Matrix cofactorization for joint unmixing and classification of hyperspectral images

Adrien Lagrange*, Mathieu Fauvel†, Stéphane May‡, José M. Bioucas-Dias◇ and Nicolas Dobigeon*
* IRIT/INP-ENSEEIHT, University of Toulouse, Toulouse, France
† Centre d'Études Spatiales de la BIOsphère (CESBIO), INRA, Toulouse, France
‡ Centre National d'Études Spatiales (CNES), DCT/SI/AP, Toulouse, France
◇ Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal
firstname.name@{enseeiht,inra,cnes,enseeiht}.fr, bioucas@lx.it.pt

***Abstract*—This paper introduces a matrix cofactorization approach to perform spectral unmixing and classification jointly. After formulating the unmixing and classification tasks as matrix factorization problems, a link is introduced between the two coding matrices, namely the abundance matrix and the feature matrix. This coupling term can be interpreted as a clustering term where the abundance vectors are clustered and the resulting attribution vectors are then used as feature vectors. The overall non-smooth, non-convex optimization problem is solved using a proximal alternating linearized minimization algorithm (PALM) ensuring convergence to a critical point. The quality of the obtained results is finally assessed by comparison to other conventional algorithms on semi-synthetic yet realistic dataset.**

***Index Terms*—supervised learning, spectral unmixing, cofactorization, hyperspectral images.**

I. Introduction

Following the fast increase of available remote sensing images, many methods have been proposed to extract information from such specific data. In particular classification algorithms received a lot of attention from the scientific community. The emergence of state-of-the-art algorithms such as convolutional neural network [1] or random forest [2] have brought unprecedented good results. In the so-called supervised classification framework, these algorithms make it possible to infer, from a reduced number of examples provided by an expert, a classification rule. This rule is then used to attribute to unknown pixels a class among a predefined set of classes. Although very efficient, classification methods remain a limited analysis of the image since it only attributes a single class to each pixel when it is sometimes possible to extract more information. In the specific case of hyperspectral images (HSI), images capture a very rich signal since each pixel is a sampling of the reflectance spectrum of the corresponding area, typically in the visible and infrared spectral domains with hundreds of measurements. To fully exploit the available information, it is interesting to resort to alternative methods of interpretation such as representation learning methods, namely spectral unmixing in the case of HSI [3]. Spectral unmixing is a physic-based model which assumes that a given pixel, i.e. a given measured spectrum, is the result of the combination of a reduced number of elementary spectra called endmembers, specific to a given material. The aim of unmixing methods is to infer the proportion of each material present in the pixel. The obtained abundance maps display the spatial distribution of the material in the observed scene.

Even if classification and spectral unmixing are two widely-used techniques, very few attempts have been made to combine them. Most of these works [4], [5] intend to improve classification results by using spectral unmixing to identify mixed pixels and then process specifically the identified mixed pixels. Instead of using the two methods sequentially, the method proposed in this paper introduces the idea of a joint unmixing and classification. This method is formulated as a cofactorization problem, which is known to produce valuable results in many application fields such as music source separation [6], and image analysis [7]. The core concept is to express the two problems of interest, namely spectral unmixing and classification, as factorization problems and then to introduce a coupling term to intertwine the two estimations. Similarly to [8], the coupling term is defined as a clustering term where the abundance vectors provided by the unmixing step are clustered and the resulting attribution vectors are then used as feature vectors for the classification. The overall optimization problem is non-convex non-smooth. Such problems are known to be challenging to solve but, building on recent advances in optimization, the PALM algorithm proposed in [9] is used as an optimization scheme, thus guaranteeing convergence to a critical point of the objective function.

The rest of this paper is organized as follows. Section II defines the two factorization tasks and introduces the global cofactorization problem. Then, the method used to minimize the resulting criterion is presented in Section III. Finally, the method is tested and compared to other unmixing and classification methods in Section IV. Section V draws some conclusions and perspectives.

II. Problem statement

As presented in Sections II-A and II-B, spectral unmixing and supervised classification are commonly expressed as factorization problems. We propose to derive a unified framework

Part of this work has been supported Centre National d'Études Spatiales (CNES), Occitanie Region, EU FP7 through the ERANETMED JC-WATER program, MapInvPlnt Project ANR-15-NMED-0002-02 and ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI).

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

by considering a global cofactorization problem. It relies on a link between the two factorization problems in order to perform a joint estimation. In the proposed model, the link is made between the abundance matrix and the feature matrix. More precisely, the coupling term is expressed as a clustering term over the abundance vectors where the attribution vectors to the clusters are also the feature vectors of the classification as detailed in Section II-C.

