experiments, the number of labeled samples in Table IV 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 in Table IV, 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 on 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 eight 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 semisupervised cluster SVMs introduced in [18]. Specifically, when 128 labeled samples (eight samples per class) are used by our proposed method, the OA of the proposed approach

TABLE IV

CLASSIFICATION (IN PARENTHESES) AND SEGMENTATION OAS (IN PERCENT) 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 $k = 0, 1, \ldots, 9$. 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 AFTER TEN 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)

is 69.79% ($U = 288$, obtained by active selection), which is comparable with the best result 69.82% reported in [18] (using 519 labeled samples). For illustrative purposes, Fig. 7(b) and (c) shows the classification and segmentation maps, respectively. These figures show the effective results without severe block artifacts. Notice that the results shown in Figs. 6 and 7 are obtained with just eight and ten samples per class, respectively. To give an idea of the quality of this result, we note that the recent semisupervised technique [18] takes approximately two times more training samples to achieve a comparable performance, if we take into account only classification results, and four times more, if we use spatial information (see Table IV).

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 (6), a condition without which the active selection would fail to provide good results [33].

VII. CONCLUSION AND FUTURE LINES

In this paper, we have introduced a new semisupervised 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 improve the estimation of the class distributions. By adopting a spatial multilevel logistic prior and computing the MAP 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 improves significantly with respect to the classification results proposed by the same algorithm using only the learned 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 toward testing the proposed segmentation approach in different analysis scenarios dominated by the limited availability of training samples *a priori*.

ACKNOWLEDGMENT

The authors would like to thank Prof. D. Landgrebe for making the AVIRIS Indian Pines hyperspectral data set available to the community, and Dr. L. Johnson and Dr. J. A. Gualtieri for providing the AVIRIS Salinas data set used in our experiments.

REFERENCES

[1] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.

[2] 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 Sens. Environ.*, vol. 113, pp. 110–122, Sep. 2009.

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

[4] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Proc. 16th Annu. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2002.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. New York: Springer-Verlag, 2007.

[6] B. Scholkopf and A. Smola, *Learning With Kernels—Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press, 2002.

[7] D. Böhning, "Multinomial logistic regression algorithm," *Ann. Inst. Stat. Math.*, vol. 44, no. 1, pp. 197–200, Mar. 1992.

[8] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[9] M. Fauvel, J. Benediktsson, J. Chanussot, and J. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[10] J. Plaza, A. Plaza, and C. Barra, "Multi-channel morphological profiles for classification of hyperspectral images using support vector machines," *Sensors*, vol. 9, no. 1, pp. 196–218, 2009.

[11] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.

[12] J. Borges, J. Bioucas-Dias, and A. Marçal, "Evaluation of Bayesian hyperspectral imaging segmentation with a discriminative class learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Barcelona, Spain, 2007, pp. 3810–3813.

[13] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo, "On semi-supervised classification," in *Proc. 18th Annu. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2004, pp. 721–728.

[14] O. Chapelle, M. Chi, and A. Zien, "A continuation method for semi-supervised SVMs," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 185–192.

[15] Y. Zhong, L. Zhang, B. Huang, and P. Li, "An unsupervised artificial immune classifier for multi/hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 2, pp. 420–431, Feb. 2006.

[16] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for the semisupervised classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3363–3373, Nov. 2006.

[17] G. Camps-Valls, T. Bandos, and D. Zhou, "Semisupervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.

[18] D. Tuia and G. Camps-Valls, "Semisupervised hyperspectral image classification with cluster kernels," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 224–228, Apr. 2009.

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

[20] Y. D. Rubinstein and T. Hastie, "Discriminative vs. informative learning," in *Proc. ACM KDD*, 1997, pp. 49–53.

[21] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[22] J. Bernardo and A. Smith, *Bayesian Theory*. Chichester, U.K.: Wiley, 1994.

[23] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[24] K. Lange, D. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *J. Comput. Graph. Stat.*, vol. 9, pp. 1–59, 2000.

[25] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[26] S. Z. Li, *Markov Random Field Modeling in Computer Vision*. London, U.K.: Springer-Verlag, 1995.

