# Alternating direction optimization for image segmentation using hidden Markov measure field models

José Bioucas-Dias[a,b], Filipe Condessa[a,b,c,e], and Jelena Kovačević[c,d,e]

[a] Instituto de Telecomunicações, Lisboa, Portugal
[b] Instituto Superior Técnico, Lisboa, Portugal
[c] Department of ECE, Carnegie Mellon University, Pittsburgh, PA, USA
[d] Department of BME, Carnegie Mellon University, Pittsburgh, PA, USA
[e] Center for Bioimage Informatics, Carnegie Mellon University, Pittsburgh, PA, USA

## ABSTRACT

Image segmentation is fundamentally a discrete problem. It consists of finding a partition of the image domain such that the pixels in each element of the partition exhibit some kind of similarity. The solution is often obtained by minimizing an objective function containing terms measuring the consistency of the candidate partition with respect to the observed image, and regularization terms promoting solutions with desired properties. This formulation ends up being an integer optimization problem that, apart from a few exceptions, is NP-hard and thus impossible to solve exactly. This roadblock has stimulated active research aimed at computing "good" approximations to the solutions of those integer optimization problems. Relevant lines of attack have focused on the representation of the regions (*i.e.*, the partition elements) in terms of functions, instead of subsets, and on convex relaxations which can be solved in polynomial time.

In this paper, inspired by the "hidden Markov measure field" introduced by Marroquin et al. in 2003, we sidestep the discrete nature of image segmentation by formulating the problem in the Bayesian framework and introducing a hidden set of real-valued random fields determining the probability of a given partition. Armed with this model, the original discrete optimization is converted into a convex program. To infer the hidden fields, we introduce the Segmentation via the Constrained Split Augmented Lagrangian Shrinkage Algorithm (SegSALSA). The effectiveness of the proposed methodology is illustrated with simulated and real hyperspectral and medical images.

**Keywords:** Image segmentation, integer optimization, Markov random fields, hidden Markov measure fields, hidden fields, alternating optimization, Constrained Split Augmented Lagrangian Shrinkage Algorithm (SALSA), semi-supervised segmentation.

## 1. INTRODUCTION

Image segmentation plays a crucial role in many imaging and computer vision applications. Relevant examples are biomedical imaging (*e.g.*, quantification of tissue volumes, diagnosis, localization of pathology, study of anatomical structure, treatment planning, partial volume correction of functional imaging data, and computer integrated surgery[1]), remote sensing (*e.g.*, elaboration of thematic maps in hyperspectral imaging[2] and oil spill detection[3]), and computer vision (*e.g.*, stereo matching[4] and photo and video editing[5]).

The image segmentation problem consists in finding a partition of the image domain such that the image properties in a given partition element, expressed via image features or cues, are similar in some sense. Because image segmentation is almost invariably an ill-posed inverse problem, some form of regularization (a prior in Bayesian terms) is usually imposed on the solution with the objective of promoting solutions with desirable characteristics.

José Bioucas-Dias: E-mail: bioucas@lx.it.pt
Filipe Condessa: E-mail: condessa@cmu.edu
Jelena Kovacevic: E-mail: jelenak@cmu.edu

Thus, the features, the regularization, and the estimation criteria are main ingredients in image segmentation. The design of informative image features is problem dependent. Classical examples include color (spectral) vectors, morphological profiles, Gabor features, wavelet-based features, and combinations of local statistics. See Randen et al.[6] for a comparative study of relevant filtering approaches to texture feature extraction. The type of regularization and the estimation criteria used to infer a partition are relates issues. In the Bayesian framework, the segmentation is often obtained by computing the *maximum a posteriori probability* (MAP) estimate of the partition, which maximizes the product of likelihood function (*i.e.*, the probability of the observed image given the partition) with the prior probability for the partition, usually a Markov Random Field[7,8]. In the variational framework (*e.g.* active contours/snakes, geodesic active contours, level sets[9,10]), the image segmentation is obtained by finding the partition boundaries that minimize the sum of data misfit terms (interpretable as the negative likelihood in Bayesian terms) and regularization terms, such as length and area of the boundary. In the graph-based methods[11], the segmentation is obtained by formulating image segmentation as a graph partitioning problem, where the regularization is implicit in the definition of the partition of the graph.

Images of integers are natural representations for partitions. With this representation, the MAP segmentation, or the equivalent variational approach, is an integer optimization problem that, apart from a few exceptions, is NP-hard and thus impossible to solved exactly. In the last decade, a large class of powerful integer minimization methods based on graph cuts[5,12–14] and based on convex relaxations[15–17] has been proposed to solve approximately MAP estimation problems of discrete MRFs.

In this paper, inspired by the "hidden Markov measure fields" introduced by Marroquin et al.[18], we sidestep the hurdles raised by the discrete nature of image segmentation by (a) formulating the problem in the Bayesian framework and (b) introducing a hidden set of real-valued random fields conditioning the probability of a given partition. Armed with this model, we compute the marginal MAP (MMAP) estimate of the hidden fields, which is, under suitable conditions, a convex program. From the MMAP estimate of the hidden fields and the conditional probability of the partition, we obtain a soft and a hard estimate of the partition.

In the hidden field model, the prior on the partition is indirectly expressed by the prior on the hidden fields. In this paper, we use a form of vectorial total variation (VTV)[19,20], which promotes piecewise smooth segmentations and promotes sharp discontinuities between in the estimated partition.

### 1.1 Contributions

The main contributions of the paper are (a) the proposal of the VTV as prior on the hidden fields, (b) the proof that the MMAP estimate of the partition is a convex program, and (c) the introduction of an instance of the SALSA[21] algorithm, which we term SegSALSA, to compute the exact MMAP estimate of the partition with $O(Kn \ln n)$ complexity, where $K$ is the cardinality of the partition and $n$ the number of image pixels.

In this paper, we assume a supervised scenario in which the probability density functions of the feature vectors conditioned to a given partition element, also termed *class*, are known or were learned from a training set. We will provide, however, a qualitative discussion on the extension of the proposed methodology to unsupervised or semi-supervised scenarios via *expectation maximization* (EM)[22].

### 1.2 Related work

The work by Figueiredo[23] also approaches the image segmentation problem following closely the "hidden Markov measure fields" paradigm[18]. The main difference is in the statistical link and the prior on the hidden fields. The former is based on the multinomial logistic model and the latter on wavelets. The main advantage of using the multinomial logistic model is that it automatically enforces the nonnegativity and sum-to-one constraints linked with the probability of the classes given the hidden fields. However, this is true only for the EM algorithm proposed there and not for the supervised scenario considered here; we will discuss this in more detail in Section 2.4.

In Lellmann et al.[24] a multi-class labeling is approximately solved using tools from convex optimization. The approach therein proposed has links with ours in that it also uses a VTV regularizer and the optimization imposes constraints similar to ours. However, the data terms are different: ours is derived in under the a Bayesian framework whereas theirs is introduced heuristically. In addition, our optimization algorithm exploits the SALSA

splitting flexibility to avoid double loops as those shown in the Douglas-Rachford Splitting algorithm proposed in Lellmann et al.[24]

Finally, we mention the work by Sun et al.[25], which also uses non-isotropic TV as a regularizer and imposes constraints similar to ours. However, as in Lellmann et al.[24], the data term, introduced heuristically, measures the quadratic error norm between the probability vector obtained with sparse multinimial logistic regression and the optimization variables. The optimization problem is convex and solved in via ADMM.

The paper is organized as follows. Section 2 formulates the problem, introduces the hidden fields, the MMAP of the hidden fields, the statistical link between the class labels and the hidden fields, and the VTV prior. Section 3 presents the SegSALSA algorithm, which is an instantiation of SALSA to the problem in hand. Section 4 presents a number of experimental results with simulated and real hyperspectral and medical images. Finally, Section 5 present as few concluding remarks and pointers to future work.

