# Hyperspectral Image Segmentation Using a New Bayesian Approach With Active Learning

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

***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 algorithm. Second, we use the information acquired in the previous step to segment the hyperspectral image using a multilevel 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 paper is the introduction of a new active sampling approach, called modified breaking ties, which is able to provide an unbiased sampling. Furthermore, 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 $\alpha$-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*—Active learning, graph cuts, hyperspectral image segmentation, ill-posed problems, integer optimization, mutual information (MI), sparse multinomial logistic regression (MLR).**

## I. Introduction

WITH THE recent developments in remote sensing instruments, hyperspectral images are now widely used in different application domains [1]. The special characteristics of hyperspectral data sets bring difficult processing problems. Obstacles, such as the Hughes phenomenon [2], 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 [3]–[5]. 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 [6]–[8], effectively tackle the aforementioned difficulties. Specifically, support vector machines (SVMs) [9] 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 semisupervised classifications of hyperspectral data using limited training samples [1], [3], [10]–[15]. 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 directly models the posterior densities instead of the joint probability distributions. The distinguishing features of discriminative classifiers have been reported in the literature before [7], [8], [17]. For instance, effective sparse MLR (SMLR) methods are available in the literature [18]. These ideas have been applied to hyperspectral image classification [5], [19], [20], 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, multiple classifiers, and decision fusion [21], among many others [1]. Active learning, which is another active research topic, has been widely studied in the literature [22]–[28]. These studies are based on different principles, such as the evaluation of the disagreement between a committee of classifiers [25], the use of hierarchical classification frameworks [24], [27], unbiased query by bagging [28], or the exploitation of a local proximity-based data regularization framework [26].

In this paper, 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 regard 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

Manuscript received August 5, 2010; revised December 28, 2010 and February 25, 2011; accepted March 3, 2011. Date of publication May 11, 2011; date of current version September 28, 2011. This work was supported in part by the European Commission under the Marie Curie training Grant MEST-CT-2005-021175, by the Instituto de Telecomunicações under the IT Grant, and by the MRTN-CT-2006-035927 and AYA2008-05965-C04-02 projects.

J. Li and A. Plaza are with the Hyperspectral Computing Laboratory (HyperComp), Department of Technology of Computers and Communications, University of Extremadura, E-10071 Caceres, Spain (e-mail: junli@unex.es; aplaza@unex.es).

J. M. Bioucas-Dias is with the Instituto de Telecomunicações, Instituto Superior Técnico, Universidade Técnica de Lisboa, 1049-001 Lisboa, Portugal (e-mail: bioucas@lx.it.pt).

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

Digital Object Identifier 10.1109/TGRS.2011.2128330

is whether the considered classifier (in this paper, 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 [1], [5], [13], [29]. 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 paper, 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: 1) learning stage, using the MLR via variable splitting and augmented Lagrangian (LORSAL) [30] algorithm to infer the class distributions, and 2) segmentation stage, which infers the labels from a posterior distribution built on the learned class distributions and on a multilevel logistic (MLL) prior [31]. 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 $\alpha$-expansion algorithm [32]. 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 paper, different strategies are used to implement active learning in addition to random sampling (RS): 1) the mutual information (MI) between the MLR regressors and the class labels [22], [23]; 2) a criterion called breaking ties (BT) [33]; and 3) our proposed version called modified BT (MBT), which is also intended to guarantee unbiased samplings among the classes.

The remainder of this paper is organized as follows. Section II formulates the hyperspectral image segmentation problem. Section III describes the proposed approach. Section IV presents the active learning algorithms considered in this paper. Section V reports segmentation results based on both simulated and real hyperspectral data sets in several ill-posed scenarios. Comparisons with state-of-the-art algorithms are also included and thoroughly described in this section. Finally, Section VI concludes with a few remarks and hints at plausible future research lines.

## II. 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 $\mathcal{D}_L \equiv \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_L, y_L)\} \in (\mathbb{R}^d \times \mathcal{L})^L$ be a training set where $L$ denotes the total number of available labeled samples. With the aforementioned 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[1]

$$p(\mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \tag{1}$$

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 independence 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:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \frac{1}{p(\mathbf{x})} p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \\ &= \alpha(\mathbf{x}) \prod_{i=1}^{i=n} \frac{p(y_i|\mathbf{x}_i)}{p(y_i)} p(\mathbf{y}) \end{aligned} \tag{2}$$

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

$$\widehat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{L}^n} \left\{ \sum_{i=1}^{n} (\log p(y_i|\mathbf{x}_i) - \log p(y_i)) + \log p(\mathbf{y}) \right\}. \tag{3}$$

In the present approach, the densities $p(y_i|\mathbf{x}_i)$ are modeled as MLRs [16], whose regressors are learned via the LORSAL algorithm [30]. As prior $p(\mathbf{y})$ on the labelings $\mathbf{y}$, we adopt an MLL Markov random field (MRF) [31], which encourages neighboring pixels to have the same label. The MAP labeling/segmentation $\widehat{\mathbf{y}}$ is computed via the $\alpha$-expansion algorithm [34], a min-cut-based tool to efficiently solve a class of integer optimization problems of which the MAP segmentation in (3) is an example.

## III. Proposed Approach

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

$$p(y_i = k|\mathbf{x}_i, \boldsymbol{\omega}) \equiv \frac{\exp\left(\boldsymbol{\omega}^{(k)} \mathbf{h}(\mathbf{x}_i)\right)}{\sum_{k=1}^{K} \exp\left(\boldsymbol{\omega}^{(k)} \mathbf{h}(\mathbf{x}_i)\right)} \tag{4}$$

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

where $\mathbf{h}(\mathbf{x}) \equiv [h_1(\mathbf{x}), \ldots, h_l(\mathbf{x})]^{\mathrm{T}}$ is a vector of $l$ fixed functions of the input, often termed *features*, and $\boldsymbol{\omega} \equiv [\boldsymbol{\omega}^{(1)^{\mathrm{T}}}, \ldots, \boldsymbol{\omega}^{(K)^{\mathrm{T}}}]^{\mathrm{T}}$ denotes the logistic regressors. Since the density in (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., $\boldsymbol{\omega} \equiv [\boldsymbol{\omega}^{(1)^{\mathrm{T}}}, \ldots, \boldsymbol{\omega}^{(K-1)^{\mathrm{T}}}]^{\mathrm{T}}$.

It should be noted that function $\mathbf{h}$ may be linear, i.e., $\mathbf{h}(\mathbf{x}_i) = [1, x_{i,1}, \ldots, x_{i,d}]^{\mathrm{T}}$, 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 $\mathbf{h}(\mathbf{x}_i) = [1, K_{\mathbf{x}_i,\mathbf{x}_1}, \ldots, K_{\mathbf{x}_i,\mathbf{x}_l}]^{\mathrm{T}}$, 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 [11]. If we denote by $\gamma$ 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.