A. Spectral unmixing

Each pixel of an HSI is a L-dimensional measurement of a reflectance spectrum. Physics models this spectrum as a combination of R elementary spectrum, gathered in the so-called endmember matrix $\mathbf{M} \in \mathbb{R}^{L\times R}$, each characterizing a specific material. The spectral unmixing task aims at retrieving the so-called abundance vectors $\mathbf{a}_p \in \mathbb{R}^R$, with $R \ll L$, from the spectrum $\mathbf{y}_p \in \mathbb{R}^L$ of the pth pixel ($p \in \mathcal{P}$ where $\mathcal{P} \triangleq \{1,\ldots,P\}$ is the set of pixel indexes). These abundance vectors describe the mixture contained in the pixel. Using the conventional linear mixture model, the spectral unmixing problem can be expressed as follow

$$\min_{\mathbf{M},\mathbf{A}} \frac{1}{2}\left\|\mathbf{Y}-\mathbf{M}\mathbf{A}\right\|_{\mathrm{F}}^2 + \lambda_a \left\|\mathbf{A}\right\|_1 + \imath_{\mathbb{R}_+^{R\times P}}(\mathbf{A}) \tag{1}$$

where matrix $\mathbf{Y} \in \mathbb{R}^{L\times P}$ gathers the P pixel spectra and $\mathbf{A} \in \mathbb{R}^{R\times P}$ the abundance vectors. In addition to the data fitting term, two penalization terms are considered in the proposed unmixing model. The term $\imath_{\mathbb{R}_+^{R\times P}}(\mathbf{A})$ enforces a nonnegativity constraint, ensuring an additive decomposition of the spectra. The second penalization $\lambda_a \left\|\mathbf{A}\right\|_1$ is a sparsity penalization promoting the concept that only a few endmembers are active in a given pixel. In the following work, the choice has been made to discard the estimation of the endmember matrix for the sake of simplicity. The endmember matrix is assumed to be known or estimated beforehand.

B. Classification

In the context of supervised classification, a subset of pixels is available with their corresponding groundtruth. The index subset of labeled pixel is denoted hereafter $\mathcal{L}$ while the index subset of unlabeled pixel is $\mathcal{U}$ ($\mathcal{L} \cap \mathcal{U} = \emptyset$ and $\mathcal{L} \cup \mathcal{U} = \mathcal{P}$). Classification intends to assign one of the C classes to each pixel. In practice, classifying can be formulated as estimating a $C \times P$ matrix $\mathbf{C}$ whose columns correspond to unknown C-dimensional attribution vectors $\mathbf{c}_p$ ($p \in \mathcal{U}$). Each vector is made of 0 except for $c_{i,p} = 1$ when the pth pixel is assigned the ith class. Numerous decision rules have been proposed to carry out classification. Most of them rely on the use of feature vectors $\mathbf{z}_p \in \mathbb{R}^K$ ($p \in \mathcal{P}$) associated with the P pixels, gathered in the matrix $\mathbf{Z} \in \mathbb{R}^{K\times P}$. Considering a linear classifier parametrized by the matrix $\mathbf{Q} \in \mathbb{R}^{C\times K}$, a vector-wise nonlinear mapping $\phi(\cdot)$, such as a sigmoid or a softmax operator, is then applied to the output of the classifier. Finally the classification rule can be expressed as the matrix factorization problem

$$\min_{\mathbf{Q},\mathbf{C}_{\mathcal{U}}} \mathcal{J}_{\mathrm{c}}(\mathbf{C},\phi(\mathbf{QZ})) + \imath_{\mathbb{S}_C^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}}) \tag{2}$$

where $\mathcal{J}_{\mathrm{c}}(\cdot,\cdot)$ is a cost function measuring the quality of the estimated attribution vectors $\phi(\mathbf{Q}\mathbf{z}_p)$ and and $\mathbb{S}_C$ is the C-dimensional probability simplex ensuring nonnegativity and sum-to-one constraints of the attribution vectors. In this work, the cost function $\mathcal{J}_{\mathrm{c}}(\cdot,\cdot)$ has been chosen as the cross-entropy, defined in a multi-class problem as

$$\mathcal{J}_{\mathrm{c}}(\mathbf{C},\hat{\mathbf{C}}) = -\sum_{p\in\mathcal{P}} d_p \sum_{i\in\mathcal{C}} c_{i,p} \log\left(\hat{c}_{i,p}\right) \tag{3}$$

