Hyperspectral segmentation with active learning

Jun Li, José M. Bioucas-Dias, and Antonio Plaza

Abstract

This paper introduces a new supervised Bayesian approach to hyperspectral image segmentation, with two main steps: (a) learning, for each class label, the posterior probability distributions, based on a multinomial logistic regression model; (b) segmenting the hyperspectral image, based on the posterior probability distribution learnt in step (a) and on a multi-level logistic prior encoding the spatial information. The multinomial logistic regressors are learnt by using the recently introduced LORSAL (logistic regression via splitting and augmented Lagrangian) algorithm. The maximum a posterior segmentation is efficiently computed by the α-Expansion min-cut based integer optimization algorithm. Aiming at reducing the costs of acquiring large training sets, active learning is performed using a mutual information based criterion. State-of-the-art performance of the proposed approach is illustrated with simulated and real hyperspectral data sets in a number of experimental comparisons with recently introduced hyperspectral classification methods.

Index Terms

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

I. Introduction

In recent years, with the development of remote sensing sensors, hyperspectral images are widely available. The special characteristics of hyperspectral data sets bring difficult processing problems. Obstacles, *e.g.*, Hughes phenomenon [1], come out as the data

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

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

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 have been applied to extract relevant information from hyperspectral data sets [2–4]. However, although many progresses have been made, the difficulty in learning high dimensional densities from a limited number of training samples, *i.e.*, ill-posed problems, is still an active area of research.

The discriminative approach, which learns the class distributions in high dimensional spaces by inferring the boundaries between classes in the feature space [5–7], tackles effectively the above mentioned difficulties. Support vector machines (SVMs) [8] are among the state-of-the-art discriminative techniques in ill-posed classification problems. Due to their ability to deal with large input spaces efficiently and to produce sparse solutions, SVMs have been successfully used for hyperspectral supervised and semi-supervised classification with limited training samples [2, 9–15]. 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. Effective sparse MLR (SMLR) methods are available [17]. These ideas have been applied to hyperspectral classification [4, 18] yielding state-of-the-art performance.

In order to improve the classification accuracies obtained by SVMs and MLR-based techniques, a recent trend is to integrate spatial contextual information with spectral information in hyperspectral data interpretation [4, 10, 13, 19]. These methods exploit, in a way or another, the continuity, in probability sense, of neighboring labels: it is very likely that, in an hyperspectral image, two neighboring pixels have the same label.

More recently, in order to reduce the size of the training sets, active learning has been widely studied in the literature [20–25]. These studies based on maximum entropy [23],

on an extension of SVM margin sampling [23], on a hierachical classification framework [22, 25], on a local proximity based data regularization framework [24], etc., give evidence that the active learning procedure leads systematically to noticeable improvements in the classification results.

In this paper, we introduce a new supervised Bayesian segmentation approach which exploits both the spectral and spatial information in the interpretation of hyperspectral data. The algorithm implements two main steps: (a) learning step, which uses the multinomial logistic regression via variable splitting and augmented (LORSAL) [26] algorithm to infer the class distributions; (b) segmentation step, which infers the labels from a posterior distribution built on the learned class distributions and on a multi-level logistic (MLL) prior [27], where the maximum a posterior (MAP) segmentation is computed via a min-cut based integer optimization algorithm. Furthermore, aiming at a reduction in the size of the training set, we implement an active learning technique based on the mutual information (MI) between the MLR regressors and the class labels [20, 21].

The remainder of the paper is organized as follows. Section II formulates the problem. Section III describes the proposed approach. Section IV reports segmentation results based on simulated and real hyperspectral datasets in several ill-posed scenarios; comparisons with state-of-the-art competitors are also included. Finally, section V concludes with a few remarks.

II. problem formulation

Let $\mathcal{S} \equiv \{1, \ldots, n\}$ denote a set of integers indexing the pixels of a n-size image, $\mathcal{L} \equiv \{1, \ldots, K\}$ be a set of K labels, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ be an image of d-dimensional feature vectors, $\mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{L}^n$ be an image of labels, and $\mathcal{D}_L \equiv \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_L, y_L)\} \in \left(\mathbb{R}^d \times \mathcal{L}\right)^L$ be a training set where L denotes the number of labeled samples.

With the above definitions in place, the goal of classification is to assign a label $y_i \in \mathcal{L}$ to each pixel $i \in \mathcal{S}$, based on the vector $\mathbf{x}_i$, resulting in an image of class labels $\mathbf{y}$. We call this assignment a *labeling*. The goal of segmentation is, based on the observed image $\mathbf{x}$, to compute 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 property, for example to 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 $\mathcal{S}$. On the other way around, 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. There is, therefore, a one-to-one relation between labelings and segmentations. Nevertheless, in this paper, we use the term classification when there is no spatial information and segmentation when the spatial prior is being considered.

Inference in a Bayesian framework is often carried out by maximizing the posterior distribution[1]

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

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

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

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

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

[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.

In the present approach, the densities $p(y_i|\mathbf{x}_i)$ are modeled as MLRs [16], whose regressors are learnt via the LORSAL algorithm [26]. As prior $p(\mathbf{y})$ on the labelings, $\mathbf{y}$, we adopt an MLL Markov random field (MRF) [27], which encourages neighboring pixels to have the same label. The MAP labeling/segmentation $\widehat{\mathbf{y}}$ is computed via the α-Expansion algorithm [28], a min-cut based tool to efficiently solve a class of integer optimization problems of which (2) is an example.

III. Proposed approach

The MLR model is formally given by [16],

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

where $\mathbf{h}(\mathbf{x}) \equiv [h_1(\mathbf{x}), ..., h_l(\mathbf{x})]^T$ is a vector of l fixed functions of the input, often termed as features, $\boldsymbol{\omega} \equiv [\boldsymbol{\omega}^{(1)^T}, ..., \boldsymbol{\omega}^{(K)^T}]^T$ denotes the logistic regressors. Since the density (3) does not depend on translations on the regressors $\boldsymbol{\omega}^{(K)}$, we take $\boldsymbol{\omega}^{(K)} = \mathbf{0}$ and remove it from $\boldsymbol{\omega}$, *i.e.*, $\boldsymbol{\omega} \equiv [\boldsymbol{\omega}^{(1)^T}, ..., \boldsymbol{\omega}^{(K-1)^T}]^T$.

It should be noted that function $\mathbf{h}$ may be linear, *i.e.*, $\mathbf{h}(\mathbf{x}_i) = [1, x_{i,1}, ..., x_{i,d}]^T$, where $x_{i,j}$ is the j-th component of $\mathbf{x}_i$, or nonlinear. Kernels, *i.e.*, $\mathbf{h}(\mathbf{x}_i) = [1, K_{\mathbf{x},\mathbf{x}_1}, ..., K_{\mathbf{x},\mathbf{x}_l}]^T$, where $K_{\mathbf{x}_i,\mathbf{x}_j} \equiv K(\mathbf{x}_i, \mathbf{x}_j)$ and $K(\cdot,\cdot)$ is some symmetric kernel function, 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) $K(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2/(2\rho^2))$ kernel, which is widely used in hyperspectral image classification [11]. Defining γ as the dimension of $\mathbf{h}(\mathbf{x})$, then we have $\gamma = d + 1$ for the above linear case and $\gamma = L + 1$ for the RBF kernel (recall that L is the number of samples in the training set $\mathcal{D}_L$).