### A. LORSAL

In our context, learning the class densities amounts to estimating the logistic regressors $\boldsymbol{\omega}$. Following the principles of the SMLR algorithm [18], the estimation of $\boldsymbol{\omega}$ amounts to computing the MAP estimate

$$\widehat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \ell(\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}) \tag{5}$$

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

$$\ell(\boldsymbol{\omega}) \equiv \log \prod_{i=1}^{L} p(y_i|\mathbf{x}_i, \boldsymbol{\omega}) \tag{6}$$

$$p(\boldsymbol{\omega}) \propto \exp\left(-\lambda \|\boldsymbol{\omega}\|_1\right) \tag{7}$$

is a Laplacian prior promoting the sparsity on $\boldsymbol{\omega}$ ($\|\boldsymbol{\omega}\|_1$ denotes the $l_1$ norm of $\boldsymbol{\omega}$) with $\lambda$ 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 (5) is difficult because the term $\ell(\boldsymbol{\omega})$ is nonquadratic and the term $\log p(\boldsymbol{\omega})$ is nonsmooth. A majorization–minimization framework [35] has recently been used in [18], [20], [23], and [36] to decompose the problem in (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 data sets with a large number of features, with a large number of classes, or both. The fast SMLR (FSMLR) [19] 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, particularly for hyperspectral data sets with high-dimensional features.

In this paper, we resort to the recently introduced LORSAL algorithm [30] to learn the MLR regressors given by (5). By replacing $\log p(\boldsymbol{\omega})$ in (4) 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 nonsmooth 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 $\gamma K^2$ and $\gamma$, respectively.

### B. 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 [37] and has been widely used in image segmentation problems (see, e.g., [5], [20], [36], and [38]).

According to the Hammersly–Clifford theorem [39], the density associated with an MRF is a Gibbs' distribution [37]. Thus, the prior model has the structure

$$p(\mathbf{y}) = \frac{1}{Z} e^{\left(-\sum_{c \in \mathcal{C}} V_c(\mathbf{y})\right)} \tag{8}$$

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

$$-V_c(\mathbf{y}) = \begin{cases} \upsilon_{y_i}, & \text{if } |c| = 1 \text{ (single clique)} \\ \mu_c, & \text{if } |c| > 1 \text{ and } \forall_{i,j \in c} y_i = y_j \\ -\mu_c, & \text{if } |c| > 1 \text{ and } \exists_{i,j \in c} y_i \neq y_j \end{cases} \tag{9}$$

where $\mu_c$ is a nonnegative constant.

The potential function in (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 $\upsilon_{y_i}$ and $\mu_c$. For example, the model generates

[2]A clique is a single term that denotes a set of pixels that are neighbors of one another.

texturelike regions if $\mu_c$ depends on $c$ and bloblike regions otherwise [31]. In this paper, we take $e^{\upsilon_{y_i}} = c^{te}$, implying that we are assuming equiprobable classes, and $\mu_c = (1/2)\mu > 0$ and assume that a clique consists either of a single pixel, i.e., $c = \{i\}$, or a pair of neighboring pixels, i.e., $c = \{i, j\}$, where $i$ and $j$ are neighbors; then, (8) can be rewritten as

$$p(\mathbf{y}) = \frac{1}{Z} e^{\mu \sum\limits_{\{i,j\} \in \mathcal{C}} \delta(y_i - y_j)}, \tag{10}$$

where $\delta(y)$ is the unit impulse function.[3] 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 $\mu$ controls the degree of smoothness.

### C. 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 (3), under the equiprobable class assumption, the MAP segmentation is finally given by

$$\widehat{\mathbf{y}} = \arg \min_{\mathbf{y} \in \mathcal{L}^n} \sum_{i \in \mathcal{S}} -\log p(y_i|\widehat{\boldsymbol{\omega}}) - \mu \sum_{i,j \in C} \delta(y_i - y_j) \tag{11}$$

where $p(y_i|\widehat{\boldsymbol{\omega}}) \equiv p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$ computed at $\widehat{\boldsymbol{\omega}}$. Minimization of (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 [32], [34], [40], loopy belief propagation [41], [42], and tree-reweighed message passing [43] are efficient tools to tackle this class of optimization problems. In this paper, we use the $\alpha$-expansion algorithm [34] to solve our integer optimization problem [44]. 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 pseudocode for the proposed supervised segmentation algorithm with discriminative class learning and MLL prior is shown in Algorithm 1.

**Algorithm 1** Supervised Segmentation Algorithm (LORSAL-MLL)

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

### D. 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 4 (see Appendix), which has a complexity of $O(\gamma^2 K)$, and by the $\alpha$-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] $\delta(0) = 1$ and $\delta(y) = 0$ for $y \neq 0$.

## IV. Active Learning

In this paper, 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: 1) MI-based criterion [22], [23]; 2) BT algorithm [33]; and 3) our proposed MBT scheme.

### A. MI-Based Active Learning

The first active learning scheme considered is an MI-based criterion [22], [23] that maximizes the MI between the MLR regressors and the class labels. Let $I(\boldsymbol{\omega}; y_i|\mathbf{x}_i)$ denote the MI between the MLR regressors and the class label $y_i$. Following [22], the new vector $\mathbf{x}_i$ is selected according to

$$\widehat{\mathbf{x}}_i^{\text{MI}} = \arg \max_{\mathbf{x}_i, i \in \mathcal{S}_U} \quad I(\boldsymbol{\omega}; y_i|\mathbf{x}_i), \tag{12}$$

where (see [22] for more details)

$$I(\boldsymbol{\omega}; y_i|\mathbf{x}_i) = (1/2) \log \left( |\mathbf{H}^{\text{MI}}| / \mathbf{H} \right). \tag{13}$$

Here, $\mathbf{H}$ is the posterior precision matrix, i.e., the Hessian of minus the log-posterior $\mathbf{H} \equiv \nabla^2(-\log p(\widehat{\boldsymbol{\omega}}|\mathcal{D}_L))$ [45] and $\mathbf{H}^{\text{MI}}$ 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 [23], [46]. Under this assumption, we can compute $\mathbf{H}^{\text{MI}}$ as follows:

$$\mathbf{H}^{\text{MI}} = \mathbf{H} + \left( \text{diag}\left(\mathbf{p}_i(\widehat{\boldsymbol{\omega}})\right) - \mathbf{p}_i(\widehat{\boldsymbol{\omega}}) \mathbf{p}_i(\widehat{\boldsymbol{\omega}})^{\text{T}} \right) \otimes \mathbf{h}(\mathbf{x}_i) \mathbf{h}(\mathbf{x}_i)^{\text{T}} \tag{14}$$