## 2. PROBLEM FORMULATION

To formulate the segmentation problem in mathematical terms, we start by introducing notation. Let $\mathcal{S} \equiv \{1, \cdots, n\}$ denote a set of integers indexing the $n$ pixels of an image and $\mathbf{x} \equiv [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ a $d \times n$ matrix holding the $d$-dimensional image feature vectors. Given $\mathbf{x}$, the goal of image segmentation is to find a partition $P \equiv \{R_1, \ldots, R_K\}$ of $\mathcal{S}$ such that the feature vectors with indices in a given set $R_i$, for $i = 1, \ldots, K$, be *similar* in some sense*. Associated with a partition $P$, we introduce the image of class labels, also termed segmentation, $\mathbf{y} \equiv (y_1, \cdots, y_n) \in \mathcal{L}^n$, where $\mathcal{L} \equiv \{1, \ldots, K\}$ , such that $y_i = k$ if and only if $i \in R_k$. We remark that there is a one-to-one correspondence between partitions and segmentations.

### 2.1 Maximum *a posteriori* probability segmentation

We adopt a Bayesian perspective to the image segmentation problem. Under this perspective, the MAP segmentation is given by

$$\begin{aligned} \widehat{\mathbf{y}}_{MAP} &= \arg\max_{\mathbf{y} \in \mathcal{L}^n} p(\mathbf{y}|\mathbf{x}) \\ &= \arg\max_{\mathbf{y} \in \mathcal{L}^n} p(\mathbf{x}|\mathbf{y}) p(\mathbf{y}), \end{aligned} \tag{1}$$

where $p(\mathbf{y}|\mathbf{x})$ is the posterior probability† of $\mathbf{y}$ given $\mathbf{x}$, $p(\mathbf{x}|\mathbf{y})$ is the observation model, and $p(\mathbf{y})$ is the prior probability for the labeling $\mathbf{y}$.

An usual assumption in many low-level image problems is that of *conditional independence*[26], that is,

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \prod_{i=1}^{n} p(\mathbf{x}_i|y_i) \\ &= \prod_{k=1}^{K} \prod_{i \in R_k} p_k(\mathbf{x}_i), \end{aligned} \tag{2}$$

where $p_k(\mathbf{x}_i) = p(\mathbf{x}_i|y_i = k)$. For now, we assume that the class densities $p_k$, for $k \in \mathcal{L}$ are known or learned from a training set in a supervised fashion. Later, we discuss the semi-supervised scenario in which the class densities depend on parameters to be learned from a small training set and a large unlabeled set.

Using the Bayes law, we may write $p(\mathbf{x}_i|y_i) = p(y_i|\mathbf{x}_i)p(\mathbf{x}_i)/(p(y_i))$. Assuming that the *a priori* class probabilities $p(y_i)$, for $y_i \in \mathcal{L}$, are known, we conclude therefore that the class densities $p(\mathbf{x}_i|y_i)$ in (2) may be replaced with the discriminative probability model $p(y_i|\mathbf{x}_i)$ normalized by the respective *a priori* class probabilities $p(y_i)$.

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

†To keep the notation light, we denote both probability densities and probability distributions with $p(\cdot)$. Furthermore, the random variable to which $p(\cdot)$ refers is to be understood from the context.

The relevance of this replacement is linked with the fact that the discriminative models are usually less complex and yield, in the case small size training sets, better performance than the corresponding generative ones[27].

Various forms of Markov random fields (MRFs) have been widely used as prior probability for the class labels $\mathbf{y}$. A paradigmatic example is the multilevel logistic/Potts model (MLL)[7], which corresponds to the Ising model in the case of two classes. These models promote piecewise smooth segmentations, *i.e.*, segmentations in which it is more likely to have neighboring labels of the same class than the other way around.

The minimization in (1) is an integer optimization problem. In the case of MLL priors, the exact solution for $K = 2$ was introduced by mapping the problem into the computation of a min-cut on a suitable graph.[28] However, for $K > 2$, the computation of $\widehat{\mathbf{y}}_{MAP}$ in (1) is NP-hard and, therefore, impossible to solve exactly. Various algorithms to approximate $\widehat{\mathbf{y}}_{MAP}$ have been introduced in the last decade of which we highlight the graph cuts based $\alpha$-expansion[12], the sequential tree-reweighted message passing (TRW-S)[29], and the max-product loopy belief propagation (LBP)[30], and convex relaxations[15–17]. See Szeliski et al.[31] for an extensive comparison of these methods.

### 2.2 Hidden fields

The MAP formulation to image segmentation in terms of the image of class labels $\mathbf{y}$ raises a series of difficulties regarding (a) the high computational complexity involved in computing the solution of the integer optimization problem (1), (b) the selection of prior $p(\mathbf{y})$, which is often constrained to the availability of an effective minimization algorithm, and (c) the learning of unknown parameters $\boldsymbol{\theta}$ parameterizing the model $p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$, owing to the complexity usually involved in computing statistics with respect to $\mathbf{y}$.

These roadblocks have stimulated research on several fronts. A powerful approach, introduced in Marroquin et al.[18] reformulates the original segmentation problem in terms of real-valued hidden fields conditioning the random field $\mathbf{y}$ and endowed with a Gaussian MRF prior promoting smooth fields. The segmentation is obtained by computing the *marginal MAP* (MMAP) estimate of the hidden fields, which corresponds to a soft segmentation. The distinctive features of this approach are that it converts a hard integer optimization problem into a smooth and, under suitable conditions, constrained convex problem, thus much simpler to solve exactly using convex optimization tools.

### 2.3 Marginal MAP estimate of the hidden fields

To formulate the hidden field concept, and following closely Marroquin et al.,[18] let $\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{K \times n}$ denote a $K \times n$ matrix holding a collection of hidden random vectors, $\mathbf{z}_i \in \mathbb{R}^K$, for $i \in \mathcal{S}$ (one per pixel), and define the joint probability

$$p(\mathbf{y}, \mathbf{z}) = p(\mathbf{y}|\mathbf{z})p(\mathbf{z}),$$

with

$$p(\mathbf{y}|\mathbf{z}) = \prod_{i=1}^{n} p(y_i|\mathbf{z}_i).$$

With these definitions in place, the joint probability of $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is given by

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}|\mathbf{z})p(\mathbf{z}) \tag{3}$$

from which we may write the marginal density with respect to $(\mathbf{x}, \mathbf{z})$ as

$$p(\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} \left\{ \sum_{y_i \in \mathcal{L}} p(\mathbf{x}_i|y_i)p(y_i|\mathbf{z}_i) \right\} p(\mathbf{z}). \tag{4}$$

The MMAP estimate of the of the hidden field $\mathbf{z}$ is then given by

$$\begin{aligned}\widehat{\mathbf{z}}_{MMAP} &= \arg\max_{\mathbf{z}\in\mathbb{R}^{K\times n}} p(\mathbf{x},\mathbf{z}) \\ &= \arg\max_{\mathbf{z}\in\mathbb{R}^{K\times n}} \prod_{i=1}^{n}\left\{\sum_{y_i\in\mathcal{L}} p(\mathbf{x}_i|y_i)p(y_i|\mathbf{z}_i)\right\} p(\mathbf{z}).\end{aligned} \tag{5}$$

From $\widehat{\mathbf{z}}_{MMAP}$, we obtain the soft segmentation $p(\mathbf{y}|\widehat{\mathbf{z}}_{MMAP})$. A hard segmentation may be then obtained by computing

$$\widehat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{L}^n} p(\mathbf{y}|\widehat{\mathbf{z}}_{MMAP}).$$

### 2.4 The statistical link between the class labels and the hidden fields

The conditional probabilities $p(y_i|\mathbf{z}_i)$, for $i\in\mathcal{S}$, play a central role in our approach. As in Marroquin et al.[18], we adopt the following model

$$p(y_i = k|\mathbf{z}_i) \equiv [\mathbf{z}_i]_k \quad i\in\mathcal{S}, \quad k\in\mathcal{L}, \tag{6}$$

where the $[\mathbf{a}]_k$ stands for the $k$-th element of vector $\mathbf{a}$. Given that $[\mathbf{z}_i]_k$, for $k\in\mathcal{L}$, represents a probability distribution, then the hidden vectors $\mathbf{z}_i$, for $i\in\mathcal{S}$, shall satisfy the nonnegativity constraint $\mathbf{z}_i \geq 0$, where the inequality is to be understood componentwise, and the sum-to-one constraint $\mathbf{1}_K^T\mathbf{z}_i = 1$, where $\mathbf{1}_K$ stands for a column vectors of size $K$ containing only ones.