A. LORSAL

In the present problem, learning the class densities amounts to estimating the logistic regressors $\boldsymbol{\omega}$. Following the SMLR algorithm [17], the estimation of $\boldsymbol{\omega}$ amounts to computing the MAP estimate

$$\widehat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \quad \ell(\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}), \tag{4}$$

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

$$\ell(\boldsymbol{\omega}) \equiv \log \prod_{i=1}^{L} p(y_i|\mathbf{x}_i, \boldsymbol{\omega}), \tag{5}$$

and $p(\boldsymbol{\omega}) \propto \exp(-\lambda\|\boldsymbol{\omega}\|_1)$ is a sparsity promoting prior ($\|\boldsymbol{\omega}\|_1$ denotes the l_1 norm of $\boldsymbol{\omega}$) where λ is a regularization parameter. The prior $p(\boldsymbol{\omega})$ forces many components of $\boldsymbol{\omega}$ to be zero, thus controlling the classifier complexity and. consequently, enhancing its generalization capacity.

Optimization (4), although convex, is difficult to solve because the term $\ell(\boldsymbol{\omega})$ is non-quadratic and the term $\log p(\boldsymbol{\omega})$ is non-smooth. The majorization-minimization framework [29] has recently been used in [17, 21, 30, 31] to convert (4) into a sequence of quadratic problems. The computational cost of the SMLR algorithm [17] involved in solving each quadratic problem is $O((\gamma K)^3)$, which is prohibitive when dealing with datasets with large number of features, or with large number of classes, or both. FSMLR [18], a fast version of SMLR, implements a block Gauss-Seidual iterative procedure to calculate $\boldsymbol{\omega}$ which is $O(K^2)$ faster than the original SMLR algorithm [17]. Thus, the FSMLR algorithm extends the SMLR capability to handle datasets with large number of classes. However, with an overall complexity of $O(\gamma^3 K)$, for hyperspectral data with large number of features, FSMLR complexity is still unbearable in many cases.

In this paper, we use the recently introduced LORSAL algorith, [26] to learn the MLR regressors. LORSAL replaces a difficult non-smooth convex problem with a sequence of

quadratic plus diagonal l_2-l_1 problems very easy to solve. In practice, the total cost of the LORSAL algorithm is $O(\gamma^2 K)$ per iteration, which is in contrast with the $O((\gamma K)^3)$ and $O(\gamma^3 K)$ complexities of, respectively, SMLR and FSMLR algorithms. As a result, the reduction of computational complexity is of the order of γK^2 and γ, respectively. The LORSAL algorithm is briefly reviewed in appendix.

B. Active learning

In order to reduce the acquisition of large amount of labeled samples, active learning is performed in this paper. The basic idea of active learning is that of iteratively enlarging the training set by requesting an expert to, in each iteration, label samples from the unlabeled set, $\{\mathbf{x}_i, i \in \mathcal{S}_U\}$, where $\mathcal{S}_U$ is the set of unlabeled feature vectors, *i.e.*, spectral vectors in the current application. The relevant question is, of course, what vectors should be chosen. In this paper, we use an MI based criterion [20, 21] that maximizes the MI between the MLR regressors and the class labels. The proposed approach uses a Laplace approximation of the posterior $p(\boldsymbol{\omega}|\mathcal{D}_L) \simeq \mathcal{N}(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}}, \mathbf{H}^{-1})$, where $\mathbf{H}$ is the posterior precision matrix, *i.e.*, the Hessian of minus the log-posterior $\mathbf{H} \equiv \nabla^2(-\log p(\widehat{\boldsymbol{\omega}}|\mathcal{D}_L))$. Let $\mathbf{x}_i$ be an unlabeled sample and y_i be its label. Assume that the MAP estimate $\widehat{\boldsymbol{\omega}}$ remains unchanged after including y_i. This assumption is clearly not true at the beginning of the active learning procedure. Nevertheless, it was empirically observed that it leads to very good approximations [21]. Under this assumption, the posterior precision matrix changes to

$$\mathbf{H}' = \mathbf{H} + \left(\mathrm{diag}(\mathbf{p}_i(\widehat{\boldsymbol{\omega}})) - \mathbf{p}_i(\widehat{\boldsymbol{\omega}})\mathbf{p}_i(\widehat{\boldsymbol{\omega}})^T\right) \otimes \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)^T, \tag{6}$$

where $\mathbf{p}_i(\widehat{\boldsymbol{\omega}}) \equiv [p_{i,1}, \ldots, p_{i,K}]^T$, $p_{i,k} \equiv p(y_i = k|\mathbf{x}_i, \widehat{\boldsymbol{\omega}})$, for $k = 1, \ldots, K$, and $\otimes$ is the Kronecker product. As shown in [20], the MI $I(\boldsymbol{\omega}; y_i)$ between the MLR regressors and the class label

y_i is given by

$$\begin{aligned} I(\boldsymbol{\omega}; y_i) &= (1/2)\log(|\mathbf{H}'|/\mathbf{H}) \\ &= (1/2)\log\left(1 + \prod_{k=1}^{K} p_{i,k}\, \mathbf{x}_i^T \mathbf{H}^{-1} \mathbf{x}_i\right). \end{aligned} \tag{7}$$

The function (7) is maximized for $p_{i,k} \approx 1/K$, *i.e.*, for samples near the classifier boundaries, corresponding to probability vectors $\mathbf{p}_i$ with maximum entropy. Algorithm 1 shows the pseudo-code of an iterative active learning scheme, where u is the number of new samples considered per iteration, $\beta \geq 0$ is the augmented Lagrangian LORSAL parameter (see appendix). Although the expression (7) has been derived for the selection of just one sample, we consider the inclusion of a number $u \geq 1$ in each iteration of Algorithm 1. This is, of course, a sub-optimal procedure. Nevertheless, we have found out experimentally that it still leads to very good results with the advantage of being u times faster.

Algorithm 1 LORSAL using active learning (LORSAL-AL)

Input: $\widehat{\boldsymbol{\omega}}$, $\mathcal{D}_L$, u, λ, β

1: **repeat**
2: $\quad\widehat{\mathbf{P}} := [\mathbf{p}_i(\widehat{\boldsymbol{\omega}})]$ for $i \in \mathcal{S}$
3: $\quad H_i := H[\mathbf{p}_i(\widehat{\boldsymbol{\omega}}]$ for $i \in \mathcal{S} - \{1, \ldots, L\}$ ($*$ compute entropy of $\mathbf{p}_i$ $*$)
4: $\quad i_1, i_2, \ldots, i_{n-L} :=$ permutation of $\mathcal{S} - \{1, \ldots, L\}$ such that H_{i_k} is decreasing
5: $\quad \mathcal{D}_L := \mathcal{D}_L \bigcup \{(\mathbf{x}_{i_i}, y_{i_1}), \ldots, (\mathbf{x}_{i_u}, y_{i_u})\}$
6: $\quad \widehat{\boldsymbol{\omega}} :=$ LORSAL$(\mathcal{D}_L, \lambda, \beta)$
7: **until** some stopping criterion is meet

C. The Multi-Level Logistic spatial prior

In order to encourage piecewise smooth segments and promote solutions in which adjacent pixels are likely to belong to the same class, we include the contextual spatial information 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 [32] and has been widely used in image segmentation problems (see *e.g.*, [4, 30, 31, 33]).

According to the Hammersly-Clifford theorem [34], the density associated with an MRF

is a Gibbs's distribution [32]. 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 μ_c is a non-negative constant.

The potential function in (9) encourages neighbors to have the same class label. The considered MLL prior offers great flexibility by varying the set of cliques and the parameters υ_{y_i} and μ_c. For example, the model generates texture-like regions if μ_c depends on c and blob-like regions otherwise [27]. By taking $e^{\upsilon_{y_i}} \propto p(y_i)$, denoting $\mu_c = \frac{1}{2}\mu > 0$, and assuming that the cliques consistes either of a single pixel, *i.e.*, $c = \{i\}$, or of a pair of neighboring pixels, *i.e.*, $c = \{i, j\}$ where i and j are neighbors, then the equation (8) can be rewritten as

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

where $\delta(y)$ is the unit impulse function[3]. This choice gives no preference to any direction. A straightforward computation of $p(y_i)$, *i.e.*, the marginal of $p(\mathbf{y})$ with respect to y_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

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

around. In this way, the MLL prior promotes piecewise smooth segmentations, where μ controls the degree of smoothness.