with

$$d_p = \begin{cases} \frac{1}{|\mathcal{L}_i|}, & \text{if } p \in \mathcal{L}_i, \\ \frac{1}{|\mathcal{U}|}, & \text{if } p \in \mathcal{U}, \end{cases} \tag{4}$$

where $\mathcal{L}_i$ is the subset of labeled pixels belonging to class i, $\hat{\mathbf{c}}_p$ is the estimated attribution vector and $\mathbf{c}_p$ the true one. The weighing coefficients d_p adjust the cost function with respect to the sizes of the training and test sets, in particular in the case of unbalanced classes. This particular loss function has been extensively used in the context of neural networks [10]. Moreover, the nonlinear mapping $\phi(\cdot)$ is chosen as a sigmoid, which makes the proposed classifier interpretable as a one layer neural network.

To consider a more elaborate case, it is also possible to add a set of penalizations/constraints. In particular, a penalization of the classifier parameters $\mathbf{Q}$ is considered to prevent an artificial decrease of the loss function. This penalization is based on a Frobenius-norm and is well-known in the neural network community where it is referred to as *weight decay*. The second considered penalization is a spatial regularization enforced through a smoothed weighted vectorial total variation norm (vTV). This regularization promotes a piece-wise constant solution for the classification map $\mathbf{C}$. The overall resulting problem can be written

$$\begin{aligned} \min_{\mathbf{Q},\mathbf{C}_{\mathcal{U}}} & -\sum_{p\in\mathcal{P}} d_p \sum_{i\in\mathcal{C}} c_{i,p} \log\left(\frac{1}{1+\exp(-\mathbf{q}_{i:}\mathbf{z}_p)}\right) \\ & + \lambda_q \left\|\mathbf{Q}\right\|_F^2 + \lambda_c \left\|\mathbf{C}\right\|_{\mathrm{vTV}} + \imath_{\mathbb{S}_C^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}}) \end{aligned} \tag{5}$$

where $\mathbf{q}_{i:}$ is the i-th line of $\mathbf{Q}$, λ_q and λ_c weight the regularization terms and

$$\left\|\mathbf{C}\right\|_{\mathrm{vTV}} = \sum_{p=1}^{P} \beta_p \sqrt{\left\|\left[\nabla_{\mathrm{h}}\mathbf{C}\right]_p\right\|_2 + \left\|\left[\nabla_{\mathrm{v}}\mathbf{C}\right]_p\right\|_2 + \epsilon} \tag{6}$$

where $\epsilon > 0$ is a smoothing parameter and $\left[\nabla_{\mathrm{h}}(\cdot)\right]_p$ and $\left[\nabla_{\mathrm{v}}(\cdot)\right]_p$ denote horizontal and vertical discrete gradients[1]

$$\begin{aligned} \left[\nabla_{\mathrm{h}}\mathbf{C}\right]_{(m,n)} &= \mathbf{c}_{(m+1,n)} - \mathbf{c}_{(m,n)} \\ \left[\nabla_{\mathrm{v}}\mathbf{C}\right]_{(m,n)} &= \mathbf{c}_{(m,n+1)} - \mathbf{c}_{(m,n)}. \end{aligned}$$

The weighting coefficients $\beta_{m,n}$ are introduced to account for the natural boundaries present in the image. They are computed beforehand using external data containing information on the spatial structures, e.g., a panchromatic image or a LIDAR image [11]. An example of such weights is described in Section IV.

[1] With a slight abuse of notations, $\mathbf{c}_{(m,n)}$ refers to the pth column of $\mathbf{C}$ where the pth pixel is spatially indexed by (m,n).

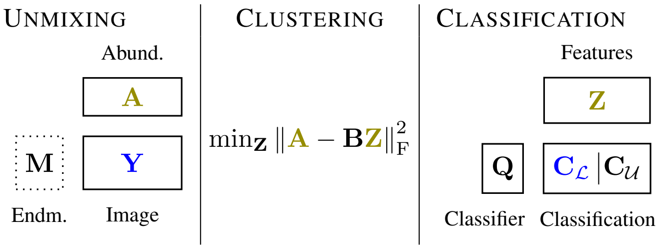

UNMIXING
Abund.
A
M
Y
Endm.
Image
CLUSTERING
$\min_{\mathbf{Z}} \|\mathbf{A} - \mathbf{B}\mathbf{Z}\|_{\mathrm{F}}^2$
CLASSIFICATION
Features
Z
Q
$\mathbf{C}_{\mathcal{L}} | \mathbf{C}_{\mathcal{U}}$
Classifier
Classification

Fig. 1. Structure of the cofactorization model. Variables in *blue* stand for observations or available external data. Variables in *olive green* are linked through the clustering term. The variable in a dotted box is assumed to be known beforehand.