As explained in Section 3, the negative loglikelihood of the terms inside brackets in (6) are convex. This is of paramount importance, for it implies that the optimization (5) may be converted into a convex program, provided that the negative likelihood of the prior is also convex. To add even more interest to model (4) endowed with the statistic link (6), it also leads to convex terms in a semi-supervised scenario where the model parameters are learned via a suitable *expectation maximization algorithm* (EM), as explained in Section (2.6).

At this stage, we make reference to the work by Figueiredo[23] that has also approached the image segmentation problem following closely Marroquin et al.[18] The main difference concerns the statistical link and the class the prior. The former is based in the multinomial logistic model and the latter on wavelets. According to Figueiredo[23] , the main advantage of using the multinomial logistic model is that it automatically enforces the nonnegativity and sum-to-one constraints. However, this is true only for the EM algorithm proposed there and not for the supervised scenario where the terms inside brackets in (6) are nonconvex for the multinomial logistic model.

### 2.5 The prior

In this paper, we adopt form of vector total variation (VTV)[19,20] regularizer defined as

$$-\ln p(\mathbf{z}) \equiv \lambda_{TV}\sum_{n\in\mathcal{S}}\sqrt{\left\|\mathbf{D}_h\mathbf{z}[n]\right\|^2 + \left\|\mathbf{D}_v\mathbf{z}[n]\right\|^2} + c^{te}, \tag{7}$$

where and $\lambda_{TV} > 0$ is a regularization parameter controlling the strength of the prior, $\|\cdot\|$ is the standard Euclidean norm, and $\mathbf{D}_h, \mathbf{D}_v : \mathbb{R}^{K\times n} \mapsto \mathbb{R}^{K\times n}$ are linear operators computing horizontal and vertical first order backward differences, respectively; that is

$$\begin{aligned}\mathbf{D}_h\mathbf{z}[n] &\equiv \mathbf{z}_n - \mathbf{z}_{h(n)} \\ \mathbf{D}_v\mathbf{z}[n] &\equiv \mathbf{z}_n - \mathbf{z}_{v(n)}\end{aligned}$$

where $h(n)$ and $v(n)$ denote, respectively, horizontal and vertical backward neighbors of pixel $n$. Here, we assume cyclic boundaries.

The regularizer (7) has a number of desirable properties: (a) it promotes piecewise smooth hidden fields; (b) as any total variation regularizer, it tends to preserve discontinuities and, owing to the coupling among the classes introduced by the terms $\sqrt{\left\|\mathbf{D}_h\mathbf{z}[n]\right\|^2 + \left\|\mathbf{D}_v\mathbf{z}[n]\right\|^2}$, it tends to align the discontinuities among classes; (c) it is convex, although not strictly, and amenable to optimization via proximal methods relying on Moreau proximity operators[32].

### 2.6 Semi-supervised segmentation

As we have already referred to, we assume in this paper a supervised scenario in which the class densities $p(\mathbf{x}_i|y_i = k)$, for $k \in \mathcal{L}$, are known or were learned from a training set. In an unsupervised or semi-supervised scenario, those densities are often of the form $p(\mathbf{x}_i|y_i = k, \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k$, for $k \in \mathcal{S}$, are unknown vector parameters to be learned. We provide a brief discussion on the extension of the proposed methodology to unsupervised or semi-supervised scenarios.

The MMAP estimate of the couple $(\mathbf{z}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_1 \ldots, \boldsymbol{\theta}_K)$, is given by

$$\begin{aligned}(\widehat{\mathbf{z}}, \widehat{\boldsymbol{\theta}})_{MMAP} &= \arg\max_{\mathbf{z},\boldsymbol{\theta}} p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) \\ &= \arg\max_{\mathbf{z},\boldsymbol{\theta}} \prod_{i=1}^{n} \left\{ \sum_{y_i \in \mathcal{L}} p(\mathbf{x}_i|y_i, \boldsymbol{\theta}_{y_i}) p(y_i|\mathbf{z}_i) \right\} p(\mathbf{z}) p(\boldsymbol{\theta}),\end{aligned} \tag{8}$$

where $p(\boldsymbol{\theta})$ is the prior $\boldsymbol{\theta}$. A possible line of attack to solve the optimization (8) is applying alternating optimization with respect to $\mathbf{z}$ and to $\boldsymbol{\theta}$. The optimization with respect to $\mathbf{z}$ is as that in (5) and may be solved with the SegSALSA algorithm proposed in the next section. However, the optimization with respect to $\boldsymbol{\theta}$ is rather involving because neither $p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})$ nor any simple modification of it is decoupled with respect to $\boldsymbol{\theta}_k$, for $k \in \mathcal{S}$. This is true even under the assumption that those class vector parameters are statistically independent.

To circumvent the above difficulties, we propose an EM based algorithm[22] in which $\mathbf{x}$ is the observed data, $\mathbf{y}$ is the missing data, and the pair $(\mathbf{z}, \boldsymbol{\theta})$ is the entity to be inferred. At the $t$-th iteration, the E-step and M-step of the EM algorithm amount to compute, respectively,

$$\begin{aligned}\text{E-step:} \quad & Q(\mathbf{z}, \boldsymbol{\theta}; \mathbf{z}^t, \boldsymbol{\theta}^t) \equiv \mathbb{E}[\ln p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \boldsymbol{\theta})|, \mathbf{z}^t, \boldsymbol{\theta}^t] \\ \text{M-step:} \quad & (\mathbf{z}^{t+1}, \boldsymbol{\theta}^{t+1}) \equiv \arg\max_{\mathbf{z},\boldsymbol{\theta}} Q(\mathbf{z}, \boldsymbol{\theta}; \mathbf{z}^t, \boldsymbol{\theta}^t).\end{aligned}$$

Having in mind that $p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{S}} p(\mathbf{x}_i|y_i, \boldsymbol{\theta}_{y_i}) p(y_i|\mathbf{z}_i) p(\mathbf{z})$ and the statistical link $p(y_i|\mathbf{z}_i)$ given by (6), the function $Q(\mathbf{z}, \boldsymbol{\theta}; \mathbf{z}^t, \boldsymbol{\theta}^t)$ is convex with respect to $\mathbf{z}$ and is decoupled with respect to $\boldsymbol{\theta}_k$, for $k \in \mathcal{S}$, provided that these vector parameters are independent. These characteristics enable the design of effective *generalized* EM (GEM) algorithms in which the optimization with respect to $\mathbf{z}$ is a convex program similar to that of the supervised scenario and the optimization with respect to $\boldsymbol{\theta}_k$ is decoupled, for $k \in \mathcal{S}$. Further details of the proposed GEM approach are out of the scope of this paper and will be left for future work.

## 3. OPTIMIZATION ALGORITHM

Having in mind the model (6) and the prior (7), we may write the MMAP estimation of $\mathbf{z}$ as

$$\begin{aligned}\widehat{\mathbf{z}}_{MMAP} = \arg \min_{\mathbf{z} \in \mathbb{R}^{K \times n}} & \sum_{i=1}^{n} -\ln\left(\mathbf{p}_i^T \mathbf{z}_i\right) + \lambda_{TV} \sum_{n \in \mathcal{S}} \sqrt{\left\|\mathbf{D}_h \mathbf{z}[n]\right\|^2 + \left\|\mathbf{D}_v \mathbf{z}[n]\right\|^2} \\ \text{subject to:} \quad & \mathbf{z} \geq 0, \quad \mathbf{1}_K^T \mathbf{z} = \mathbf{1}_n^T,\end{aligned} \tag{9}$$

where $\mathbf{p}_i \equiv [p(\mathbf{x}_i|y_i = 1), \ldots, p(\mathbf{x}_i|y_i = K)]^T$ and it was assumed that $\mathbf{p}_i^T \mathbf{z}_i > 0$ for $\mathbf{z}_i$ in the feasible set[‡]. A straightforward calculus of Hessian matrix of $-\ln\left(\mathbf{p}_i^T \mathbf{z}_i\right)$ yields

$$\frac{\partial^2}{\partial \mathbf{z}_i \partial \mathbf{z}_i^T} \left(-\ln\left(\mathbf{p}_i^T \mathbf{z}_i\right)\right) = \frac{\mathbf{p}_i \mathbf{p}_i^T}{(\mathbf{p}_i^T \mathbf{z}_i)^2},$$

which is a semipositive definite matrix. We conclude therefore that (9) is convex.