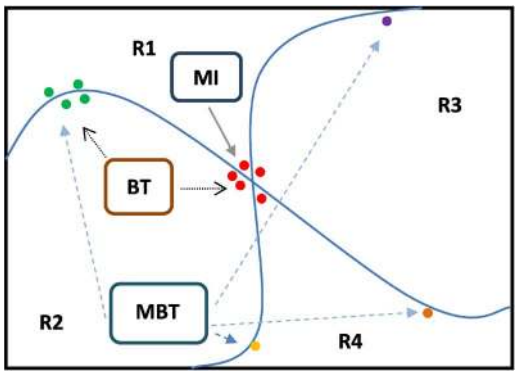


Fig. 1. Graphical illustration of the MI, BT, and MBT active learning approaches using a toy example.

where $\mathbf{p}_i(\widehat{\boldsymbol{\omega}}) \equiv [p_{i,1}, \ldots, p_{i,K}]^{\mathrm{T}}$, $p_{i,k} \equiv p(y_i = k|\mathbf{x}_i, \widehat{\boldsymbol{\omega}})$ for $k = 1, \ldots, K$, and $\otimes$ is the Kronecker product. Therefore, (13) turns to

$$I(\boldsymbol{\omega}; y_i|\mathbf{x}_i) = (1/2)\log\left(1 + \prod_{k=1}^{K} p_{i,k}\mathbf{x}_i^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{x}_i\right). \quad (15)$$

According to (15), the function in (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 shown in Fig. 1, in which a toy example with four simulated regions is used for demonstration purposes. As shown by Fig. 1, the MI focuses on the most complex area (boundary between the four regions).

### B. BT Active Learning

The BT active learning algorithm [33] was proposed to achieve diversity in the sampling, thus alleviating the bias in the MI-based sampling. The decision criterion is

$$\widehat{\mathbf{x}}_i^{\mathrm{BT}} = \arg\min_{\mathbf{x}_{i,i}\in\mathcal{S}_U} \left\{ \max_{k\in\mathcal{L}} p(y_i = k|\mathbf{x}_i, \widehat{\boldsymbol{\omega}}) - \max_{k\in\mathcal{L}\setminus\{k^+\}} p(y_i = k|\mathbf{x}_i, \widehat{\boldsymbol{\omega}}) \right\} \quad (16)$$

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 Fig. 1, which shows 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 section, we propose a new modified scheme (called MBT) which promotes even more diversity in the sampling process.

### C. 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:

$$\begin{aligned} &\textbf{do} \\ &\quad s = \text{next class} \\ &\quad \text{select } \mathcal{S}_{U_s} \\ &\quad \widehat{\mathbf{x}}_i^{\mathrm{MBT}} = \arg\max_{\mathbf{x}_i, i\in\mathcal{S}_{U_s}, k\in\mathcal{L}\setminus\{s\}} p(y_i = k|\mathbf{x}_i, \widehat{\boldsymbol{\omega}}), \\ &\textbf{while } \text{stop rule} \end{aligned} \quad (17)$$

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 (17), both intended to promote diversity in the selection process as compared with the BT criterion.

1) 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.
2) 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 Fig. 1, the main advantage of the proposed MBT with regard 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 paper, i.e., MI, BT, and MBT, it is now important to emphasize that (12), (16), and (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, round$(u/K) + 1$ new samples per class are selected according to (17), where function round$(\cdot)$ simply rounds toward the nearest integer value. Second, we run (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 multiclass problems, the three considered strategies may lead to different labelings. In turn, when $u$ is very small, the performances of BT and MBT become similar.

To conclude this section, Algorithm 2 shows the pseudocode 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.

**Algorithm 2** LORSAL Using Active Learning (LORSAL-AL)

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

**Algorithm 3** Supervised Segmentation Algorithm Using Active Learning (LORSAL-AL-MLL)

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

## V. 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.[4] 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 $\rho$. 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, for example, $L_k$, 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 ten independent Monte Carlo runs with respect to $\mathcal{D}_{L_i}$.

The remainder of this section is organized as follows. Section V-A reports experiments with simulated data, with Section V-A.1 conducting an evaluation of the LORSAL algorithm, Section V-A.2 evaluating the impact of the spatial prior, and Section V-A.3 evaluating the impact of the active learning approaches. Section V-B evaluates the performance of the proposed algorithm using four real hyperspectral scenes collected by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS), operated by NASA Jet Propulsion Laboratory, and by the Reflective Optics Imaging Spectrometer System (ROSIS), operated by the German Aerospace Agency.

### A. 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 $128 \times 128$ MLL distribution with $\mu = 2$. The feature vectors are simulated according to

$$\mathbf{x}_i = \mathbf{m}_{y_i} + \mathbf{n}_i, \qquad i \in \mathcal{S}, y_i \in \mathcal{L} \tag{18}$$

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 Sections V-A.1 and A.2, we will not consider the active learning procedure (i.e., $L = L_i$) because our focus in these two sections 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, Section V-A.3 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 suboptimal 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.

*1) Evaluation of the LORSAL Algorithm:* In this section, we generate the simulated hyperspectral data according to the model in (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[5] with $d = 224$, $K = 10$, $L = 1000$, and $\sigma = 1$.

[4]The normalization is simply given by $\mathbf{x}_i := \mathbf{x}_i / (\sqrt{\sum \|\mathbf{x}_i\|^2})$, for $i = 1, \ldots, n$, where $\mathbf{x}_i$ is a spectral vector.

[5]Available online: http://speclab.cr.usgs.gov.

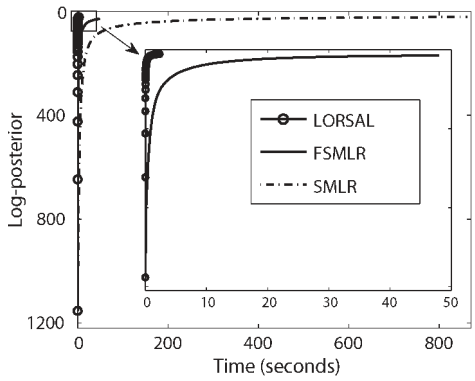


Fig. 2. Evaluation of the log-posterior in (5) as a function of the computing time (measured in a desktop PC with Intel Core 2 Duo CPU at 2.40 GHz and 4 GB of RAM memory) for LORSAL, FSMLR, and SMLR algorithms.

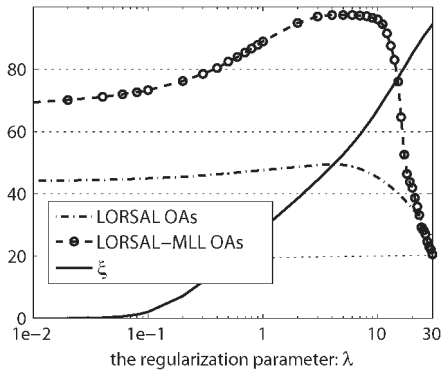


Fig. 3. Evaluation of the impact of the regularization parameter $\lambda$ on the OA and on the level of sparsity $\xi$.

In our first experiment, we illustrate the computational efficiency of the LORSAL algorithm. Fig. 2 shows 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 Fig. 2, LORSAL is by far the fastest algorithm. For a similar log-posterior, the LORSAL algorithm took about 2 s 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 around 48 and 880 s, respectively, in the same computing environment.