D. 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 (2), the MAP segmentation is finally given by

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

where $p(y_i|\widehat{\boldsymbol{\omega}}) \equiv p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$, computed at $\widehat{\boldsymbol{\omega}}$. Minimization of expression (11) is a combinatorial optimization problem, involving unary and pairwise interaction terms, which is difficult to compute. Energy minimization algorithms like Graph cuts [28,35,36], Loopy Belief Propagation [37, 38], and tree-reweighed message passing [39] developed recently are efficient tools to tackle this class of optimization problems. In this work, we use the α-Expansion algorithm [28] to solve our integer optimization problem [40], which yields very good approximations to the MAP segmentation and is efficient from the computational point of view, being the practical computational complexity $O(n)$. The pseudo-codes for the proposed supervised segmentation algorithms with discriminative class learning and MLL prior are shown in Algorithm 2, without active learning, and in Algorithm 3, with active learning.

Algorithm 2 Supervised segmentation algorithm (LORSAL-MLL)

Input: $\mathcal{D}_L$, λ, β

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

Algorithm 3 Supervised segmentation algorithm using active learning (LORSAL-AL-MLL)

Input: $\widehat{\boldsymbol{\omega}}$, $\mathcal{D}_L$, u, λ, β

repeat
  $\widehat{\mathbf{P}} := [\mathbf{p}_i(\widehat{\boldsymbol{\omega}})]$ for $i \in \mathcal{S}$
  $H_i := H[\mathbf{p}_i(\widehat{\boldsymbol{\omega}}]$ for $i \in \mathcal{S} - \{1, \ldots, L\}$ ($*$ compute entropy of $\mathbf{p}_i$ $*$)
  $i_1, i_2, \ldots, i_{n-L} :=$ permutation of $\mathcal{S} - \{1, \ldots, L\}$ such that H_{i_k} is decreasing
  $\mathcal{D}_L := \mathcal{D}_L \bigcup \{(\mathbf{x}_{i_i}, y_{i_1}), \ldots, (\mathbf{x}_{i_u}, y_{i_u})\}$
  $\widehat{\boldsymbol{\omega}} :=$ LORSAL$(\mathcal{D}_L, \lambda, \beta)$
  $\widehat{\mathbf{y}} := \alpha$-Expansion$(\widehat{\mathbf{P}}, \mu)$
until some stopping criterion is meet

E. Overall complexity

The overall complexity is dominated by the supervised learning of the MLR regressors through the LORSAL algorithm, shown in appendix (Algorithm 4), which has a complexity of $O(\gamma^2 K)$, and by the α-Expansion algorithm used to determine the MAP segmentation, which has a practical complexity of $O(n)$. In conclusion, if $\gamma^2 K \gg n$ (*e.g*, $\mathbf{h}(\mathbf{x})$ are kernels and the number of classes is large), then the algorithm is dominated by the computation of the MLR regressors, whereas if $\gamma^2 K \ll n$, the algorithm complexity is dominated by the $\alpha-$Expansion algorithm.

IV. 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 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.

This section is organized as follows. Section A reports experiments with simulated data and it has the following structure. Subsection A.1 evaluates the LORSAL algorithm. Subsection A.2 evaluates the impact of the spatial prior. Finally, subsection A.3 evaluate the

impact of the active learning approach. Section B evaluates the performance of the proposed algorithm using four real hyperspectral scenes collected by AVIRIS over agricultural fields located at Indian Pines, Indiana [41], and by the ROSIS sensor, operated by DLR (German Aerospace Agency) over the town of Pavia, Italy. In this section, the results are compared with state-of-the-art algorithms for hyperspectral image processing presented in [10], as recent advances in hyperspectral techniques have been introduced in this paper.

It should be noted that, in all experiments, except in Section A.1, which uses the linear model to evaluate the LORSAL algorithm, we use RBF Kernels $K(\mathbf{x},\mathbf{z}) \equiv \exp(-\|\mathbf{x}-\mathbf{z}\|^2/(2\rho^2))$ applied on the normalized data[4], where all spectral bands have been used. The scale parameter is set to a fixed value with $\rho = 0.6$, as this setting leads to very good estimate. Furthermore, we have noticed that there is no noticeable improvements for small variations of ρ. The regularization parameter and the LORSAL augmented Lagrangian parameter were set to $\lambda = 10^{-3}$ and $\beta = 10^{-4}$, respectively. Although this setting is not optimal, we have observed that it yields very good results in all experiments.

Let $\mathcal{D}_{L_i}$, $\mathcal{D}_L$, and $\mathcal{D}_U \equiv \mathcal{D}_{L-L_i}$ denote the initial labeled set, the final labeled set, and the actively selected samples (and labels), respectively, $\mathcal{D}_u$ be the new samples actively selected per iteration, where L_i, L, U, u are the number of samples in the respective set. In the experiments, $\mathcal{D}_{L_i}$ and $\mathcal{D}_U$ are, respectively, randomly selected and actively selected from the complete training set. In all cases, the reported values of the overall accuracy (OA) are obtained by averaging the results of 10 Monte Carlo runs, with respect to the initial labeled samples $\mathcal{D}_{L_i}$.

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

A. Experiments with simulated data

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

$$\mathbf{x}_i = \mathbf{m}_{y_i} + \mathbf{n}_i, \quad i \in \mathcal{S}, \;\; y_i \in \mathcal{L} \tag{12}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the spectral vector observed at pixel i, $\mathbf{m}_{y_i}$, for $y_i \in \mathcal{L}$, 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 the subsections of A.1 and A.2, we do not consider the active learning procedure, *i.e.*, $L = L_i$, because our focus is on the competitiveness of the LORSAL algorithm and on the role of the spatial prior, independently of the active learning mechanism. The training set $\mathcal{D}_L$ is randomly selected from the ground truth image. The remaining samples are considered as the validation set.

A.1 On the effectiveness of the LORSAL algorithm

In this experiment, we illustrate the effectiveness of the LORSAL algorithm. We generate the spectral vector according to the above model (12), where spectral vectors $\mathbf{m}_i$, for $i = 1, \ldots, K$, were selected (randomly) from the U.S. Geological Survey (USGS) digital spectral library[5] with $d = 224$, $K = 10$, $L = 1000$, and $\sigma = 1$. Fig.1 plots the evaluation of the log-posterior $\ell(\boldsymbol{\omega}) - \lambda\|\boldsymbol{\omega}\|_1$ as a function of time for LORSAL, FSMLR, and SMLR algorithms. LORSAL is, by far, the fastest algorithm. For a similar log-posterior, LORSAL algorithm takes about 2 seconds while FSMLR algorithm takes around 48 seconds, and the SMLR algorithm takes about 880 seconds.