C. Clustering

To define a global cofactorization problem, a relation is drawn between the activation matrices of the two factorization problems, namely the abundance matrix and the feature matrix. More specifically, following the idea developed in [8], a clustering term is introduced as a coupling. Abundances vectors are clustered and the resulting attribution vectors are then used as feature vectors for the classification. Ideally, clustering attribution vectors $\mathbf{z}_p \in \mathbb{R}^K$ are filled with zeros except for $z_{k,p} = 1$ when $\mathbf{a}_p$ is associated with the kth cluster. The well-known k-means is chosen to perform this task since it is easily expressed as an optimization problem

$$\min_{\mathbf{Z},\mathbf{B}} \frac{1}{2} \|\mathbf{A} - \mathbf{B}\mathbf{Z}\|_{\mathrm{F}}^2 + \imath_{\mathbb{S}_K^P}(\mathbf{Z}) + \imath_{\mathbb{R}_+^{R\times K}}(\mathbf{B}) \tag{7}$$

where columns of $\mathbf{B} \in \mathbb{R}^{R\times K}$ stands for the centroids of the K clusters. Two constraints are considered in this k-means clustering problem: *i)* a positivity constraint on $\mathbf{B}$ since centroids are expected to be interpretable as mean abundance vectors and *ii)* the vectors $\mathbf{z}_p$ ($p \in \mathcal{P}$) are assumed to be defined on the K-dimensional probability simplex $\mathbb{S}_K$. Thus, the resulting clustering method is a particular instance of k-means where the attribution vectors are relaxed and can be interpreted as the collection of probabilities to belong to each of the clusters.

D. Multi-objective problem

The two factorization problems corresponding to the spectral unmixing and classification tasks have been expressed and the link between these two problems has been set up through the clustering term. The global cofactorization problem, illustrated in Figure 1, is finally formulated as

$$\begin{aligned}\min_{\substack{\mathbf{A},\mathbf{Q},\mathbf{Z}\\ \mathbf{C}_{\mathcal{U}},\mathbf{B}}} & \frac{\lambda_0}{2} \|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda_a \|\mathbf{A}\|_1 + \imath_{\mathbb{R}_+^{R\times P}}(\mathbf{A}) \\ & - \frac{\lambda_1}{2} \sum_{p\in\mathcal{P}} d_p \sum_{i\in\mathcal{C}} c_{i,p} \log\left(\frac{1}{1+\exp(-\mathbf{q}_{i:}\mathbf{z}_p)}\right) \\ & + \frac{\lambda_q}{2} \|\mathbf{Q}\|_{\mathrm{F}}^2 + \lambda_c \|\mathbf{C}\|_{\mathrm{vTV}} + \imath_{\mathbb{S}_C^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}}) \\ & + \frac{\lambda_2}{2} \|\mathbf{A} - \mathbf{B}\mathbf{Z}\|_F^2 + \imath_{\mathbb{S}_K^P}(\mathbf{Z}) + \imath_{\mathbb{R}_+^{R\times K}}(\mathbf{B})\end{aligned} \tag{8}$$

where λ_0, λ_1 and λ_2 are introduced to weight the contribution of the various terms.

III. Optimization Scheme

The proposed global optimization problem (8) is non-convex and non-smooth. Such problem are usually very challenging to solve. To handle it, we propose to resort to the PALM algorithm proposed in [9]. PALM algorithm ensures convergence to a critical point, i.e., a local minimum of the objective function. To apply PALM, the objective is rewritten as a sum of independent non-smooth terms $f_j(\cdot)$ ($j \in \{1,\ldots,3\}$) and a smooth coupling term $g(\cdot)$

$$\min_{\substack{\mathbf{A},\mathbf{B},\mathbf{Z},\\ \mathbf{Q},\mathbf{C}_{\mathcal{U}}}} f_0(\mathbf{A}) + f_1(\mathbf{B}) + f_2(\mathbf{Z}) + f_3(\mathbf{C}_{\mathcal{U}}) + g(\mathbf{A},\mathbf{B},\mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{Q})$$

where

$$f_0(\mathbf{A}) = \imath_{\mathbb{R}_+}(\mathbf{A}) + \lambda_a \|\mathbf{A}\|_1, \quad f_1(\mathbf{B}) = \imath_{\mathbb{R}_+}(\mathbf{B})$$