As already mentioned, the regularization parameter $\lambda$ in (7) controls the sparseness of the regressors, which is essential to the generalization capacity. However, an inappropriate value of $\lambda$ may lead to overfitting or underfitting scenarios. In practice, we estimate $\lambda$ by using cross-validation sampling [47] over the initial training set. Nevertheless, in our second experiment, we conduct an analysis of the impact of $\lambda$ on the achieved performance. Let $\xi = 100 \times (n_{\omega_0}/n_\omega)\%$, where $n_\omega$ and $n_{\omega_0}$ denote the number of components and zeros in $\boldsymbol{\omega}$, respectively. Fig. 3 shows the OA and $\xi$ as a function of $\lambda$, for $10^{-2} \leq \lambda \leq 30$. The impact of $\lambda$ on the sparsity of $\omega$ is clear. The higher values of OA are obtained for $\lambda \in [2, 10]$ corresponding to levels of sparsity $\xi \in [50, 60]\%$.

*2) 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 (18) is shown in Fig. 4(a), where the labels $y_i = 1, 2$ correspond to $\xi_i = -1$, $+1$, respectively. In this problem, the theoretical OA, given by $\mathrm{OA}_{\mathrm{opt}} \equiv 100(1 - P_e)\%$ and corresponding to the minimal probability of error [48] is

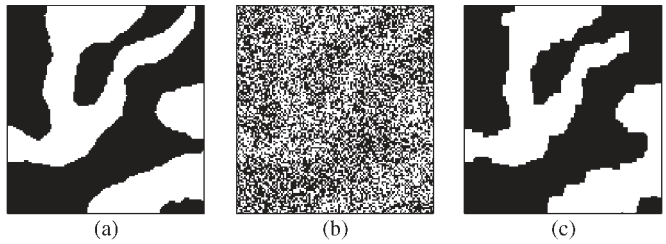


Fig. 4. Classification and segmentation results obtained with the proposed algorithm. The simulated data set was generated according to (18) with $d = 500$, $\sigma = 1.5$, and $\mu = 2$. (a) Simulated binary map. (b) Classification map produced by the LORSAL algorithm using $L = 100$ labeled samples without active learning ($\mathrm{OA} = 60.13\%$, with $\mathrm{OA}_{\mathrm{opt}} = 71.91\%$, see text). (c) Segmentation map, same as (b) but using the MLL spatial prior ($\mathrm{OA} = 92.48\%$).

$$P_e = \frac{1}{2}\mathrm{erfc}\left(\frac{1+\lambda_0}{\sqrt{2}\sigma}\right)p_0 + \frac{1}{2}\mathrm{erfc}\left(\frac{1-\lambda_0}{\sqrt{2}\sigma}\right)p_1 \qquad (19)$$

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 paper, we empirically concluded 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 shown in Fig. 5 where we studied the impact of the spatial prior. Here, Fig. 5(a) shows the OA results as a function of $\mu$. For the considered problem, with $2 \leq \mu \leq 6$, the LORSAL-ALL algorithm obtained good segmentation results. It should be noted that ten independent Monte Carlo runs were conducted in these experiments, and we report only the mean scores obtained. The following conclusions may be drawn from Fig. 5.

1) The best overall results are obtained by the proposed segmentation algorithm (in all cases, the classification accuracies and the values of $\mathrm{OA}_{opt}$ 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 Fig. 5(b)].
2) The classification OA approaches the optimal value $\mathrm{OA}_{\mathrm{opt}}$ as the number of labeled samples increases [see Fig. 5(c)]. However, the number of labeled samples needs to be relatively high in order to obtain classification accuracies which are close to optimal.
3) For a fixed number of training samples, the classification accuracy of our proposed method decreases as the number of bands increases [see Fig. 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

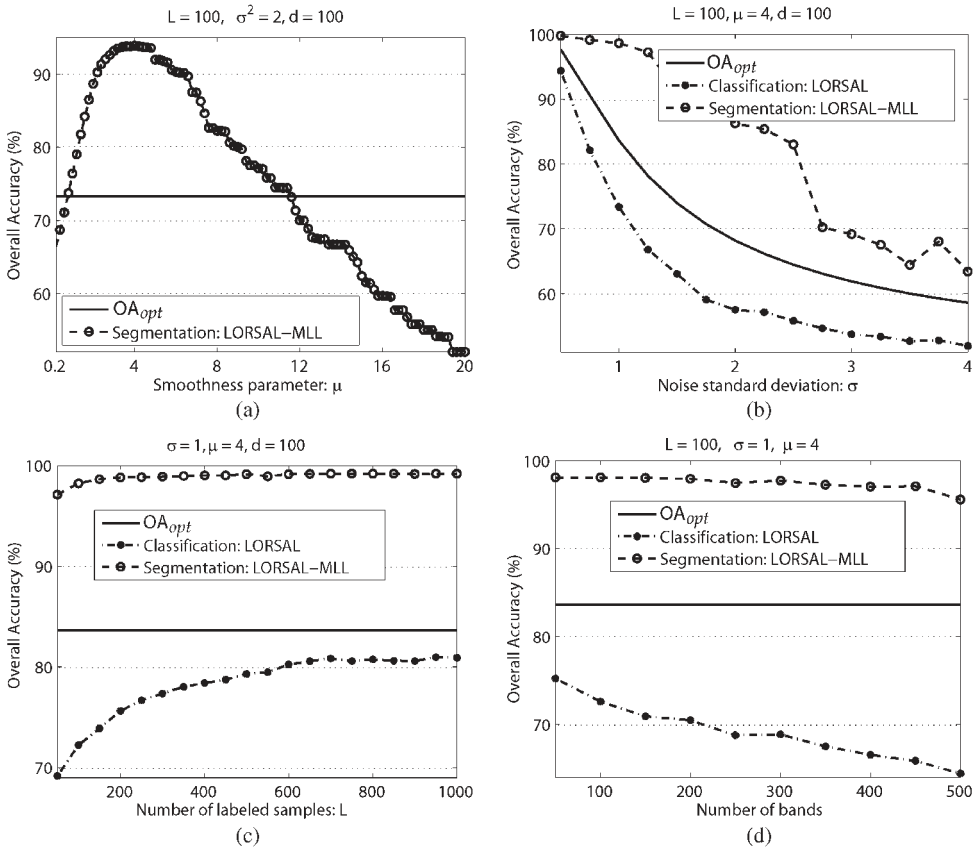


Fig. 5. OA results obtained by the proposed algorithm: (a) As a function of the spatial prior parameter $\mu$. (b) As a function of the noise standard deviation $\sigma$. (c) As a function of the number of labeled samples $L$. (d) As a function of the number of bands $d$.