[‡]This assumption may be discarded.

In this section, we develop an instance of the Split Augmented Lagrangian Shrinkage (SALSA) methodology introduced by Afonso et al.[21] to compute $\widehat{\mathbf{z}}_{MMAP}$. We start by rewriting the optimization (9) in the following equivalent format more suitable to SALSA:

$$\min_{\mathbf{z}\in\mathbb{R}^{K\times n}} \sum_{i=1}^{4} g_i(\mathbf{H}_i\mathbf{z}), \tag{10}$$

where $g_i$, for $i = 1,\ldots,4$, denote, closed, proper, and convex functions, and $\mathbf{H}_i$, for $i = 1,\ldots,4$, denote linear operators. The particular definitions of these entities for our problem are as follows:

$$\begin{array}{l} \mathbf{H}_1 = \mathbf{I},\ \mathbf{H}_2 = \begin{pmatrix}\mathbf{D}_h\\ \mathbf{D}_v\end{pmatrix},\ \mathbf{H}_3 = \mathbf{I},\ \mathbf{H}_4 = \mathbf{I},\\ g_1(\boldsymbol{\xi}) = \sum_{n\in\mathcal{S}} -\ln\left(\mathbf{p}_n^T\boldsymbol{\xi}_n\right)_+,\\ g_2(\boldsymbol{\xi}) = \lambda_{TV}\sum_{n\in\mathcal{S}}\sqrt{\left\|\boldsymbol{\xi}^h[n]\right\|^2 + \left\|\boldsymbol{\xi}^v[n]\right\|^2},\\ g_3(\boldsymbol{\xi}) = \iota_+(\boldsymbol{\xi}),\\ g_4(\boldsymbol{\xi}) = \iota_1(\mathbf{1}_K^T\boldsymbol{\xi}), \end{array} \tag{11}$$

where $\mathbf{I}$ denotes the identity operator, $\boldsymbol{\xi}$ are dummy variables whose dimensions depend on the functions $g_i$, for $i = 1,2,3,4$, $(x)_+ \equiv \max\{0,x\}$ is the positive part of $x$, and $\ln(0) \equiv +\infty$. In the case of $g_2$, we have $\boldsymbol{\xi} = [(\boldsymbol{\xi}^h)^T,(\boldsymbol{\xi}^v)^T]^T$ where $\boldsymbol{\xi}^h$ and $\boldsymbol{\xi}^v$ are in the range of $\mathbf{D}^h$ and $\mathbf{D}^v$, respectively. The function $\iota_+$ denotes the indicator in the set in $\mathbb{R}_+^{K\times n}$, *i.e.*, $\iota_+(\boldsymbol{\xi}) = 0$ if $\boldsymbol{\xi}\in\mathbb{R}_+^{K\times n}$ and $\iota_+(\boldsymbol{\xi}) = \infty$ otherwise. By the same token $\iota_1(\boldsymbol{\xi})$ is the indicator in the set $\{\mathbf{1}_n\}$.

We now introduce the variable splitting $\mathbf{u}_i = \mathbf{H}_i\mathbf{z}$, for $i = 1,2,3,4$, in (10) and convert the original optimization into the equivalent constrained form

$$\boxed{\min_{\mathbf{u},\mathbf{z}} \sum_{i=1}^{4} g_i(\mathbf{u}_i) \quad \text{subject to} \quad \mathbf{u} = \mathbf{G}\mathbf{z},} \tag{12}$$

where $\mathbf{u}_1,\mathbf{u}_3,\mathbf{u}_4 \in \mathbb{R}^{K\times n}$, $\mathbf{u}_2,\in\mathbb{R}^{2K\times n}$, $\mathbf{u} \equiv [\mathbf{u}_1^T,\ldots\mathbf{u}_4^T]^T \in \mathbb{R}^{5K\times n}$, and $\mathbf{G} : \mathbb{R}^{K\times n} \mapsto \mathbb{R}^{5K\times n}$ is the linear operator obtained by columnwise stacking the operators $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$, and $\mathbf{H}_4$.

The next step consists in applying the SALSA methodology[21] to (12). SALSA is essentially an instance of the alternating method of multipliers (ADMM)[33–35] designed to optimize sums of an arbitrary number of convex terms. The following is a simplified version of a theorem by Eckstein and Bertsekas, adapted to our setting, stating convergence of SALSA[33–35]. The notation $\mathbf{d} = [\mathbf{d}_1^T,\mathbf{d}_2^T,\mathbf{d}_3^T,\mathbf{d}_4^T]^T$ stands for scaled Lagrange multipliers associated with the equality constraint $\mathbf{u} = \mathbf{G}\mathbf{z}$, where $\dim(\mathbf{d}_i) = \dim(\mathbf{H}_i\mathbf{z})$.

THEOREM 3.1. *Assume that Null*$(\mathbf{G}) = \{\mathbf{0}\}$, i.e., *the null space of operator* $\mathbf{G}$ *is* $\{\mathbf{0}\}$, *and let* $f(\mathbf{u}) = \sum_{i=1}^{J} g_i(\mathbf{u}_i)$ *be closed, proper, and convex. Consider arbitrary* $\mu > 0$ *and* $\mathbf{z}_0,\mathbf{d}_0$. *Consider three sequences* $\{\mathbf{z}^k,\ k = 0,1,...\}$, $\{\mathbf{u}^k,\ k = 0,1,...\}$, *and* $\{\mathbf{d}^k,\ k = 0,1,...\}$ *that satisfy*

$$\mathbf{z}^{k+1} = \arg\min_{\mathbf{z}} \left\|\mathbf{G}\mathbf{z} - \mathbf{u}^k - \mathbf{d}^k\right\|_F^2, \tag{13}$$

$$\mathbf{u}^{k+1} = \arg\min_{\mathbf{u}} f(\mathbf{u}) + \frac{\mu}{2}\left\|\mathbf{G}\mathbf{z}^{k+1} - \mathbf{u} - \mathbf{d}^k\right\|_F^2, \tag{14}$$

$$\mathbf{d}^{k+1} = \mathbf{d}^k - \left[\mathbf{G}\mathbf{z}^{k+1} - \mathbf{u}^{k+1}\right], \tag{15}$$

*where* $\|\mathbf{x}\|_F \equiv \sqrt{\text{trace}\{\mathbf{x}\mathbf{x}^T\}}$ *stands for the Frobenius norm. Then, if (12) has a solution, the sequence* $\{\mathbf{z}^k\}$ *converges to it; otherwise, at least one of the sequences* $\{\mathbf{u}^k\}$ *or* $\{\mathbf{d}^k\}$ *diverges.*

Given that the linear operator $\mathbf{G}$ in (12) has $\text{Null}(\mathbf{G}) = \{\mathbf{0}\}$, that the objective functions are closed, proper, and convex and that (12) has solutions[§], then the sequence $\mathbf{z}^k$ generated by (13 – 15) converges to a solution of (12) for any $\mu > 0$.

[§]Given that the feasible set is compact, the conditions $\mathbf{p}_i^T\mathbf{z}_i > 0$, for $i\in\mathcal{S}$, for any point $\mathbf{z}$ of the feasible set implies that objective function (9) is continuous on the feasible set and thus it has a minimum point.