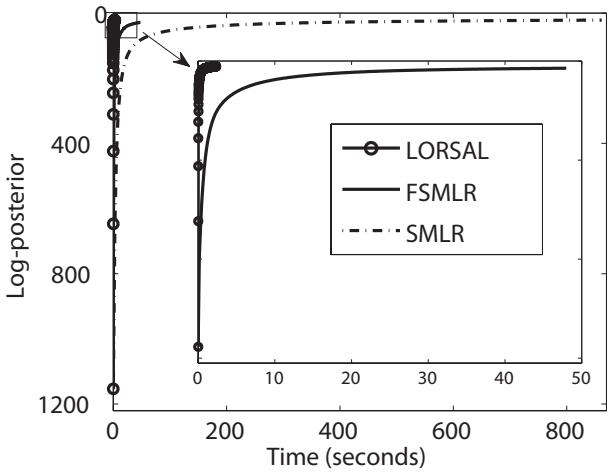
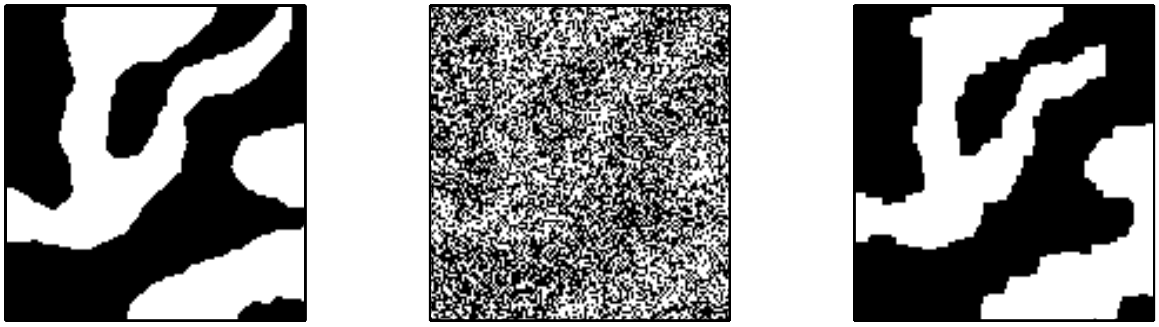


Fig. 1. Evaluation of the log-posterior (4) as a function of the time for LORSAL, FSMLR, and SMLR algorithms.



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

Fig. 2. Classification and segmentation results obtained with the proposed algorithm. The simulated data set was generedted according to (12) with $d = 500$ and $\sigma = 1.5$, $\mu = 2$. (a) simulated binary map; (b) classification map produced by the LORSAL algorithm using $L = 100$ labeled samples without active learning (OA=60.13%, with $\mathrm{OA}_{opt} = 71.91\%$, see text); (c) as in (b) but using the MLL spatial prior (OA=92.48%).

A.2 Impact of the spatial prior

In this experiment, we analyze the impact of the spatial prior on the segmentation performance on a binary problem, *i.e.*, $K = 2$. 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 (12) 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 $\mathrm{OA}_{opt} \equiv 100\,(1 - \mathrm{P}_e)\%$ and corresponding

[5] USGS is available online: http://speclab.cr.usgs.govand

to the minimal probability of error [42] is

$$\mathrm{P}_e = \frac{1}{2}\,\mathrm{erfc}\left(\frac{1+\lambda_0}{\sqrt{2}\,\sigma}\right) p_0 + \frac{1}{2}\,\mathrm{erfc}\left(\frac{1-\lambda_0}{\sqrt{2}\,\sigma}\right) p_1, \tag{13}$$

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

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

Fig.3 plots the OA results obtained with the proposed algorithms. The following conclusions may be drawn:

1. the best results are obtained, as expected, by the proposed segmentation algorithm, which are higher, in all cases, than the classification results (and also of OA_{opt}). This confirms our introspection that the inclusion of a spatial prior significantly improve the classification results provided by using only spectral information, even for very noisy scenarios [see plot (a)].
2. the classification OA approaches the optimal values OA_{opt} as the number of labeled samples increases [see plot (b)]. However, the number of labeled samples needs to be relatively high in order to obtain classification accuracies which are close to optimal.
3. by using the same size of training samples, the classification accuracy decreases as the number of bands increases [see plot (c)]. This is expectable according to the Hugues phenomenon. On the contrary, by including the spatial prior, our supervised segmentation algorithm, performs very well even with small training sets and large number of bands.
4. the segmentation results are almost insensitive to the smooth parameter μ for $\mu \geq 2$ [see plot (d)].

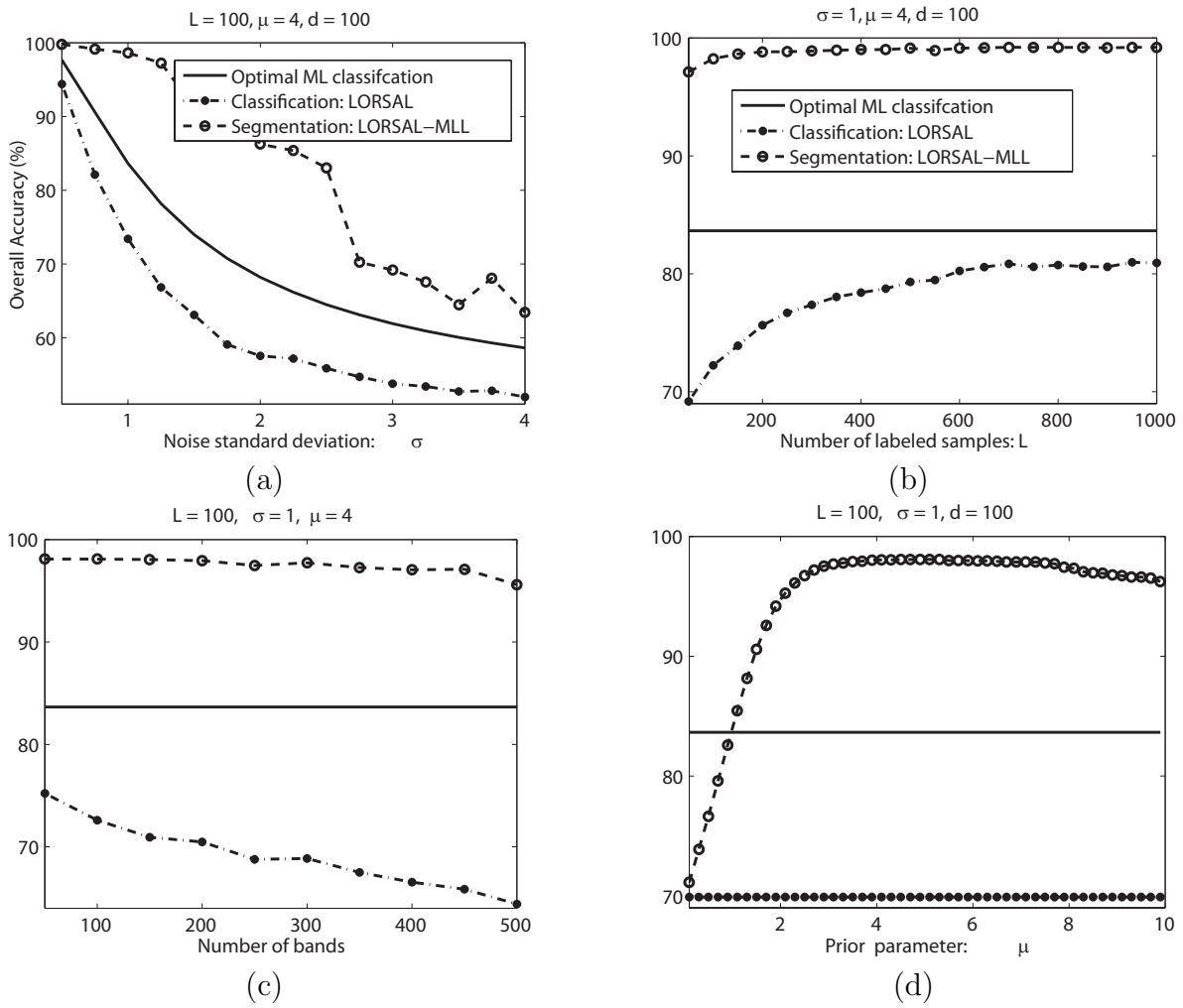


Fig. 3. OA results obtained by the proposed algorithms. Plot (a), OA results as a function of the noise standard deviation σ. Plot (b), OA results as a function of the number of labeled samples L. Plot (c), OA results as a function of the number of bands d. Plot (d), OA results as a function of the spatial prior parameter μ.