[27] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. R. Stat. Soc. B*, vol. 36, no. 2, pp. 192–236, 1974.

[28] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, 2nd ed. New York: Springer-Verlag, 2001.

[29] D. Greig, B. Porteous, and A. Seheult, "Exact maximum *a posteriori* estimation for binary images," *J. R. Stat. Soc. B*, vol. 51, no. 2, pp. 271–279, 1989.

[30] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[31] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[32] S. Bagon, Matlab Wrapper for Graph Cut, Dec. 2006. [Online]. Available: http://www.wisdom.weizmann.ac.il/~bagon

[33] D. MacKay, "Information-based objective functions for active data selection," *Neural Comput.*, vol. 4, no. 4, pp. 590–604, Jul. 1992.

[34] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[35] W. Di and M. M. Crawford, "Locally consistent graph regularization based active learning for hyperspectral image classification," in *Proc. 2nd WHISPERS*, 2010.

[36] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2000.

[37] C. Scott and R. Nowak, "Minimax-optimal classification with dyadic decision trees," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1335–1353, Apr. 2006.

[38] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1270–1279, Jul. 1993.

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

Jun Li received the B.S. degree in geographic information systems from Hunan Normal University, Hunan, China, in 2004 and the M.E. degree in remote sensing from Peking University, Beijing, China, in 2007.

Since 2007, she has been a Research Fellow with the Department of Electrical and Computer Engineering, Instituto de Telecomunicações and Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal. Her research interests include hyperspectral classification and segmentation, unmixing, signal processing, remote sensing, etc.

José M. Bioucas-Dias (S'87–M'95) received the E.E., M.Sc., Ph.D., and "Agregado" degrees in electrical and computer engineering from the Instituto Superior Técnico (IST), the Engineering School, Technical University of Lisbon (UTLisbon), Lisbon, Portugal, in 1985, 1991, 1995, and 2007, respectively.

Since 1995, he has been with the Department of Electrical and Computer Engineering, IST. He is also a Senior Researcher with the Communication Theory and Pattern Recognition Group of the Institute of Telecommunications, a private not-for-profit research institution. His research interests include signal and image processing, pattern recognition, optimization, and remote sensing. He is involved in several national and international research projects and networks, including the Marie Curie Actions "Hyperspectral Imaging Network (HYPER-I-NET)" and the "European Doctoral Program in Signal Processing (SIGNAL)."

Dr. Bioucas-Dias is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, and a Guest Editor of a special issue of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He has been a member of program/technical committees of several international conferences, including CVPR, ICPR, ICIAR, IGARSS, ICIP, SPIE, EMMCVPR, ISVC, and WHISPERS.

Antonio Plaza (M'05–SM'07) received the M.S. and Ph.D. degrees in computer engineering from the University of Extremadura, Caceres, Spain, in 1997 and 2002, respectively.

He was a Visiting Researcher with the Remote Sensing Signal and Image Processing Laboratory, University of Maryland Baltimore County, Baltimore; with the Applied Information Sciences Branch, Goddard Space Flight Center, Greenbelt, MD; and with the AVIRIS Data Facility, Jet Propulsion Laboratory, Pasadena, CA. Since 2000, he has been an Associate Professor with the Department of Technology of Computers and Communications, University of Extremadura, Caceres, Spain, where he was an Assistant Professor from 1997 to 1999. He is the Coordinator of the Hyperspectral Imaging Network (Hyper-I-Net), which is a European project designed to build an interdisciplinary research community focused on hyperspectral imaging activities. He has been a Proposal Reviewer with the European Commission, the European Space Agency, and the Spanish Government, and has also served as a Reviewer for more than 30 different journals. He is the author or coauthor of more than 230 publications on remotely sensed hyperspectral imaging, including more than 40 Journal Citation Report papers, book chapters, and conference proceeding papers. He has coedited a book on high-performance computing in remote sensing and several special issues on remotely sensed hyperspectral imaging for different journals. He has served as a Reviewer for more than 120 manuscripts submitted to the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include remotely sensed hyperspectral imaging, pattern recognition, signal and image processing, and efficient implementation of large-scale scientific problems on parallel and distributed computer architectures.

Dr. Plaza is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING on hyperspectral image analysis and signal processing.