**Algorithm** *SALSA*
1. Set $k = 0$, choose $\mu > 0$, $\mathbf{u}^0 = (\mathbf{u}_1^0, \mathbf{u}_2^0, \mathbf{z}_3^0, \mathbf{z}_4^0)$
2. Set $\mathbf{d}^0 = (\mathbf{d}_1^0, \mathbf{d}_2^0, \mathbf{d}_3^0, \mathbf{d}_4^0)$
3. **repeat**
4. ($*$ update $\mathbf{z}$ $*$)
5. $\mathbf{z}^{k+1} := \arg\min_{\mathbf{z}} \left\|\mathbf{G}\mathbf{z} - \mathbf{u}^k - \mathbf{d}^k\right\|_F^2$
6. ($*$ update $\mathbf{u}$ $*$)
7. **for** $i = 1$ **to** $i = 4$
8. **do** $\boldsymbol{\nu}_i := \mathbf{H}_i\mathbf{z}^{k+1} - \mathbf{d}_i^k$
9. ($*$ apply Moreau proximity operators $*$)
10. $\mathbf{u}_i^{k+1} := \arg\min_{\mathbf{u}_i} g_i(\mathbf{u}_i) + \frac{\mu}{2}\left\|\mathbf{u}_i - \boldsymbol{\nu}_i\right\|_F^2$
11. ($*$ update Lagrange multipliers $\mathbf{d}$ $*$)
12. $\mathbf{d}_i^{k+1} := -\boldsymbol{\nu}_i + \mathbf{u}_i^{k+1}$
13. $k \leftarrow k + 1$
14. **until** stopping criterion is satisfied.

Figure 1. Augmented Lagrangian Shrinkage Algorithm (SALSA).

Fig. 2 shows the pseudocode of the SALSA algorithm. A distinctive feature of SALSA is that optimization with respect to $\mathbf{u}$ is decoupled into optimization problems with respect to the blocks $\mathbf{u}_i$, for $i = 1, 2, 3, 4$, whose solutions are the so-called Moreau proximity operators (MPOs)[32] for the respective convex functions $g_i$, for $i = 1, 2, 3, 4$. In order to implement SALSA, we need to solve the quadratic optimization problem in line 5 and to apply the Moreau proximity operators in line 10. Below, we present the solutions to these optimization subproblems.

### 3.1 Optimization with respect to z

The solution of the quadratic optimization on line 5 is given by

$$\mathbf{z}^{k+1} = \left(\mathbf{G}^*\mathbf{G}\right)^{-1}\mathbf{G}^*\left(\mathbf{u}^k - \mathbf{d}^k\right) = \left(\mathbf{D}^*\mathbf{D} + 3\mathbf{I}\right)^{-1}\left(\sum_{i=1,3,4}\left(\mathbf{u}_i^k + \mathbf{d}_i^k\right) + \mathbf{D}^*(\mathbf{u}_2^k + \mathbf{d}_2^k)\right),$$

where the notation $(\cdot)^*$ stands for adjoint operation with respect to the Frobenius norm. Having in mind that the operator $\mathbf{D}$ is the columnwise stacking of operators $\mathbf{D}_h$ and $\mathbf{D}_h$ and that $\mathbf{D}_h\mathbf{z}$ compute independent cyclic convolutions on each image of $\mathbf{z}$, then the computation of $\mathbf{z}^{k+1}$ can be carried out efficiently in the frequency domain using the fast Fourier transform (FFT) with $O(Kn\ln n)$ complexity.

### 3.2 Moreau proximity operators

The optimization subproblems shown in line 10 correspond to evaluating the Moreau proximity operators[32] of the convex functions $g_1$, $g_2$, $g_3$, and $g_4$. In this section, we present closed form expressions for these operators.

#### 3.2.1 Moreau proximity operator for $g_1$

$$\psi_{g_1/\mu}(\boldsymbol{\nu}) = \arg\min_{\boldsymbol{\xi}}\left(-\sum_{i=1}^{n}\ln\left(\mathbf{p}_i^T\boldsymbol{\xi}_i\right)_+\right) + (\mu/2)\left\|\boldsymbol{\xi} - \boldsymbol{\nu}\right\|_F^2, \tag{16}$$

where $\boldsymbol{\nu} \equiv [\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_n] \in \mathbb{R}^{K\times n}$, $\boldsymbol{\xi} \equiv [\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n] \in \mathbb{R}^{K\times n}$. The optimization (16) is decoupled with respect to $\boldsymbol{\xi}_i$, for $i \in \mathcal{S}$, and then it follows that

$$\psi_{g_1/\mu}(\boldsymbol{\nu}) = \left(\psi_{g_1/\mu}(\boldsymbol{\nu}_1), \ldots, \psi_{g_1/\mu}(\boldsymbol{\nu}_n)\right) \tag{17}$$

with

$$\psi_{g_1/\mu}(\boldsymbol{\nu}_i) = \arg\min_{\boldsymbol{\xi}_i} -\ln\left(\mathbf{p}_i^T\boldsymbol{\xi}_i\right)_+ + (\mu/2)\left\|\boldsymbol{\xi}_i - \boldsymbol{\nu}_i\right\|^2 = \boldsymbol{\nu}_i + \frac{\mathbf{p}_i}{\mu\, a_i}, \tag{18}$$

where

$$a_i \equiv \frac{\mathbf{p}_i^T\boldsymbol{\nu}_i + \sqrt{\left(\mathbf{p}_i^T\boldsymbol{\nu}_i\right)^2 + \|\mathbf{p}_i\|^2/\mu}}{2}.$$

The expression (18) was derived by computing the positive root of the equation $\mathbf{p}_i^T\boldsymbol{\nabla}\psi_{g_1/\mu} = \mathbf{0}$ with respect to $\mathbf{p}_i^T\boldsymbol{\xi}_i$ and then using this root in the equation $\boldsymbol{\nabla}\psi_{g_1/\mu} = \mathbf{0}$. The complexity to compute $\psi_{g_1/\mu}$ is $O(KN)$.

### 3.2.2 Moreau proximity operator for $g_2$

$$\psi_{g_2\lambda_{TV}/\mu}(\boldsymbol{\nu}) = \arg\min_{\boldsymbol{\xi}} \lambda_{TV}\sum_{n\in\mathcal{S}}\sqrt{\left\|\boldsymbol{\xi}^h[n]\right\|^2 + \left\|\boldsymbol{\xi}^v[n]\right\|^2} + (\mu/2)\left\|\boldsymbol{\xi} - \boldsymbol{\nu}\right\|_F^2, \tag{19}$$

where $\boldsymbol{\nu}, \boldsymbol{\xi} \in \mathbb{R}^{2K\times n}$ and $\boldsymbol{\xi}^h, \boldsymbol{\xi}^v \in \mathbb{R}^{2K\times n}$. The optimization (19) is pixelwise decoupled and yields the vector-soft-thresholding operator[32]

$$\psi_{g_2\lambda_{TV}/\mu}(\boldsymbol{\nu})[n] = \max\left\{\mathbf{0}, \|\boldsymbol{\nu}[n]\| - \lambda_{TV}/\mu\right\}\frac{\boldsymbol{\nu}[n]}{\|\boldsymbol{\nu}[n]\|}. \tag{20}$$

The complexity to compute $\psi_{g_2\lambda_{TV}/\mu}$ is $O((K+1)N)$.

### 3.2.3 Moreau proximity operator for $g_3$

$$\psi_{g_3/\mu}(\boldsymbol{\nu}) = \arg\min_{\boldsymbol{\xi}} \iota_+(\boldsymbol{\xi}) + (\mu/2)\left\|\boldsymbol{\xi} - \boldsymbol{\nu}\right\|_F^2 = \arg\min_{\boldsymbol{\xi}\geq 0} \left\|\boldsymbol{\xi} - \boldsymbol{\nu}\right\|_F^2 = \max\{\mathbf{0}, \boldsymbol{\nu}\},$$

where $\boldsymbol{\nu}, \boldsymbol{\xi} \in \mathbb{R}^{K\times n}$. The MPO $\psi_{g_3/\mu}$ is the projection in the first orthant and has complexity $\psi_{g_3/\mu}$ is $O(KN)$.