A.3 Impact of active learning approach

In this subsection, we analyze the impact of the proposed active learning approach. A new simulated hyperspectral data set is generated according to the model (12) with $K = 4$, $\sigma = 1$, and vectors $\mathbf{m}_{y_i}$ are signatures from the USGS library with $d = 224$.

Fig. 4 reports the learning results obtained as a function of L and U with $u = U/10$ (recall that U is the number of new sample actively selected, and u is the number of new active

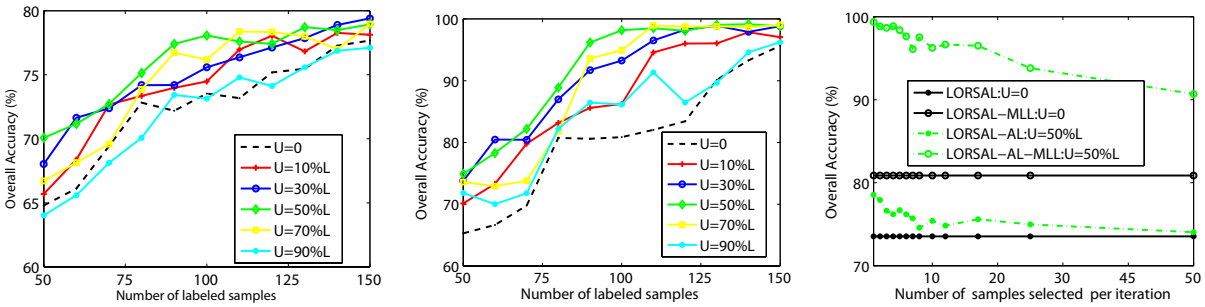


Fig. 4. OA results obtained by the proposed active learning approach: (a), classification OA results; (b), segmentation OA results; (c), OA results as a function of u with $L = 100$, $U = L_i = 50\%L$.

samples selected per iteration). Several conclusions can be obtained from Fig. 4. In general, the inclusion of the active learning improves the classification and segmentation performance. An important observation is that the advantages are less relevant as the size of the training set increases. This is expected, since the uncertainty of the classifier boundaries decreases as the training set size increases. Moreover, the contribution of the active learning depends both on the sizes of U and L_i. For $U \simeq L_i \simeq 50\%L$, the active learning leads to the best good results. It is worth noting that even with a very small amount of samples actively selected, *i.e.*, $U = 10\%L$, the proposed active learning procedure still performs better results than random selection. However, if $L_i \ll U$, in some cases, the active learning produces results even worse than the random selection. The explanation is that, with very small values of L_i, the initial estimate of the regressors $\widehat{\boldsymbol{\omega}}$ is very poor and thus, violating the active selection assumption, will suffer noticeable changes when the new label will be included. From plot (c), it can be observed that, the gain achieved by the proposed active learning approach increases as the size of u decreases. This is because the setting of $u > 1$, although speeding up the learning algorithm u times, is sub-optimal. Nevertheless, it is clear that the contribution to the segmentation performance is noticeable even with very large u. For this reason, and from a practical point of view, we set $L_i \simeq U$, $u \simeq U/10$ or $u \simeq U/5$.

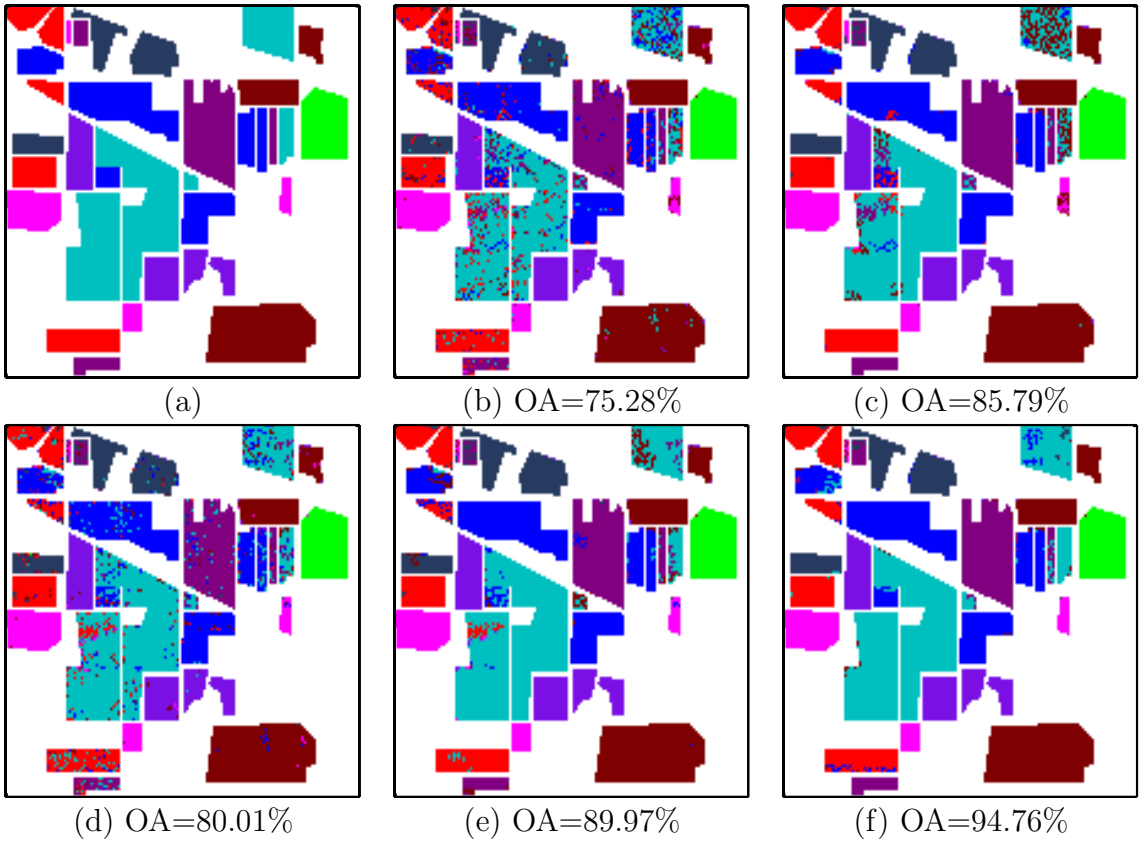
B. Experiments with real data sets

Four real hyperspectral data sets are used to evaluate our algorithm. The first one is the well-known AVIRIS Indian Pines scene, collected over Northwestern Indiana in June of 1992 [41]. The scene is available online[6] and contains 145×145 pixels and 224 spectral bands. A total of 20 spectral bands were removed prior to experiments due to noise and water absorption in those channels. The ground truth image, Fig.5 (a), contains 16 mutually exclusive ground-truth classes, 7 of which were discarded for insufficient number of training samples. The remaining 9 classes were used to generate a set of 4757 training samples, with random partition, and 4588 test samples.

The second data sets considered in this paper are based on urban hyperspectral data over the town of Pavia, Italy. The data set collected by the ROSIS sensor, operated by DLR (German Aerospace Agency) with a total of 115 spectral bands. Three different subsets of the full data set are considered in the experiments.