$$f_2(\mathbf{Z}) = \imath_{\mathbb{S}_K^P}(\mathbf{Z}), \quad f_3(\mathbf{C}_{\mathcal{U}}) = \imath_{\mathbb{S}_K^{|\mathcal{U}|}}(\mathbf{C}_{\mathcal{U}})$$

$$\begin{aligned}g(\mathbf{A},\mathbf{B},\mathbf{Z},\mathbf{C}_{\mathcal{U}},\mathbf{Q}) &= \frac{\lambda_0}{2} \|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_{\mathrm{F}}^2 \\ & - \frac{\lambda_1}{2} \sum_{p\in\mathcal{P}} d_p \sum_{i\in\mathcal{C}} c_{i,p} \log\left(\frac{1}{1+\exp(-\mathbf{q}_{i:}\mathbf{z}_p)}\right) \\ & + \frac{\lambda_2}{2} \|\mathbf{A} - \mathbf{B}\mathbf{Z}\|_{\mathrm{F}}^2 + \frac{\lambda_q}{2} \|\mathbf{Q}\|_{\mathrm{F}}^2 + \lambda_c \|\mathbf{C}\|_{\mathrm{vTV}}.\end{aligned}$$

Algorithm 1: PALM

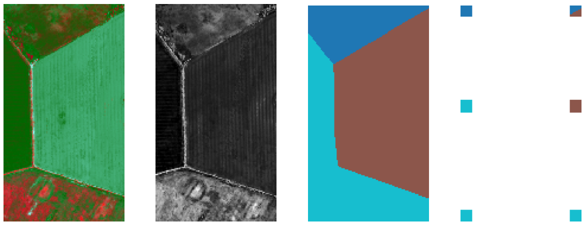
```
Initialize variables $\mathbf{A}^0$, $\mathbf{B}^0$, $\mathbf{Z}^0$, ${\mathbf{C}_{\mathcal{U}}}^0$ and $\mathbf{Q}^0$;
Set $\alpha > 1$;
while *stopping criterion not reached* do
    $\mathbf{A}^{k+1} \in \mathrm{prox}_{f_0}^{\alpha L_{\mathbf{A}}}(\mathbf{A}^k - \frac{1}{\alpha L_{\mathbf{A}}} \nabla_{\mathbf{A}} g(\mathbf{A}^k, \mathbf{B}^k, \mathbf{Z}^k, \mathbf{C}_{\mathcal{U}}^k, \mathbf{Q}^k));$
    $\mathbf{B}^{k+1} \in$
      $\mathrm{prox}_{f_1}^{\alpha L_{\mathbf{B}}}(\mathbf{B}^k - \frac{1}{\alpha L_{\mathbf{B}}} \nabla_{\mathbf{B}} g(\mathbf{A}^{k+1}, \mathbf{B}^k, \mathbf{Z}^k, \mathbf{C}_{\mathcal{U}}^k, \mathbf{Q}^k));$
    $\mathbf{Z}^{k+1} \in$
      $\mathrm{prox}_{f_2}^{\alpha L_{\mathbf{Z}}}(\mathbf{Z}^k - \frac{1}{\alpha L_{\mathbf{Z}}} \nabla_{\mathbf{Z}} g(\mathbf{A}^{k+1}, \mathbf{B}^{k+1}, \mathbf{Z}^k, \mathbf{C}_{\mathcal{U}}^k, \mathbf{Q}^k));$
    $\mathbf{Q}^{k+1} \in \mathrm{prox}_{f_3}^{\alpha L_{\mathbf{Q}}}(\mathbf{Q}^k -$
      $\frac{1}{\alpha L_{\mathbf{Q}}} \nabla_{\mathbf{Q}} g(\mathbf{A}^{k+1}, \mathbf{B}^{k+1}, \mathbf{Z}^{k+1}, {\mathbf{C}_{\mathcal{U}}}^k, \mathbf{Q}^k));$
    $\mathbf{C}_{\mathcal{U}}^{k+1} \in \mathrm{prox}_{f_4}^{\alpha L_{\mathbf{C}_{\mathcal{U}}}}(\mathbf{C}_{\mathcal{U}}^k -$
      $\frac{1}{\alpha L_{\mathbf{C}_{\mathcal{U}}}} \nabla_{\mathbf{C}_{\mathcal{U}}} g(\mathbf{A}^{k+1}, \mathbf{B}^{k+1}, \mathbf{Z}^{k+1}, \mathbf{C}_{\mathcal{U}}^k, \mathbf{Q}^{k+1}));$
end
return $\mathbf{A}^{end}, \mathbf{B}^{end}, \mathbf{Z}^{end}, \mathbf{Q}^{end}, \mathbf{C}_{\mathcal{U}}^{end}$
```