### 3.2.4 Moreau proximity operator for $g_4$

$$\begin{aligned}
\psi_{g_4/\mu}(\boldsymbol{\nu}) &= \arg\min_{\boldsymbol{\xi}} \iota_1(\mathbf{1}_K^T\boldsymbol{\xi}) + (\mu/2)\left\|\boldsymbol{\xi} - \boldsymbol{\nu}\right\|_F^2 \\
&= \arg\min_{\boldsymbol{\xi}} \left\|\boldsymbol{\xi} - \boldsymbol{\nu}\right\|_F^2 \quad \text{subject to} \quad \mathbf{1}_K^T\boldsymbol{\xi} = \mathbf{1}_n^T \\
&= \left(\mathbf{I} - \frac{\mathbf{1}_K\mathbf{1}_K^T}{K}\right)\boldsymbol{\nu} + \frac{\mathbf{1}_K\mathbf{1}_n^T}{K},
\end{aligned}$$

where $\boldsymbol{\nu}, \boldsymbol{\xi} \in \mathbb{R}^{K\times n}$. The MPO $\psi_{g_4/\mu}$ is the projection in the probability simplex and has complexity $O(KN)$.

### 3.2.5 The SegSALSA algorithm

Fig. 2 shows the pseudocode for the proposed instance of the SALSA algorithm, which we term Segmentation via Augmented Lagrangian Shrinkage Algorithm (SegSALSA). SegSALSA converges for any $\mu > 0$. However, the convergence speed is highly sensitive to the value of $\mu$. This issue is currently a hot research topic. In this work, we have implemented the selection rule discussed in [36, Ch. 3.4] and therein formalized in expression (3.13). Nevertheless, we have observed experimentally that a value of $\mu \simeq 5$ yields nearly optimum convergence speed. Regarding the stopping criterion, we impose that the primal and dual residuals be smaller than a given threshold, as suggested in [36, Ch. 3.3.2]. We have observed, however, that a fixed number of iterations of the order of 200 provides excellent results.

Having in mind the computational computational complexities involved in the computation of $\mathbf{x}$ and of the MPOs for $g_1, g_2, g_3, g_4$, we conclude that the SegSALSA computational complexity per iteration is dominated by the term $O(Kn\ln n)$, associated to the computation $\mathbf{z}^{k+1}$ shown in line 5 of SegSALSA.

**Algorithm** *SegSALSA*

Set $k = 0$, choose $\mu > 0$, $\mathbf{u}^0 = (\mathbf{u}_1^0, \mathbf{u}_2^0, \mathbf{z}_3^0, \mathbf{z}_4^0)$

Set $\mathbf{d}^0 = (\mathbf{d}_1^0, \mathbf{d}_2^0, \mathbf{d}_3^0, \mathbf{d}_4^0)$

**repeat**

($*$ update $\mathbf{z}$ $*$)

$$\mathbf{z}^{k+1} := (\mathbf{D}^*\mathbf{D} + 3\mathbf{I})^{-1} \left( \sum_{i=1,3,4} (\mathbf{u}_i^k + \mathbf{d}_i^k) + \mathbf{D}^*(\mathbf{u}_2^k + \mathbf{d}_2^k) \right)$$

($*$ update $\mathbf{u}$ using the Moreau proximity operators $*$)

$\boldsymbol{\nu}_1 := \mathbf{z}^{k+1} - \mathbf{d}_1^k$

$\mathbf{u}_1^{k+1} := \psi_{g_1/\mu}(\boldsymbol{\nu}_1)$

$\boldsymbol{\nu}_2 := \mathbf{D}_2\mathbf{z}^{k+1} - \mathbf{d}_2^k$

$\mathbf{u}_2^{k+1} := \psi_{g_2\lambda_{TV}/\mu}(\boldsymbol{\nu}_2)$

$\boldsymbol{\nu}_3 := \mathbf{z}^{k+1} - \mathbf{d}_3^k$

$\mathbf{u}_3^{k+1} := \psi_{g_3/\mu}(\boldsymbol{\nu}_3)$

$\boldsymbol{\nu}_4 := \mathbf{z}^{k+1} - \mathbf{d}_4^k$

$\mathbf{u}_4^{k+1} := \psi_{g_4/\mu}(\boldsymbol{\nu}_4)$

($*$ update Lagrange multipliers $\mathbf{d}$ $*$)

**for** $i = 1$ **to** $i = 4$

**do** $\mathbf{d}_i^{k+1} := -\boldsymbol{\nu}_i + \mathbf{u}_i^{k+1}$

$k \leftarrow k + 1$

**until** stopping criterion is satisfied.

Figure 2. Segmentation via Augmented Lagrangian Shrinkage Algorithm (SegSALSA).

## 4. RESULTS

In this section, we report experimental results that illustrate the effectiveness of the proposed method with simulated and real hyperspectral and medical images. For the simulated images, we assume known class densities, whereas for the real images, we will learn a discriminative class model using the LORSAL algorithm.[37] The segmentation performance is measured in term of pixelwise accuracy (ratio between the number of pixels correctly classified and the total number of pixels).

### 4.1 Simulated images

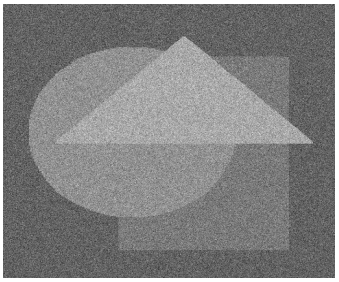
The simulated experiment (Fig. 3) shows the segmentation of a simple synthetic image with four classes and known class models. The four regions of the image follows a Gaussian distribution with means $1, 2, 3, 4$ and standard deviation 1. Knowing the class models, we obtain a maximum likelihood segmentation, a graph cut segmentation, and a SegSALSA segmentation. In the case of graph cuts, we compute the MAP segmentation given by (1), where $p(\mathbf{y})$ is an MLL[7] MRF. Both the MLL parameter, controlling the weight of the prior $p(\mathbf{y})$, and the parameter $\lambda_{TV}$ in the VTV prior/regularizer were hand tuned for optimal performance.

Two results emerge from this experiment. First, the use of a contextual prior causes an significant improvement on the performance of the segmentation. This is clear when comparing the maximum likelihood segmentation (59% accuracy) with either the graph cut segmentation (95% accuracy) or the SegSALSA segmentation (99% accuracy). Second, SegSALSA yields better performance than the graph cut segmentation. The advantage of SegSALSA is certainly due to the fact that the underlying optimization is exact, which is not the case with the graph cuts, and due to the better adequacy of the isotropic VTV prior, which allows for clearer boundary preservation. Fig. 3, part (6), illustrates the hidden field $[\mathbf{z}_i]_1$, taking a value close to 1 for pixels belonging to class 1. The ability to preserve sharp discontinuities along different directions is also illustrated.

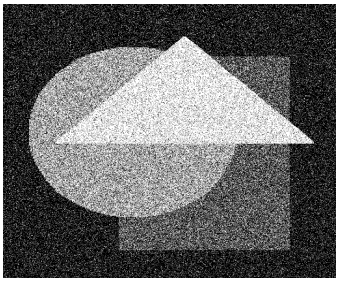
### 4.2 Real images

We now present two experiments with real images, showing the effectiveness of the segmentation obtained and exploring the effect of different weights on the VTV prior (in the medical image experiment) and different dimensions of the training set (in the hyperspectral image experiment).

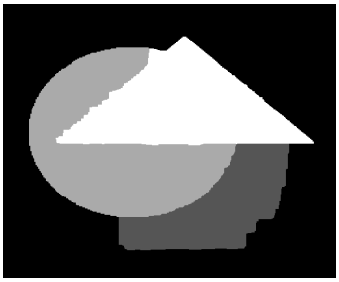
(1) True regions.



(2) Observed image.

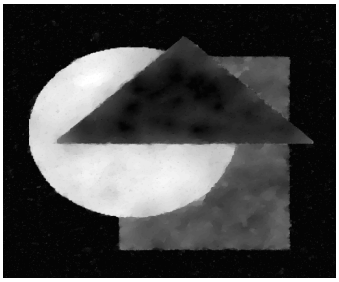


(3) ML segmentation.



(4) Graph cut segmentation.

(5) SegSALSA segmentation.



(6) Latent probabilities.

Figure 3. Segmentation of simulated image. Top row: (1) true regions, (2) observed image, (3) maximum likelihood segmentation (59% accuracy). Bottom row: (4) graph cut segmentation (95% accuracy), (5) SegSALSA segmentation (99% accuracy), (6) latent probability for light gray class.