- Subset #1, with 492×1096 pixels in size, was collected over Pavia city center, Italy. The noisy bands were removed yielding a dateset with 102 spectral channels. The ground truth image (*see* Fig. 6 (a)) contains 9 ground truth classes, 5536 training samples, and 103539 test samples.
- Subset #2, with size of 610×340 pixels, is centered at the University of Pavia. The noisy bands were removed yielding 103 spectral channels. The ground truth image (*see* Fig. 7 (a)) contains 9 ground truth classes, 3921 training samples, and 42776 test samples.
- Subset #3, which is a superset of the scene over Pavia city centre, includes a dense residential area, with 715×1096 pixels. The ground truth image(*see* Fig. 6 (d))

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



(a) (b) OA=75.28% (c) OA=85.79%

(d) OA=80.01% (e) OA=89.97% (f) OA=94.76%

Fig. 5. Classification and segmentation maps. (a) Ground truth. (b) LORSAL classification map: $L = 237$. (c) LORSAL-MLL segmentation map: $L = 237$. (d) LORSAL-AL classification map: $L = 237$, $U = 126$. (e) LORSAL-AL-MLL segmentation map: $L = 237$, $U = 120$. (f) LORSAL-AL-MLL segmentation map: $L = 475$, $U = 230$.

contains 9 ground truth classes, 7456 training samples and 148152 validation samples.

B.1 Experiment 1: AVIRIS Indiana Pines Data Set

In this experiment, we use the AVIRIS Indian Pines data set to analyze the proposed algorithm. We use training sets with 5% (237 samples), 10% (475 samples) and 25% (1189 samples) of the original training set. For the active learning procedure, we set $U \simeq L_i \simeq L/2$ and $u = U/9$. The current problem is particularly complex and ill-posed because the number of training samples is only slightly higher than (or even comparable to) the size of the number of bands. Table I shows the results in comparison with the state-of-the-art competitors. Several conclusion can be obtained from Table I. First, the proposed MLR-based algorithms

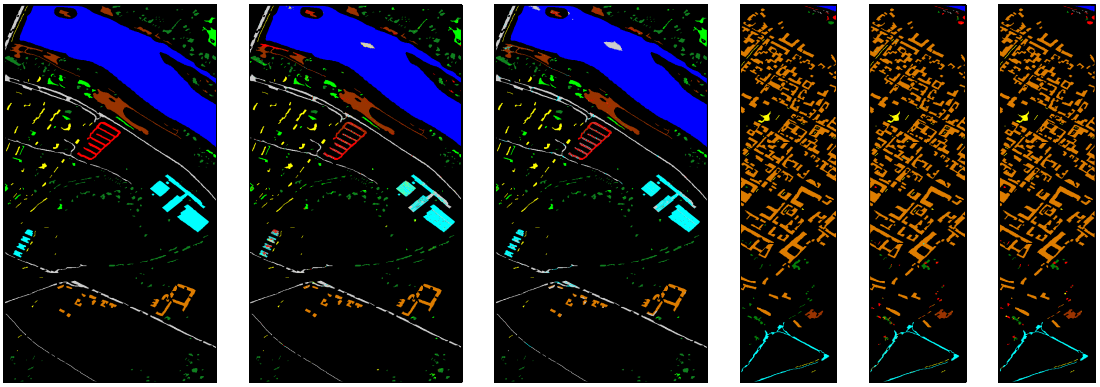


Fig. 6. Classification and segmentation maps for subset #1 and subset #3. (a) Ground truth of subset #1. (b) Classification map for subset #1 obtained by the LORSAL algorithm by using 10 labeled samples per class (OA=95.24%). (c) Classification map for subset #1 obtained by the LORSAL-AL algorithm by using 10 labeled samples per class, and $U = L/2$ (OA=96.25%). (d) Ground truth of subset #3. (e) Classification map for subset #3 obtained by the LORSAL-AL algorithm by using $L = 50$, and $U = L/2$ (OA=98.18%). (f) Segmentation map for subset #3 obtained by the LORSAL-AL-MLL algorithm (OA=98.41%).

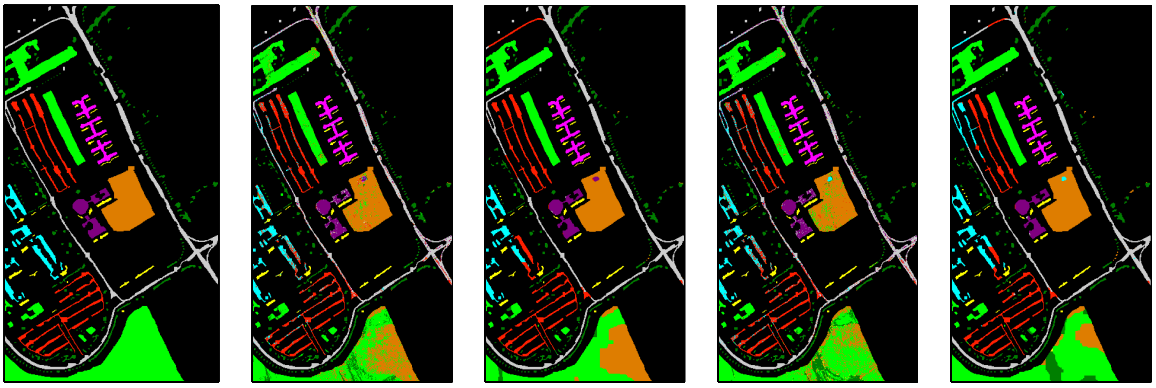


Fig. 7. Classification and segmentation maps for subset #2. (a) Ground truth. (b) Classification map for subset #2 obtained by the LORSAL algorithm with the full training set(OA=79.61%). (c) Segmentation map for subset #2 obtained by LORSAL-MLL algorithm (OA=86.60%). (d) Classification map for subset #2 obtained by the LORSAL-AL algorithm with $L = 650$, $U = 200$ (OA=81.92%). (e) Segmentation map for subset #2 obtained by LORSAL-AL-MLL algorithm (OA=89.31%).

outperforms the competitors in all cases. Our classification results obtained by the LORSAL algorithm are better than the results of TSVMs classifier [10]. This comparison is fair as all classifiers do not use spatial information and all training samples are randomly selected. Second, by including the spatial prior, the proposed segmentation method improves a lot the classification results provided by LORSAL (the improvement is always in the order

TABLE I

OA [%] RESULTS OBTAINED WITH THE PROPOSED ALGORITHM AS A FUNCTION OF THE NUMBER OF LABELED SAMPLE FOR THE AVIRIS INDIANA PINES, WITH $U \simeq L_i \simeq L/2$. THE PRIOR PARAMETER IS SET TO $\mu = 6$.

Training set		[10]		Proposed algorithm			
		Classification		Classification		Segmentation	
Percentage	L	SVMs	TSVMs	LORSAL	LORSAL-AL	LORSAL-MLL	LORSAL-AL-MLL
5%	237	73.41	76.20	76.60	80.40	85.09	89.30
10%	475	76.46	80.21	81.82	85.03	90.75	94.31
25%	1189	82.17	84.83	86.48	90.77	95.35	97.76

of 8% or higher). Furthermore, the proposed active learning procedure improves the OA results for both the classification and the segmentation algorithms. Finally, as expected the advantage of the active learning decreases as the size of the training set increases, since the uncertainty of the class boundaries decreases as the number of labeled samples increases. The effectiveness of the proposed method is illustrated in Fig. 5 where shows classification and segmentation maps.