The concept of this algorithm is to perform a proximal gradient descent according to each variable alternatively. To apply PALM, the functions $f_j(\cdot)$ have to be proper, lower semi-continuous, extended real-valued. A sufficient condition on the function $g(\cdot)$ is to be $\mathcal{C}^2$, i.e., with continuous first and second derivatives, and its partial gradients have to be globally Lipschitz. $L_{\mathbf{X}}$ denotes herein the Lipschitz constant associated to the partial gradient according to $\mathbf{X}$. The detailed steps of the algorithm are summarized in Algorithm 1 and further theoretical details are available in [9].

In practice, one needs to be able to compute the partial gradient and its associated Lipschitz constant to perform the gradient descent. It is also necessary to compute the proximal operator associated to the non-smooth terms. In the present case, the partial gradients is easily computed and all globally Lipschitz. The only problematic term is the vTV term which is not globally Lipschitz in its canonical form. To alleviate, a smoothed counterpart has been introduced in (6) with a smoothing parameter $\epsilon \in \mathbb{R}_+$. As for the proximal operators, they are are well-known [12] except for $f_0(\cdot)$. For $f_0(\cdot)$, it is necessary to resort to the composition of the proximal operators associated to the non-negative constraint and the ℓ_1-norm, which is here possible according to [13].

IV. Experiments

Data generation – The HSI used to perform the experiments is a semi-synthetic image. More specifically, the image has been generated using a real HSI. The real image has been unmixed using a fully constrained least square (FCLS) algorithm [14] using $R = 5$ endmembers extracted with the well-known VCA algorithm [15]. The obtained abundance maps have then been used to generate a new synthetic image using pure spectra from the hyperspectral library ASTER [16]. The groundtruth of the original data, composed of $C = 3$ classes has been preserved to assess the quality of the classification. A color composition, a panchromatic version and the groundtruth are presented in Figure 2. The subset of the image used as training data is as also shown in Figure 2.



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

Fig. 2. Synthetic image: (a) colored composition of the HSI $\mathbf{Y}$, (b) panchromatic image $\mathbf{y}_{\mathrm{PAN}}$, (c) classification ground-truth, (d) training set.

Initialization and convergence – As stated before, cofactorization is a non-convex problem and PALM only ensures convergence to a local minimum of the objective function. It is thus important to carefully initialize the estimated variables in order to reach a relevant solution. In the presented experiment, abundance matrix $\mathbf{A}^0$ has been initialized by solving $\min_{\mathbf{A}\in\mathbb{R}_+^{R\times P}} \|\mathbf{Y} - \mathbf{MA}\|_{\mathrm{F}}^2$ using a projected gradient algorithm. Then, a k-means algorithm has been applied to the obtained abundance vectors and the resulting centroids and attribution vectors have been used to initialize $\mathbf{B}^0$ and $\mathbf{Z}^0$. On the other hand, classifier parameters $\mathbf{Q}^0$ and classification matrix $\mathbf{C}_{\mathcal{U}}^0$ have been initialized randomly.

In order to assess the convergence of the optimization scheme, the normalized difference between two consecutive values of the objective function is monitored. When this value reach a certain threshold (10^{-4} for this experiment), the optimization process stops and the last estimation is assumed to be close enough to the solution.

Hyperparameters – Multiple hyperparameters $\lambda_.$ have been introduced in problem (8) to weight the various terms of the objective function. For practical use, these parameters have been normalized by the size and dynamics of the corresponding variables. These normalized parameters, denoted $\tilde{\lambda}_.$, have been empirically tuned to obtain consistent results ($\tilde{\lambda}_0 = \tilde{\lambda}_1 = \tilde{\lambda}_2 = 1$, $\tilde{\lambda}_a = 10^{-3}$, $\tilde{\lambda}_q = 0.15$). For the last hyperparameter $\tilde{\lambda}_c$, two values have been considered $0.$ and 0.1, standing respectively for the case without and with spatial regularization. The definition of the vTV regularization also includes parameters which has to be properly set. First, the smoothing parameter is set to $\epsilon = 0.01$ to ensure the gradient-Lipschitz property without modifying substantially the TV-norm. Secondly, it is necessary to define the weighing coefficients $\beta_{m,n}$. They have been computed from a panchromatic image $\mathbf{y}_{\mathrm{PAN}}$, shown in Figure 2, generated by normalizing hyperspectral bands by their mean and then summing them. More precisely, to account for possible homogeneous areas in the image, they are defined as follows

$$\beta_p = \frac{\tilde{\beta}_p}{\sum_q \tilde{\beta}_q} \text{ with } \tilde{\beta}_q = \left(\left\|[\nabla \mathbf{y}_{\mathrm{PAN}}]_q\right\|_2 + \sigma\right)^{-1}$$

where $\sigma = 0.01$ controls the variation of the weights and avoids numerical issues.