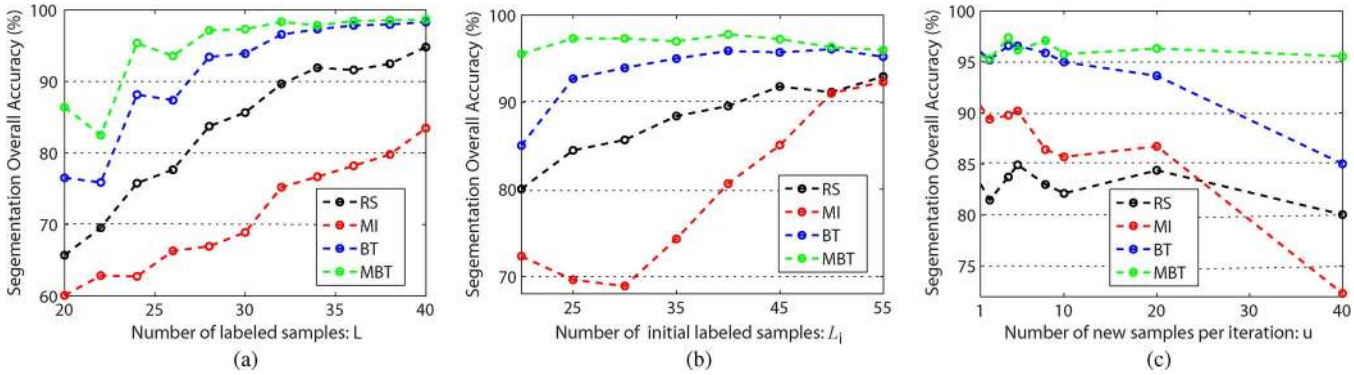


Fig. 6. Segmentation results obtained by using active learning approaches: (a) OA results as a function of $L$ with $L_i = u = L/2$. (b) OA results as a function of $L_i$ with $L = 60$ and $u = L - L_i$. (c) OA results as a function of $u$ with $L = 60$ and $L_i = 20$.

classification and LORSAL-MLL segmentation maps in Fig. 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.

*3) Impact of the Active Learning Approach:* In this section, 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 (18), with $K = 4$, $\sigma = 0.8$, and vectors $\mathbf{m}_{y_i}$ obtained from the USGS library with $d = 224$. Fig. 6 shows the learning results over 100 independent Monte Carlo runs, where we consider three different experiments: 1) OA results as a function of $L$ by using $L_i = u = L/2$; 2) OA results as a function of $L_i$ by using $L = 60$ and $u = L - L_i$; and 3) OA results as a function of $u$ by using $L = 60$ and $L_i = 20$ (five samples per class). Several conclusions can be obtained from the results shown in Fig. 6.

1) First of all, the active learning procedure improves the segmentation results as expected. In general, the MBT strategy achieves the best performance.
2) Second, as already discussed in Section IV, with a small $u$ both MBT and BT lead to very similar results.
3) 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 Fig. 6(c). However, for a large value of $u$, the MI

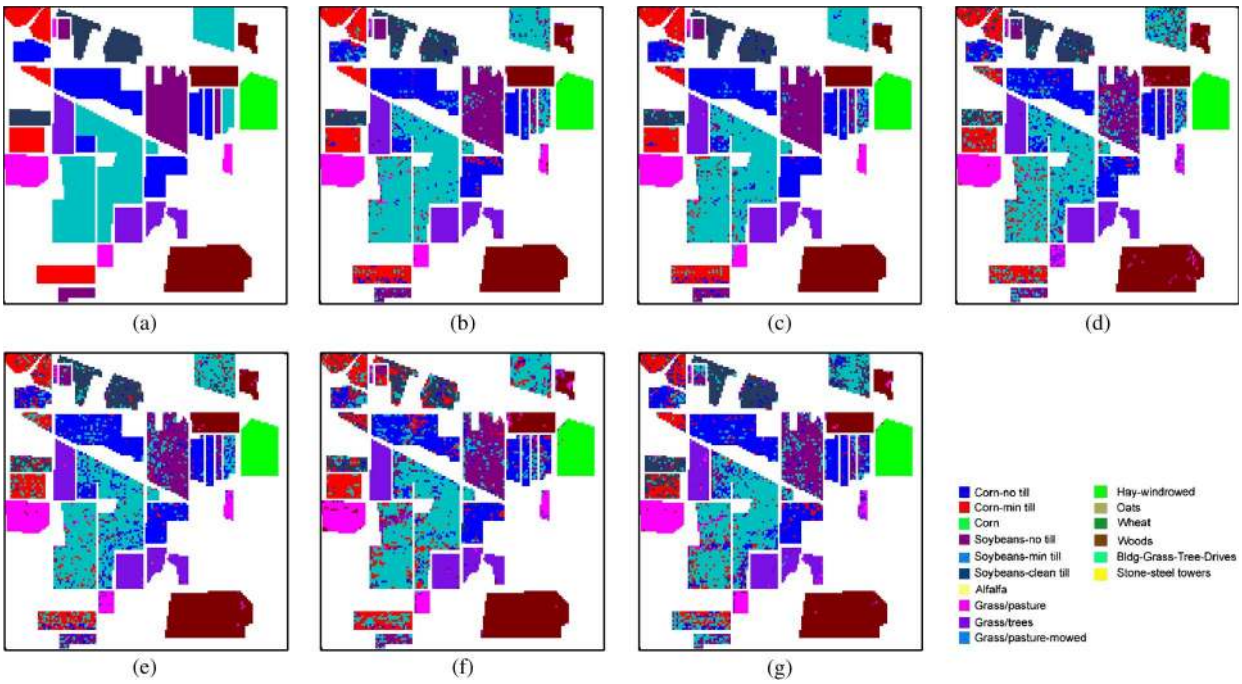


Fig. 7. Classification maps by using $L = 475$, $L_i = 235$, and $u = 60$. (a) Ground truth. (b) LORSAL-AL (RS), OA $= 84.24\%$. (c) LORSAL-AL (MBT), OA $= 86.38\%$. (d) LDA-AL (RS), OA $= 69.35\%$. (e) LDA-AL (MBT), OA $= 70.83\%$. (f) SVM (RS), OA $= 80.43\%$. (g) PCA+SVM (RS), OA $= 76.32\%$.

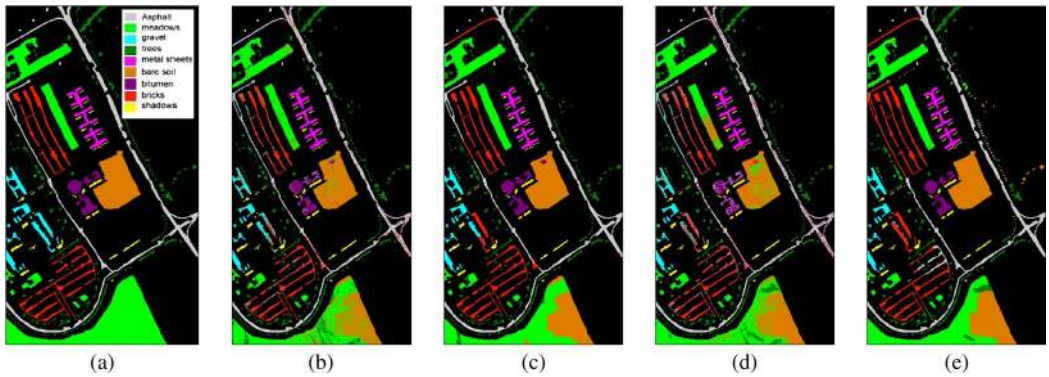


Fig. 8. Classification and segmentation maps obtained for the ROSIS subset #2 by using the whole training set ($L = 3921$). (a) Ground truth. (b) LORSAL, OA $= 80.24\%$. (c) LORSAL-MLL, OA $= 86.72\%$. (d) LDA, OA $= 73.45\%$. (e) LDA-MLL, OA $= 80.67\%$.