B.2 Experiment 2: ROSIS Pavia Data Sets

In this section, the ROSIS Pavia data sets are used to evaluate the proposed approaches. The first experiment uses the ROSIS Pavia Data subset #1. Small size of training sets, composed of 10, 20, 40, 60, 80, 100 samples per class, are randomly selected from the complete training set. In all cases, we set $U = L/2 = L_i$ and $u = U/5$. Table II summarizes the results obtained by the classification algorithms. The best results are obtained by the LORSAL-AL algorithm. For a comparable OA, the LORSAL-AL algorithm requires much less labeled samples than the competitors. Moreover, the classification algorithms, both LORSAL and LORSAL-AL, generalize very well and are quite robust to small size training sets.

In the second experiment, we use subset #2. Two different scenarios are considered

TABLE II

OA [%] CLASSIFICATION RESULTS FOR THE ROSIS PAVIA SUBSET #1 WITH THE PROPOSED ALGORITHM BY USING $U = L/2$. THE PRIOR PARAMETER μ IS SET TO $\mu = 6$. MEAN, MIN, AND MAX DENOTES THE AVERAGE, MINIMUM AND MAXIMUM OA VALUE OBTAINED OVER 10 RUNS.

		Number of labeled samples per class					
Classification Algorithms		10	20	40	60	80	100
LORSAL	mean	95.22	96.27	96.91	97.03	97.39	97.37
	min	94.06	95.81	96.33	96.55	96.89	96.97
	max	96.14	96.79	97.35	97.40	97.71	97.66
LORSAL-AL	mean	95.94	97.50	98.49	99.02	99.34	99.48
	min	94.78	96.75	97.37	98.54	98.96	99.04
	max	97.04	98.20	99.90	99.41	99.69	99.81
[10]	$k_{gaussian}$	93.85	94.51	94.51	94.71	95.29	96.45
	k_{poly}	92.34	92.77	94.20	94.07	94.81	96.03
	k_{SAM}	93.32	93.87	93.79	94.23	94.54	95.56

in this experiment: (a) we use the entire training set to learn the classifiers; (b) we use a subset of the whole training set with $L_i = 450$ and $u = 100$ to train the classifiers. OA results are shown in Table III, in comparison with the results obtained by the extended morphological profile (EMP) [10], from which we conclude that the integration of active learning produces comparable results with less labeled samples: with 1050 labeled samples, we obtain a segmentation OA of 86.15% with the LORSAL-AL-MLL algorithm, which is better than the EMP result using 3921 training samples.

In the final experiment, we consider subset #3. Table IV summarizes the results for subset #3 in comparison with DAFE/MRF and Neuro-fuzzy [10], which are unsupervised algorithms. Maximum and minimum OAs obtained are also reported. With the entire training set, obviously, both classification and segmentation results obtained by the proposed algorithms are better than those obtained by DAFE/MRF and Neuro-fuzzy. For instance,

TABLE III

OA [%] RESULTS FOR THE ROSIS PAVIA SUBSET #2 WITH THE PROPOSED ALGORITHM. THE PRIOR PARAMETER μ IS SET TO $\mu = 2$. FOR THE LORSAL-AL AND LORSAL-AL-MLL ALGORITHM, $L_i = 450$, $u = 100$. CLASS. DENOTES CLASSIFICATION. SEG. DENOTES SEGMENTATION.

	[10]		Proposed algorithms							
	Class.	Seg.	Class.				Seg.			
	Spectral	EMP	LORSAL-AL			LORSAL	LORSAL-MLL	LORSAL-AL-MLL		
L	3921	3921	650	850	1050	3921	3921	650	850	1050
OA	80.99	85.22	78.60	79.86	80.44	80.24	86.02	85.39	85.58	86.15

even without spatial information, LORSAL algorithm obtained an OA of 98.61%, which is higher than those of the competitors. However, this comparison is not fair, as the competitors are unsupervised classifiers. For a fairer comparison, we considered small size of training sets, *i.e.*, $L = \{50,\ 80,\ 102,\ 120\}$ labeled samples. In this example, we set $U = L_i = L/2$ and $u = U/5$. It should be noticed that these are very difficult and complex problems as the number of labeled samples is even smaller than the number of bands. From Table IV, we conclude that the proposed algorithm performs very well in these circumstances. For instance, with only $L = 50$ labeled samples, which is half of the number of bands, we obtain a mean OA of 95.58% and 98.01% with LORSAL and LORSAL-AL algorithms,respectively. The latter result is better than those of the competitors and almost equals to the OA obtained by the entire training set. This result is very good as only 25 samples were actively selected from the original training set, which is a very small subset of the full image.

For illustration purposes, Fig. 6 (b)-(c), (e)-(f) and Fig. 7 (b)-(c) plot the classification and segmentation maps obtained by the proposed algorithms over the Pavia data sets. Effective results can be observed from these maps.

At this point, we want to call attention to the fact that, in all experiments, the set $\mathcal{D}_U$ is actively selected from the whole training set, which usually is a small subset of the complete data set. If users have more freedom to label new samples from a larger source,

TABLE IV

OA [%] RESULTS FOR THE ROSIS PAVIA SUBSET #3 WITH THE PROPOSED ALGORITHM BY USING $U = L/2$, WHICH ARE SELECTED FROM THE COMPLETE TRAINING SET THE PRIOR PARAMETER μ IS SET TO $\mu = 2$. MEAN, MIN, AND MAXDENOTES THE AVERAGE, MINIMUM AND MAXIMUM VALUE OVER 10 RUNS, RESPECTIVELY.

			Number of labeled samples (L)				
Algorithms			50	80	102	120	All
Classification	LORSAL	mean	95.58	96.52	96.64	97.00	98.61
		min	92.28	95.49	94.82	96.48	
		max	98.14	97.50	98.20	97.80	
	LORSAL-AL	mean	98.01	98.39	98.51	98.53	
		min	96.90	98.16	97.99	98.20	
		max	98.63	98.55	98.82	98.71	
Segmentation	LORSAL-MLL	mean	96.52	97.31	97.37	97.70	98.90
		min	94.31	96.03	96.12	97.23	
		max	98.52	98.30	98.57	98.50	
	LORSAL-AL-MLL	mean	98.25	98.51	98.41	98.70	
		min	97.21	98.05	97.87	98.11	
		max	98.76	98.88	98.63	98.90	
[10] Unsupervised Segmentation	DAFE/MRF		97.27				
	Neuro-fuzzy		97.29				

the performance is expected to be better. For instance, for subset #3, by setting $L = 80$, $U = L_i = L/2 = 40$, if the initial labeled set $\mathcal{D}_{L_i}$ is randomly selected from the whole training set and $\mathcal{D}_U$ is actively selected from the complete test set, an OA of 99.32% from the LORSAL-AL algorithm and an OA of 99.45% from the LORSAL-AL-MLL algorithm would be obtained, respectively.

V. Conclusions

In this paper, LORSAL-AL-MLL, a new supervised Bayesian segmentation approach aimed at ill-posed hyperspectral segmentation/classification problems has been introduced. LORSAL-AL-MLL models the posterior class probability distributions using a multinomial logistic regression (MLR), where the MLR regressors are learnt by the logistic regression via splitting and augmented Lagrangian (LORSAL) algorithm [26]. LORSAL-AL-MLL adopts a multi-level logistic (MLL) prior to model the spatial information present the class label images. The MAP segmentation is efficiently computed by the α-Expansion graph-cut based algorithm. With respect to the classification results just based on the learned class distribution (LORSAL), the segmentation algorithm (LORSAL-MLL) greatly improves the overall accuracies. Moreover, an active learning approach based on maximum mutual information between the regressors and class label, is used to reduce the acquisition of the ground reference data. The effectiveness of the proposed LORSAL-AL and LORSAL-AL-MLL algorithm is illustrated with both simulated and real hyperspectral datasets. A comparison with state-of-the-art methods indicates that the proposed method yields better or comparable performances using less, or much less, labeled samples.