TABLE I
Unmixing and classification results.

Model	Kappa	F1-mean	RMSE($\hat{\mathbf{A}}$)	RE	Time (s)
RF	0.817	0.842	N\A	N\A	0.4
FCLS	N\A	N\A	0.0701	0.224	1.2
CBPDN	N\A	N\A	0.0792	0.229	2
D-KSVD	0.494	0.554	N\A	0.923	70
Cofact.	0.847	0.870	0.0504	0.750	180
Cofact. + vTV	0.874	0.895	0.0526	0.752	81

Compared methods – To assess the quality of the unmixing and classification results, the proposed method has been compared to several well-known unmixing and classification algorithms. Regarding classification, we considered the random forest (RF) algorithm, known to perform very well to classify HSI. Parameters of the RF (number of trees, depth) have been adjusted using gridsearch and cross-validation. The discriminative K-SVD (D-KSVD) method has been used as a benchmark [17]. This model is also a cofactorization method but with a simpler approach where the two coding matrices $\mathbf{A}$ and $\mathbf{Z}$ are imposed to be equal. In this case, the first term is not a spectral unmixing task but rather a dictionary learning task where dictionary elements are assumed to be discriminative for the classification task. Only a sparsity penalization is considered for D-KSVD using a ℓ_0-norm.

As for the unmixing comparison, we considered two methods described in [14]. The first method is the fully constrained least square method (FCLS) where the corresponding optimization problem is defined as the data fitting term with a positivity and sum-to-one constraint on abundance vectors $\mathbf{a}_p$. The second method is the constrained basis pursuit denoising (CBPDN) corresponding to problem 1. The hyperparameter λ_a, weighting the sparsity penalty is also adjusted using gridsearch and cross-validation. It should be noted that all unmixing methods use directly the correct endmember matrix $\mathbf{M}$ which has been used to generate the data. Additionally, the endmember matrix is used to initialize the dictionary of the D-KSVD method.

Results – To evaluate the unmixing results quantitatively, the reconstruction error (RE) and root global mean squared error (RMSE) are considered, i.e.,

$$\mathrm{RE} = \sqrt{\frac{1}{PL}\left\|\mathbf{Y} - \mathbf{M}\hat{\mathbf{A}}\right\|_{\mathrm{F}}^{2}}, \mathrm{RMSE} = \sqrt{\frac{1}{PR}\left\|\mathbf{A}_{\mathrm{true}} - \hat{\mathbf{A}}\right\|_{\mathrm{F}}^{2}},$$

where $\mathbf{A}_{\mathrm{true}}$ and $\hat{\mathbf{A}}$ are the actual and estimated abundance matrices. To evaluate the classification accuracy, two conventional metrics are used, namely Cohen's kappa coefficient and the averaged F1-score over all classes [18]. The results have been averaged over 20 trials. A different random noise has been added to the image for each trial such that the $\mathrm{SNR} = 30\mathrm{dB}$.

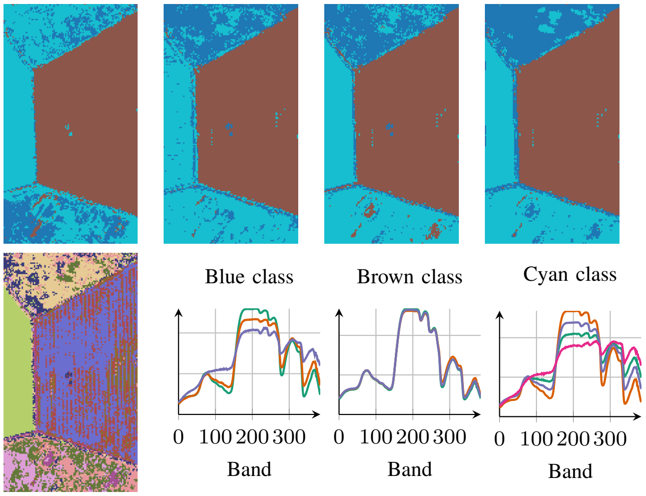


Fig. 3. Results: (1st row) classification maps for D-KSVD, RF, Cofact., Cofact. + vTV, (2nd row) cluster map and spectral centroids recovered by the proposed method (Cofact. + vTV).