#### 4.2.1 Histology image

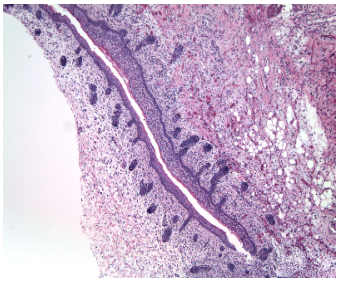
In this experiment (Fig. 4), we apply the SegSALSA algorithm to the task of digital pathology image classification, specifically classification of Hematoxilyn & Eosin (H&E) stained teratoma tissues. The classification of this class of images is extremely difficult at a pixel or small patch level, as there is very high intraclass variability and low interclass variability. Furthermore, a large number of training samples are needed to accurately classify the images.

In this experiment we classify a $1600 \times 1200$ Hematoxilyn & Eosin stained teratoma tissue imaged at 40X magnification, with 4 different classes and using a very small number of training samples. To this extent, we use $4 \times 4$ non-overlapping patches, with 100 randomly selected training patches per class, amounting to 0.3% of the data set used for training purposes. The class models are learned using the LORSAL algorithm, which has shown good classification performance in digital pathology[38].

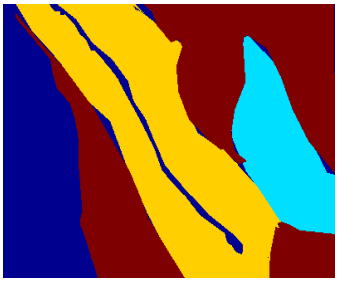
Our aim in this experiment is both to show the performance of the algorithm on hard image classification problems with small training samples, and to show the effect of the weight of the VTV prior ($\lambda_{TV}$) on the resulting classification. As seen in the bottom row of Fig. 4, it is possible to obtain a smoother classification with larger values of $\lambda_{TV}$ without loss of sharp boundaries between the classes. The value of accuracy = 84% obtained with $\lambda_{TV} = 4$ is considered state-of-the-art.

#### 4.2.2 Hyperspectral image

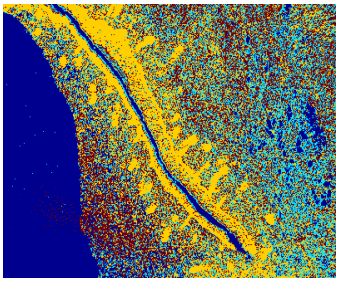
In this experiment (Fig. 5), we use the SegSALSA algorithm to classify the ROSIS Pavia scene, an hyperspectral image widely used in hyperspectral image classification. This hyperspectral image was acquired by the ROSIS optical sensor on the University of Pavia, Italy. It is a $610 \times 340$ image with a spatial resolution of $1.3m$/pixel, and 103 spectral bands. The image contains nine exclusive land-cover classes, with the accuracy of the classification being measured on those nine classes. The class models are learned using the LORSAL algorithm.



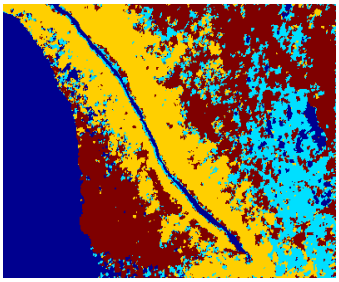
(1) H&E stained teratoma tissue.



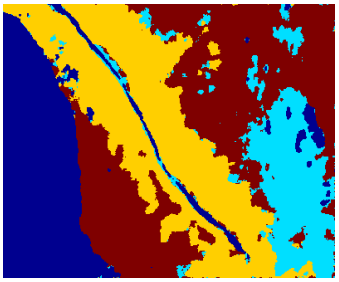
(2) Ground truth.



(3) ML classification.



(4) SegSALSA classification, ($\lambda_{TV} = 1$).



(5) SegSALSA classification, ($\lambda_{TV} = 2$).

(6) SegSALSA classification, ($\lambda_{TV} = 4$).

Figure 4. H&E stained sample of teratoma tissue imaged at 40X magnification containing the following classes: background (dark blue), fat (light blue), mesenchyme (dark red), and skin (yellow). Top row: (1) original image, (2) ground truth, (3) ML classification (56% accuracy). Bottom row: (4) SegSALSA classification with $\lambda_{TV} = 1$ (73% accuracy), (4) SegSALSA classification with $\lambda_{TV} = 2$ (81% accuracy), (6) SegSALSA classification with $\lambda_{TV} = 4$ (84% accuracy).

We run the SegSALSA algorithm for three training sets of different dimensions (20, 40, 200, and 500 samples per class randomly selected). The accuracy is computed from 10 Monte Carlo runs. For 20 samples per class we obtain an accuracy of $89.33\% \pm 3.53$; for 40 samples per class we obtain an accuracy of $92.30\% \pm 1.90$; for 200 samples per class we obtain an accuracy of $97.54\% \pm 0.50$; for 500 samples per class we obtain an accuracy of $98.50\% \pm 0.28$. The value of accuracy $= 97.54\% \pm 0.50\%$ obtained with 200 samples per class is considered state-of-the-art.[2]

## 5. CONCLUDING REMARKS

In this paper, we introduce a new approach to supervised image segmentation that avoids the discrete nature of problem present in many formulations. This is achieved by leveraging on the "hidden Markov measure field" introduced by Marroquin et al. in 2003. The proposed approach relies on four main ingredients: (a) formulating the image segmentation in the Bayesian framework; (b) introducing a hidden set of real-valued random fields determining the probability of a given partition; (c) adopting an form of isotropic vector total variation; and (d) introducing the Segmentation via the Constrained Split Augmented Lagrangian Shrinkage Algorithm (SegSALSA) to effectively solve the convex program which constitutes the marginal MAP inference of the hidden field. The effectiveness of the proposed methodology is illustrated with simulated and real hyperspectral and medical images. In addition, we provide a discussion on how to extend the proposed methodology to unsupervised and semi-supervised scenarios.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge support from the Fundação para a Ciência e Tecnologia, project PEst-OE/EEI/0008/2013, from Fundação para a Ciência e Tecnologia and the CMU-Portugal (ICTI) program, grant