TABLE I
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 |

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.

4) Finally, the improvements in performance due to active learning are less relevant as the size of the training set increases, e.g., see Fig. 6(a). This is expected, since the uncertainty in the determination of classifier boundaries decreases as the training set size increases.

### B. 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

TABLE II
PARAMETER SETTINGS IN OUR EXPERIMENTS WITH REAL HYPERSPECTRAL DATA SETS. FOR SUBSET #1, WE ONLY RUN CLASSIFICATION EXPERIMENTS; THEREFORE, NO $\mu$ IS USED

| Dataset | Indian Pines | Subset #1 | Subset #2 | Subset #3 |
|---|---|---|---|---|
| $\lambda$ | 0.001 | 0.001 | 0.001 | 0.001 |
| $\mu$ | 4 | - | 2 | 1 |

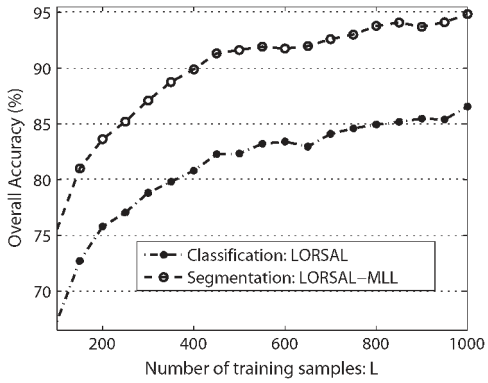


Fig. 9. OA results as a function of the number of labeled samples for the AVIRIS Indian Pines data set.

Indian Pines scene, collected over Northwestern Indiana in June 1992 [49]. The scene is available online[6] and contains 145 $\times$ 145 pixels and 224 spectral bands between 0.4 and 2.5 $\mu$m. A total of 20 spectral bands were removed prior to experiments due to noise and water absorption in those channels. The ground-truth image shown in Fig. 7(a), contains 16 mutually exclusive classes, seven of which were discarded for their small size which resulted in insufficient training samples. The remaining nine 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 $\mu$m. Three different subsets of the full data set are considered in our experiments.

1) Subset #1, with 492 $\times$ 1096 pixels in size, collected over the Pavia city center. The noisy bands were removed, yielding a data set with 102 spectral bands. The ground truth image contains 9 ground-truth classes, 5536 training samples, and 103 539 test samples.
2) Subset #2, with size of 610 $\times$ 340 pixels, centered at the University of Pavia in Italy. The noisy bands were removed, yielding 103 spectral bands. The ground truth image in Fig. 8(a), contains 9 ground-truth classes, 3921 training samples, and 42 776 test samples.
3) Subset #3 includes a dense residential area, with 715 $\times$ 1096 pixels. The ground-truth image contains 9 ground-truth classes, 7456 training samples, and 148 152 test samples.

In our experiments, we compare our proposed approach with LDA [8] and SVMs [11], using feature extraction based on PCA [48] and hyperspectral signal identification by minimum error (HySime) [50]. 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 I 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 II shows the parameter used for each data set. Although these parameter settings may be suboptimal, 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.

*1) 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, Fig. 9 shows the obtained OA results as a function of the number of labeled samples after ten Monte Carlo runs (without active sampling). Here, the training samples are randomly selected from the original training set. From the results shown in Fig. 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 III shows the classification results obtained after ten Monte Carlo runs, along with those provided by SVMs and LDA. From Table III, it can be observed that the proposed MLR-based algorithms obtain good results when compared with 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 shown in Fig. 7 in which the classification maps obtained are displayed along with their associated OA scores.

*2) 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 IV summarizes the results obtained after ten Monte Carlo runs by the considered classification algorithms in comparison with the same standard methods used for reference in the previous section. We emphasize the good classification performance achieved by the proposed LORSAL and LORSAL-AL

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

TABLE III
OA (IN PERCENT) AND $\kappa$ STATISTIC (IN 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 IV
OA (IN PERCENT) $\kappa$ STATISTIC (IN PARENTHESES) FOR 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) |

TABLE V
OA (IN PERCENT) AND $\kappa$ STATISTIC (IN PARENTHESES) OBTAINED FOR 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) |

algorithms. Moreover, Table IV reveals that the MBT sampling procedure further improves the OA results and the $\kappa$ statistic.

In our second experiment, we use subset #2 of the Pavia ROSIS data to evaluate the proposed segmentation algorithm. Table V illustrates the OA results obtained after ten Monte Carlo runs by using the entire training set. Notice the good performances achieved by the proposed LORSAL and LORSAL-MLL algorithms (see Table V), where the segmentation result obtained by the LORSAL-MLL algorithm is comparable with that reported in previous work for an SVM classifier using extended morphological profiles as input features in [1]. 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 with those of 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. Fig. 10 shows the OA results as a function of $L$, with $L_i = 450$ and $u = 20$. From Fig. 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 one 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 the dimensionality of the feature space. Fig. 11 shows 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 Fig. 11, the AL further improves the learning results and, eventually, MI, BT, and MBT converge to very similar OA results. For illustrative purposes, Fig. 8 shows the classification and segmentation maps obtained by the considered algorithm configurations in comparison with other methods) using the ROSIS Pavia University data set.

## VI. CONCLUSION

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 MLR, where the MLR regressors are learned by the LORSAL algorithm. The algorithm adopts an MLL prior to model the spatial information present the class label images. The MAP segmentation is efficiently computed

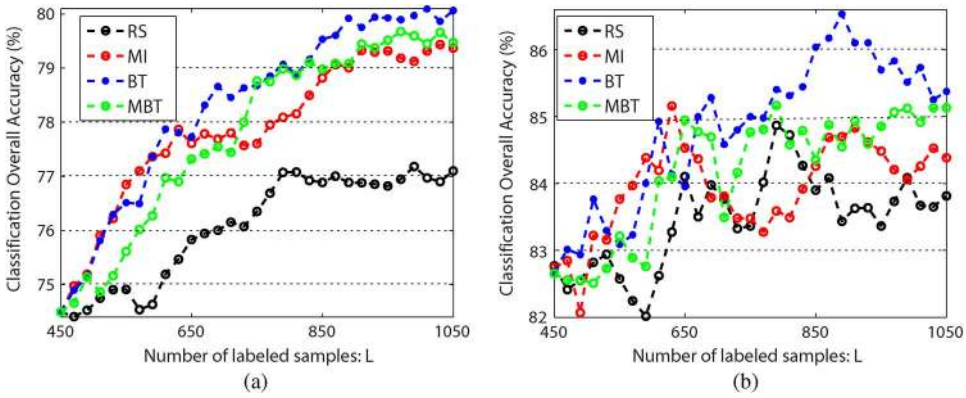


Fig. 10. OA (in percent) results as a function of the number of labeled samples for ROSIS subset #2. (a) LORSAL-AL results. (b) LORSAL-AL-MLL results.