Results reported in Table I show that the proposed cofactorization framework outperforms both RF and D-KSVD in term of classification. Regarding spectral unmixing, FCLS and CBPDN reach lower REs which is expected since both methods exhibit more degrees of freedom. Note however this metrics only evaluates the quality of the reconstructed data. However, RMSE is lower for the cofactorization methods than for FCLS and CBPDN. It is finally interesting to underline the improvement due to the vTV. As expected, classification results improve when the spatial regularization is considered but additionally the convergence time of the PALM algorithm is significantly reduced. Processing time is indeed higher for the proposed cofactorization method than for RF, FCLS and CBPDN. However, it simultaneously tackles two problems instead of one, and processing time still remains comparable to the one of D-KSVD method.

In terms of qualitative results, Figure 3 presents the classification maps which appear consistent with the quantitative results. Moreover, results of the clustering task are shown with spectral centroids of the clusters for each classes. In particular, this cluster maps exhibits some heterogeneity, which can be explained by the multi-modality of some classes. This additional result is a very interesting byproduct of the proposed method to interpret the scene.

V. Conclusion and perspective

This paper introduces a unified framework to perform jointly spectral unmixing and classification by the mean of a cofactorization problem. Coding matrices of two problems of interest are linked thanks to a clustering term. The overall cofactorization task is formulated as a non-convex nonsmooth optimization problem whose solution was approximated thanks to a PALM algorithm which ensured some convergence guarantees. Even if the proposed method appears to perform significantly well compared to other state-of-the-art methods, further improvements are yet to be considered. In particular, the additional learning of the endmember matrix could be considered. It would be also relevant to exploit the supervised information on the spectral unmixing part which is very rarely available in a conventional unmixing problem.

References

[1] Y. Chen, H. Jiang *et al.*, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, 2016.

[2] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[3] J. M. Bioucas-Dias, A. Plaza *et al.*, "Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 5, no. 2, pp. 354–379, April 2012.

[4] A. Villa, J. Chanussot *et al.*, "Spectral unmixing for the classification of hyperspectral images at a finer spatial resolution," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 3, pp. 521–533, 2011.

[5] I. Dópido, J. Li *et al.*, "A new hybrid strategy combining semisupervised classification and unmixing of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 7, no. 8, pp. 3619–3629, 2014.

[6] J. Yoo, M. Kim *et al.*, "Nonnegative matrix partial co-factorization for drum source separation," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference On*. IEEE, 2010, pp. 1942–1945.

[7] N. Yokoya, T. Yairi *et al.*, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, 2012.

[8] A. Lagrange, M. Fauvel *et al.*, "Hierarchical Bayesian image analysis: From low-level modeling to robust supervised learning," *Pattern Recognit.*, vol. 85, pp. 26–36, 2019.

[9] J. Bolte, S. Sabach *et al.*, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, Aug. 2014.

[10] I. Goodfellow, Y. Bengio *et al.*, *Deep Learning*. MIT press Cambridge, 2016, vol. 1.

[11] T. Uezato, M. Fauvel *et al.*, "Hyperspectral image unmixing with LiDAR data-aided spatial regularization," *IEEE Trans. Geosci. Remote Sens.*, 2018.

[12] L. Condat, "Fast projection onto the simplex and the l1 ball," *Math. Program.*, vol. 158, no. 1-2, pp. 575–585, 2016.

[13] Y.-L. Yu, "On decomposing the proximal map," in *Adv. Neural Inf. Process. Syst.*, 2013, pp. 91–99.

[14] J. M. Bioucas-Dias and M. A. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *Proc. IEEE Workshop Hyperspectral Image SIgnal Process.: Evolution in Remote Sens. (WHISPERS)*. IEEE, 2010, pp. 1–4.

[15] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, April 2005.

[16] A. M. Baldridge, S. J. Hook *et al.*, "The ASTER spectral library version 2.0," *Remote Sens. Environ.*, vol. 113, no. 4, pp. 711–715, 2009.

[17] Z. Jiang, Z. Lin *et al.*, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *CVPR 2011*, June 2011, pp. 1697–1704.

[18] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. CRC press, 2008.