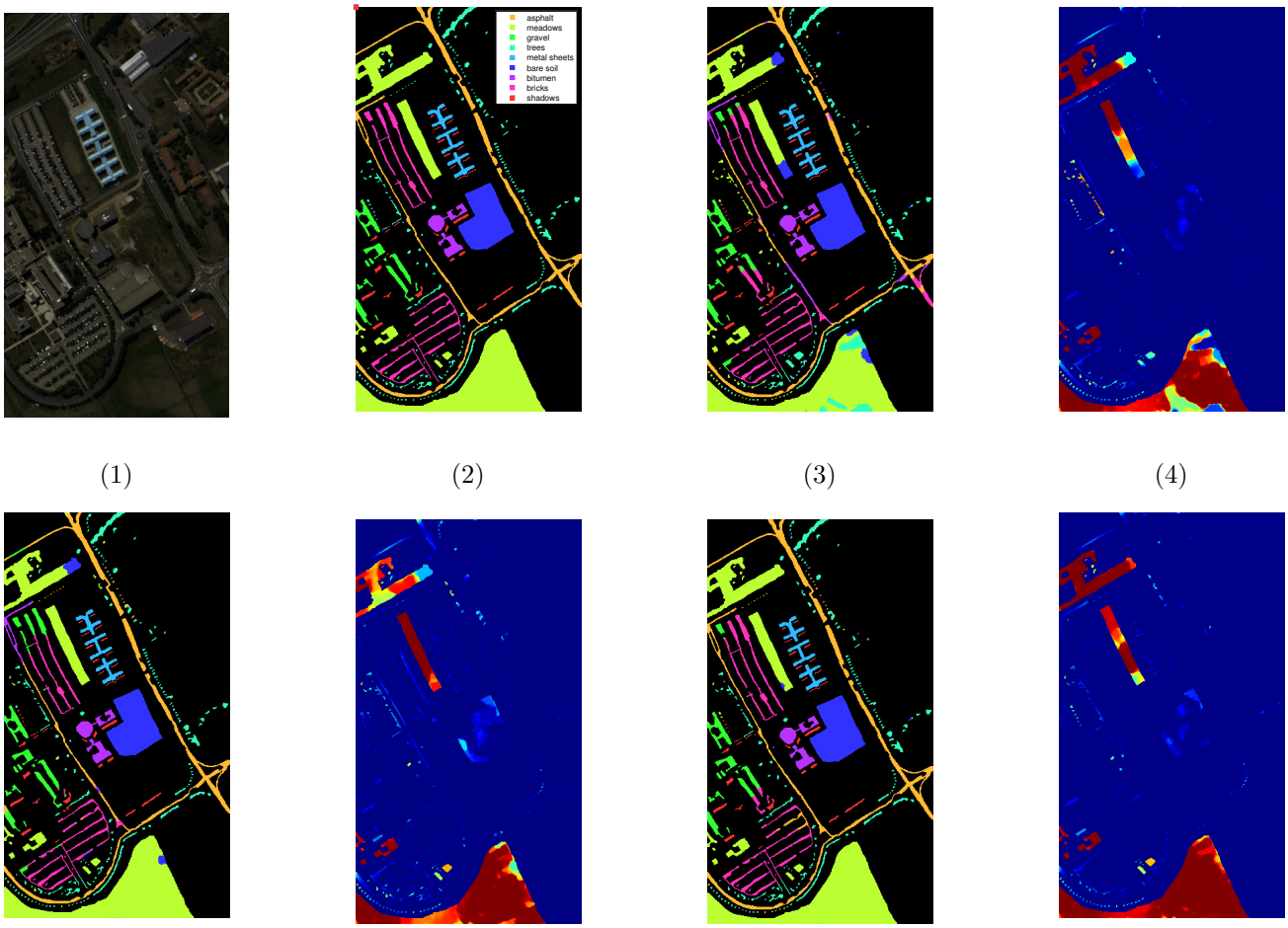



Figure 5. Classification of the ROSIS Pavia scene with varying dimension of the training set. Top row: (1) False color composition of the ROSIS Pavia scene, (2) ground truth containing the 9 mutually exclusive land-cover classes, (3) classification for 20 training samples per class (89.3% accuracy), (4) latent probabilities for "meadow" class for 20 training samples per class. Bottom row: (5) Classification for 40 training samples per class (92.3% accuracy), (6) latent probabilities for "meadow" class for 40 training samples per class, (7) classification for 200 training samples per class (98.5% accuracy), (8) latent probabilities for "meadow" class for 200 training samples per class.

SFRH/BD/51632/2011, and from and NSF through award 1017278 and the CMU CIT Infrastructure Award.

The authors would like to thank Prof. P. Gamba for providing the ROSIS Pavia scene, along with the training and test sets, Dr. C. Castro and Dr. J. Ozolek for providing the images of H&E-stained samples of teratoma tissues along with the ground truth.

## REFERENCES

1. D. Pham, C. Xu, and J. Prince, "Current methods in medical image segmentation 1," *Annual Review of Biomedical Engineering* **2**(1), pp. 315–337, 2000.
2. J. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *Geoscience and Remote Sensing Magazine, IEEE* **1**(2), pp. 6–36, 2013.
3. C. Brekke and A. Solberg, "Oil spill detection by satellite remote sensing," *Remote Sensing of Environment* **95**(1), pp. 1–13, 2005.

4. L. Hong and G. Chen, "Segment-based stereo matching using graph cuts," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, **1**, pp. I–74, IEEE, 2004.
5. Y. Boykov and G. Funka-Lea, "Graph cuts and efficient ND image segmentation," *International Journal of Computer Vision* **70**(2), pp. 109–131, 2006.
6. T. Randen and J. Husoy, "Filtering for texture classification: A comparative study," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **21**(4), pp. 291–310, 1999.
7. S. Li, *Markov random field modeling in computer vision*, Springer-Verlag New York, Inc., 1995.
8. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (6), pp. 721–741, 1984.
9. L. Vese and T. Chan, "A multiphase level set framework for image segmentation using the Mumford and Shah model," *International Journal of Computer Vision* **50**(3), pp. 271–293, 2002.
10. T. Chan, S. Esedoglu, and M. Nikolova, "Algorithms for finding global minimizers of image segmentation and denoising models," *SIAM Journal on Applied Mathematics* **66**(5), pp. 1632–1648, 2006.
11. J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**(8), pp. 888–905, 2000.
12. Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **23**(11), pp. 1222–1239, 2001.
13. V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **26**(2), pp. 147–159, 2004.
14. C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer, "Optimizing binary MRFs via extended roof duality," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.
15. M. Kumar, V. Kolmogorov, and P. Torr, "An analysis of convex relaxations for MAP estimation of discrete MRFs," *The Journal of Machine Learning Research* **10**, pp. 71–106, 2009.
16. N. Komodakis, N. Paragios, and G. Tziritas, "MRF energy minimization and beyond via dual decomposition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **33**(3), pp. 531–552, 2011.
17. A. Martins, M. Figueiredo, P. Aguiar, N. Smith, and E. Xing, "An augmented Lagrangian approach to constrained MAP inference," in *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*, pp. 169–176, 2011.
18. J. L. Marroquin, E. A. Santana, and S. Botello, "Hidden Markov measure field models for image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **25**(11), pp. 1380–1387, 2003.
19. X. Bresson and T. Chan, "Fast dual minimization of the vectorial total variation norm and applications to color image processing," *Inverse Problems and Imaging* **2**(4), pp. 455–484, 2008.
20. B. Goldluecke, E. Strekalovskiy, and D. Cremers, "The natural vectorial total variation which arises from geometric measure theory," *SIAM Journal on Imaging Sciences* **5**(2), pp. 537–563, 2012.
21. M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *Image Processing, IEEE Transactions on* **20**(3), pp. 681–695, 2011.
22. A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)* , pp. 1–38, 1977.
23. M. A. Figueiredo, "Bayesian image segmentation using wavelet-based priors," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, **1**, pp. 437–443, IEEE, 2005.
24. J. Lellmann, J. Kappes, J. Yuan, F. Becker, and C. Schnörr, "Convex multi-class image labeling by simplex-constrained total variation," in *Scale Space and Variational Methods in Computer Vision*, X.-C. Tai, K. Morken, M. Lysaker, and K.-A. Lie, eds., *Lecture Notes in Computer Science* **5567**, pp. 150–162, Springer Berlin Heidelberg, 2009.
25. L. Sun, Z. Wu1, J. Liu, and Z. Wei, "Supervised hyperspectral image classification using sparse logistic regression and spatial-tv regularization," in *IEEE Geoscience and Remote Sensing Symposium (IGARSS'13), Melbourne, Australia, 2013*, 2013.

26. J. Besag, "On the statistical analysis of dirty images," *Journal of Royal Statistics Society* **48**, pp. 259–302, 1986.
27. A. Ng and M. Jordan, "On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes," *Advances in Neural Information Processing Systems* **14**, p. 841, 2002.
28. D. Greig, B. Porteous, and A. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society. Series B (Methodological)* , pp. 271–279, 1989.
29. V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **28**(10), pp. 1568–1583, 2006.
30. P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision* **70**(1), pp. 41–54, 2006.
31. R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov random fields with smoothness-based priors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **30**(6), pp. 1068–1080, 2008.
32. P. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212, Springer, 2011.
33. D. B. J. Eckstein, "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming* **55**(1-3), pp. 293–318, 1992.
34. D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications* **2**(1), pp. 17–40, 1976.
35. R. Glowinski and A. Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires," *Revue Française dAutomatique, Informatique et Recherche Operationnelle* **9**, pp. 41–76, 1975.
36. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning* **3**(1), pp. 1–122, 2011.
37. J. Li, J. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new bayesian approach with active learning," *Geoscience and Remote Sensing Magazine, IEEE* **49**, pp. 3947–3960, Oct. 2011.
38. F. Condessa, J. Bioucas-Dias, C. Castro, J. Ozolek, and J. Kovačević, "Classification with rejection option using contextual information," in *Proceedings of the 2013 IEEE International Symposium on Biomedical Imaging*, (San Francisco), Apr. 2013.