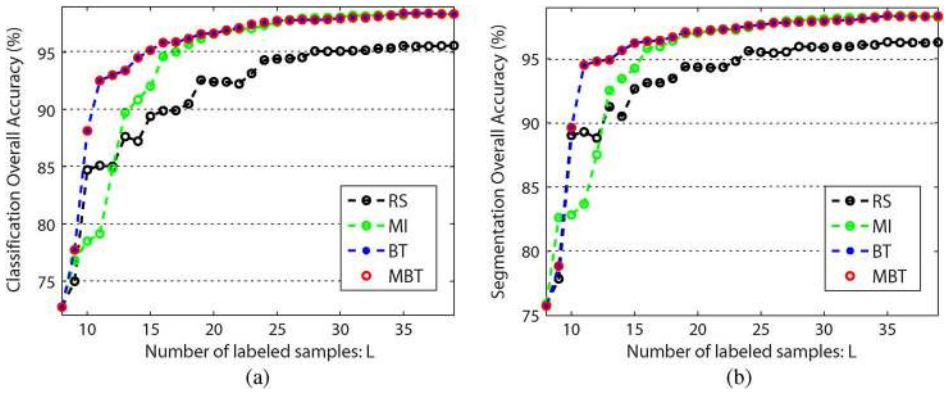


Fig. 11. OA results as a function of the number of labeled samples for ROSIS subset #3. (a) LORSAL-AL results. (b) LORSAL-AL-MLL results.

by the $\alpha$-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 paper 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 MI-based criterion, a BT strategy, and a newly developed method called MBT, are integrated in the developed classification (LORSAL) and segmentation (LORSAL-MLL) methods, resulting in two new methods with active learning, called LORSAL-AL and LORSAL-MLL-AL, respectively. The effectiveness of the proposed algorithms is illustrated in this paper using both simulated and real hyperspectral data sets. 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 toward testing the proposed approach in other different analysis scenarios dominated by the limited availability of training samples.

## APPENDIX

The problem described in (5) is equivalent to

$$(\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\nu}}) = \arg\min_{\boldsymbol{\omega},\boldsymbol{\nu}} -\ell(\boldsymbol{\omega}) + \lambda\|\boldsymbol{\nu}\|_1 \quad \text{subject to : } \boldsymbol{\omega} = \boldsymbol{\nu}. \tag{20}$$

By applying the alternating direction method of multipliers [51] (see also [52] and references therein) to solve the problem in (20), we get the iterative Algorithm 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 (20), for any $\beta \geq 0$ [51].

**Algorithm 4** LORSAL

**Input**: $\boldsymbol{\omega}^{(0)}, \boldsymbol{\nu}^{(0)}, \mathbf{b}^{(0)}, \lambda, \beta$

1: $t := 0$

2: **repeat**

3: $$\widehat{\boldsymbol{\omega}}^{(t+1)} \in \arg\min_{\boldsymbol{\omega}} -\ell(\boldsymbol{\omega}) + \frac{\beta}{2}\left\|\boldsymbol{\omega} - \boldsymbol{\nu}^{(t)} - \mathbf{b}^{(t)}\right\|^2 \tag{21}$$

4: $$\widehat{\boldsymbol{\nu}}^{(t+1)} \in \arg\min_{\boldsymbol{\nu}} \lambda\|\boldsymbol{\nu}\|_1 + \frac{\beta}{2}\left\|\boldsymbol{\omega}^{(t+1)} - \boldsymbol{\nu} - \mathbf{b}^{(t)}\right\|^2 \tag{22}$$

5: $\mathbf{b}^{(t+1)} := \mathbf{b}^{(t)} - \boldsymbol{\omega}^{(t+1)} + \boldsymbol{\nu}^{(t+1)}$

6: $t := t + 1$

7: **until** some stopping criterion is met

It should be noted that the solution of the optimization problem in (21) (line 3 of Algorithm 4) is still a difficult problem because $\ell(\boldsymbol{\omega})$, although strictly convex and smooth, is

nonquadratic and often very large. We tackle this difficulty by replacing $\ell(\boldsymbol{\omega})$ with a quadratic lower bound given by [16]

$$\ell(\boldsymbol{\omega}) \geq \ell\left(\boldsymbol{\omega}^{(t)}\right) + \left(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)}\right)^{\mathrm{T}} \mathbf{g}\left(\boldsymbol{\omega}^{(t)}\right) + \frac{1}{2}\left(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)}\right)^{\mathrm{T}} \mathbf{B}\left(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)}\right) \quad (23)$$

where $\mathbf{B} \equiv -(1/2)[\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/K] \otimes \sum_{i=1}^{L} \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)^{\mathrm{T}}$ (symbol $\mathbf{1}$ denotes a vector column of ones) and $\mathbf{g}(\boldsymbol{\omega}^{(t)})$ is the gradient of $\ell$ at $\boldsymbol{\omega}^{(t)}$. Since the system matrix involved in the optimization of (23), with $\ell(\boldsymbol{\omega})$ replaced with the quadratic bound given in (17), is fixed, its inverse can be precomputed, provided that $\gamma$—the dimension of $\mathbf{h}(\mathbf{x}_i)$—is below, for example, a few thousands. Under mild conditions, the convergence of Algorithm 4 with the aforementioned modification still holds [51], [52].

On the other hand, the solution of the optimization problem in (22) (line 4 of Algorithm 4) is simply the soft-threshold rule [53] 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 componentwise. As a final note, we reiterate that the complexity of each iteration of the LORSAL algorithm is $O(\gamma^2 K)$, which is faster than $O((\gamma K)^3)$ for the SMLR algorithm [18], and $O(\gamma^3 K)$ for the FSMLR algorithm [19].

## ACKNOWLEDGMENT

The authors would like to thank Prof. D. Landgrebe for making the AVIRIS Indian Pines hyperspectral data set available to the community, Prof. P. Gamba for providing the ROSIS data over Pavia, Italy, along with the training and test set, Prof. V. Kolmogorov for making the max-flow/min-cut C++ code available to the community, and the three anonymous reviewers for their highly constructive and outstanding remarks, which greatly helped us to improve the technical quality and presentation of our manuscript significantly.