Appendix

Problem (4) is equivalent to

$$(\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\nu}}) = \arg \min_{\boldsymbol{\omega},\, \boldsymbol{\nu}} \; -\ell(\boldsymbol{\omega}) + \lambda \|\boldsymbol{\nu}\|_1 \quad \text{subject to: } \boldsymbol{\omega} = \boldsymbol{\nu}. \tag{14}$$

By applying the alternating direction method of multipliers (ADMM) [43] (see also [44] and references therein) to solve the optimization (14), we get the following iterative algorithm:

Algorithm 4 Logistic regression via variable splitting and augmented Lagrangian (LORSAL)

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

1: t := 0
2: **repeat**
3: $\widehat{\boldsymbol{\omega}}^{(t+1)} \in \arg\min_{\boldsymbol{\omega}} \; -\ell(\boldsymbol{\omega}) + \frac{\beta}{2}\|\boldsymbol{\omega} - \boldsymbol{\nu}^{(t)} - \mathbf{b}^{(t)}\|^2$ (15)
4: $\widehat{\boldsymbol{\nu}}^{(t+1)} \in \arg\min_{\boldsymbol{\nu}} \; \lambda\|\boldsymbol{\nu}\|_1 + \frac{\beta}{2}\|\boldsymbol{\omega}^{(t+1)} - \boldsymbol{\nu} - \mathbf{b}^{(t)}\|^2$ (16)
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 meet

where $\beta \geq 0$ sets the augmented Lagrangian weight. Under mild conditions, the above sequence $\widehat{\boldsymbol{\omega}}^t$, for $t = 0, 1, 2 \ldots$ converges to a minimizer of (14), for any $\beta \geq 0$ [43].

The solution of the optimization (15) in line 3 is still a difficult problem because $\ell(\boldsymbol{\omega})$, although strictly convex and smooth, is non-quadratic and often very large. We tackle this difficulty by replacing $\ell(\boldsymbol{\omega})$ with a quadratic lower bound given by [16]

$$\ell(\boldsymbol{\omega}) \leq \ell(\boldsymbol{\omega}^{(t)}) + (\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^T \mathbf{g}(\boldsymbol{\omega}^{(t)}) + \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^T \mathbf{B}(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)}), \quad (17)$$

where $\mathbf{B} \equiv -(1/2)[\mathbf{I} - \mathbf{1}\mathbf{1}^T/K] \otimes \sum_{i=1}^{L} \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)^T$ (symbol $\mathbf{1}$ denotes a columns vector of ones) and $\mathbf{g}(\boldsymbol{\omega}^{(t)})$ is the gradient of ℓ at $\boldsymbol{\omega}^{(t)}$. Since the system matrix involved in the optimization (15), with $\ell(\boldsymbol{\omega})$ replaced with the quadratic bound given in (17), is fixed, its inverse can be pre-computed, provided that γ, the dimension of $\mathbf{h}(\mathbf{x}_i)$, is below, say, a few thousands. Under mild conditions, the converge of Algorithm 4 with the just referred modification still holds [43, 44].

The solution of the optimization in (16) in line 4 is simply the soft-threshold rule [45] given by $\widehat{\boldsymbol{\nu}}^{(t+1)} = \max\{\mathbf{0}, \text{abs}(\mathbf{u})\}\text{signal}(\mathbf{u})$, 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 refer that the complexity of each iteration of the LORSAL algorithm

is $O(\gamma^2 K)$, which is must faster than $O((\gamma K)^3)$, for the SMLR algorithm [17], and $O(\gamma^3 K)$, for FSMLR algorithm [18].

VI. Acknowledgments

This research was supported by the Marie Curie training Grant MEST-CT-2005-021175 from the European Commission and the IT Grant from Instituto de Telecomunicações. Funding from MRTN-CT-2006-035927 and AYA2008-05965-C04-02 projects is also gratefully acknowledged. The authors would like to thank Prof. D. Landgrebe for making the AVIRIS Indian Pines hyperspectral data set available to the community, Dr Paolo Gamba for providing the ROSIS data over Pavia, Italy, along with the training and test set, and Vladimir Kolmogorov for the max-flow/min-cut C++ code made available.

References

[1] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory, IT*, vol. 14, no. 1, pp. 55–63, 1968.

[2] M. Chi and L. Bruzzone, "Semi-supervised classification of hyperspectral images by svms optimized in the primal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, pp. 1870–1880, 2007.

[3] G. Camps-Valls, L. Gomez-Chova, J. Muoz-Mar, J. Vila-Francs, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, Jan 2006.

[4] J. Borges, J. Bioucas-Dias, and A. Marçal, "Evaluation of Bayesian hyperspectral imaging segmentation with a discriminative class learning," in *Proc. IEEE International Geoscience and Remote sensing Symposium*, Barcelona, Spain, 2007.

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

[6] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Proc. 16th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2002.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, 2007.

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

[9] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive svm for the semisupervised classification of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363–3373, 2006.

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

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

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

[13] M. Fauvel, J. Benediktsson, J. Chanussot, and J. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 11, pp. 3804–3814, 2008.

[14] M. Chi and L. Bruzzone, "A semi-labeled-sample driven bagging technique for ill-posed classification problems," *IEEE Geoscience and Remote Sensing Letters*, vol. 2, no. 1, pp.

69–73, 2005.

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

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

[17] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957–968, 2005.

[18] J. Borges, J. Bioucas-Dias, and A. Marçal, "Fast sparse multinomial regression applied to hyperspectral data," in *International Conference on Image Analysis and Recognition-ICIAR*, 2006.

[19] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. Benediktsson, "Svm and mrf- based method for accurate classification of hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, pp. 640–736, 2010.

[20] D. Mackay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, pp. 590–604, 1992.

[21] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo, "On semi-supervised classification," in *Proc. 18th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2004.

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

[23] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning

methods for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, 2009.

[24] W. Di and M. M. Crawford, "Locally consistend graph regularization based active learning for hyperspectral image classification," in *2rd WHISPERS*, 2010.

[25] G. Jun and J. Ghosh, "An efficient active learning algorithm with knowledge transfer for hyperspectral data analysis," in *Proc. IGARSS 2008, vol*, 2008.

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

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

[28] Y. Boykov, O. Veksler, and R. Zabih, "Efficient approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1222–1239, November 2001.

[29] H. DR and L. K., "A tutorial on MM algorithms," *The American Statistician*, vol. 58, pp. 30–37, 2004.

[30] J. Li, J. Bioucas-Dias, and A. Plaza, "Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Transactions on Geoscience and Remote Sensing (accepted)*, 2009.

[31] ——, "Semi-supervised hyperspectral image classification based on a markov random field and sparse multinomial logistic regression," in *Proc. IEEE International Geoscience and Remote sensing Symposium*, 2009.

[32] S. Geman and D. Geman, "Stochastic relaxation, gibbs distribution, andthe bayesian restoration of images," *IEEE TPAMI*, vol. 6, pp. 721–741, 1984.

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

UK, 1995.

[34] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. Royal Statistical Society B*, vol. 36, pp. 192–236, 1974.

[35] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision." *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, September 2004.

[36] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147–159, February 2004.

[37] J. Yedidia, W. Freeman, and Y.Weiss, "Understanding belief propagation and and its generalizations," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2001.

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

[39] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28.

[40] S. Bagon, "Matlab wrapper for graph cut," December 2006. [Online]. Available: http://www.wisdom.weizmann.ac.il/ bagon

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

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

[43] J. Eckstein and D. P. Bertsekas, "On the douglas-rachford splitting method and the

proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, pp. 293–318, 1992.

[44] M. V. Afonso, J. Bioucas-Dias, and M. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Transactions on Image Processing (accepted)*, 2010.

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