## REFERENCES

[1] 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.

[2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[3] M. Chi and L. Bruzzone, "Semi-supervised classification of hyperspectral images by SVMs optimized in the primal," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1870–1880, Jun. 2007.

[4] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[5] 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.

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

[7] 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, pp. 841–848.

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

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

[10] 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.

[11] 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.

[12] M. Chi and L. Bruzzone, "An ensemble-driven k-NN approach to ill-posed classification problems," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 301–307, Mar. 2006.

[13] 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.

[14] M. Chi and L. Bruzzone, "A semi-labeled-sample driven bagging technique for ill-posed classification problems," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 1, pp. 69–73, Jan. 2005.

[15] M. Chi, R. Feng, and L. Bruzzone, "Classification of hyperspectral remote sensing data with primal support vector machines," *Adv. Space Res.*, vol. 41, no. 11, pp. 1793–1799, 2008.

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

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

[18] 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.

[19] J. Borges, J. Bioucas-Dias, and A. Marçal, "Fast sparse multinomial regression applied to hyperspectral data," in *Proc. ICIAR*, 2006, pp. 700–709.

[20] J. Li, J. Bioucas-Dias, and A. Plaza, "Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, Nov. 2010.

[21] S. Prasad, L. Bruce, and H. Kalluri, "A robust multi-classifier decision fusion framework for hyperspectral, multi-temporal classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2008, pp. 3048–3051.

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

[23] 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.

[24] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, Apr. 2008.

[25] 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.

[26] W. Di and M. M. Crawford, "Locally consistend graph regularization based active learning for hyperspectral image classification," in *Proc. 2nd IEEE Workshop Hyperspectral Image Signal Process.: Evolution Remote Sens.*, 2010, pp. 1–4.

[27] G. Jun and J. Ghosh, "An efficient active learning algorithm with knowledge transfer for hyperspectral data analysis," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2008, pp. I52–I55.

[28] L. Copa, D. Tuia, M. Volpi, and M. Kaneski, "Unbiased query-by-bagging active learning for VHR image classification," in *Proc. SPIE Eur. Remote Sens.*, 2010, p. 783 00K.

[29] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. Benediktsson, "SVM and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.

[30] J. Bioucas-Dias and M. Figueiredo, "Logistic regression via variable splitting and augmented Lagrangian tools," Instituto Superior Técnico, TULisbon, Lisbon, Portugal, 2009, Tech. Rep.

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

[32] 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.

[33] T. Luo, K. Kramer, D. B. Goldgof, S. Samson, A. Remsen, T. Hopkins, and D. Cohn, "Active learning to recognize multiple types of plankton," *J. Mach. Learn. Res.*, vol. 6, pp. 589–613, 2005.

[34] Y. Boykov, O. Veksler, and R. Zabih, "Efficient approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1222–1239, Nov. 2001.

[35] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statistician*, vol. 58, no. 1, pp. 30–37, Feb. 2004.

[36] 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 *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2009, pp. III-817–III-820.
[37] 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.
[38] S. Z. Li, *Markov Random Field Modeling in Computer Vision*. London, U.K.: Springer-Verlag, 1995.
[39] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. R. Stat. Soc. B*, vol. 36, no. 2, pp. 192–236, 1974.
[40] 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.
[41] J. Yedidia, W. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," in *Proc. Int. Joint Conf. Artif. Intell.*, 2001, pp. 239–269.
[42] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.
[43] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568–1583, Oct. 2006.
[44] S. Bagon, Matlab Wrapper for Graph Cut, Dec. 2006. [Online]. Available: http://www.wisdom.weizmann.ac.il/~bagon
[45] M. E. Tipping and A. Smola, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.
[46] J. Li, J. Bioucas-Dias, and A. Plaza, "Semi-supervised hyperspectral classification using active label selection," in *Proc. SPIE Eur. Remote Sens.*, 2009, vol. 7477, pp. 74770F1–74770F8.
[47] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 1995, pp. 1137–1143.
[48] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification,* 2nd ed. Hoboken, NJ: Wiley-Interscience, 2000.
[49] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.
[50] J. Bioucas-Dias and J. Nascimento, "Hyperspectral subspace identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 8, pp. 2435–2445, Aug. 2008.
[51] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, no. 3, pp. 293–318, Jun. 1992.
[52] M. V. Afonso, J. Bioucas-Dias, and M. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–2356, Sep. 2010.
[53] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

**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.

From 2007 to 2010, she was a Marie Curie Research Fellow with the Departamento de Engenharia Electrotécnica e de Computadores, Instituto de Telecomunicações, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal, in the framework of the European Doctorate for Signal Processing (SIGNAL) under the joint supervision of Prof. José M. Bioucas-Dias and Prof. Antonio Plaza. Currently, she is 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. Her research interests include hyperspectral image classification and segmentation, spectral unmixing, signal processing, and remote sensing. She has been a Reviewer of several journals, including *Optical Engineering* and *Inverse Problems and Imaging*.

Ms. Li is a Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

**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 Instituto Superior Técnico (IST), the Engineering School of the Universidade Técnica de Lisboa, Lisboa, Portugal, in 1985, 1991, 1995, and 2007, respectively.

Since 1995, he has been with the Department of Electrical and Computer Engineering, IST, where he is also a Senior Researcher with the Communication Theory and Pattern Recognition Group, Instituto de Telecomunicações, a private not-for-profit research institution. 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)." His research interests include signal and image processing, pattern recognition, optimization, and remote sensing.

Dr. Bioucas-Dias is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He 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 the IEEE Conference on Computer Vision and Pattern Recognition, the International Conference on Pattern Recognition, the International Conference on Image Analysis and Recognition, the IEEE International Geoscience and Remote Sensing Symposium, the International Conference on Image Processing, SPIE, the International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition, the International Symposium on Visual Computing, 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.

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. He is currently an Associate Professor with the Department of Technology of Computers and Communications, University of Extremadura, Caceres, Spain, where he is the Head of the Hyperspectral Computing Laboratory (HyperComp). He was the Coordinator of the Hyperspectral Imaging Network (Hyper-I-Net), 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. He is the author or coauthor of more than 260 publications on remotely sensed hyperspectral imaging, including more than 50 Journal Citation Report papers, book chapters, and conference proceeding papers. 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 has coedited a book on high-performance computing in remote sensing and guest edited four special issues on remotely sensed hyperspectral imaging for different journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (for which he serves as Associate Editor on hyperspectral image analysis and signal processing since 2007), the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATION AND REMOTE SENSING, the *International Journal of High Performance Computing Applications*, and the *Journal of Real-Time Image Processing*. He has served as a reviewer for more than 240 manuscripts submitted to more than 40 different journals, including more than 120 manuscripts reviewed for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He is a recipient of the recognition of Best Reviewers of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2009 and a recipient of the recognition of Best Reviewers of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2010. He is currently serving as Director of Education activities for the IEEE Geoscience and Remote Sensing Society.