Hierarchical Bayesian image analysis: From low-level modeling to robust supervised learning[☆]

Adrien Lagrange[a,*], Mathieu Fauvel[b], Stéphane May[c], Nicolas Dobigeon[a,d]

[a] University of Toulouse, IRIT/INP-ENSEEIHT Toulouse, BP 7122, Toulouse Cedex 7 31071, France
[b] INRA, DYNAFOR, BP 32607, Auzeville-Tolosane, 31326 Castanet Tolosan, France
[c] CNES, DCT/SI/AP, 18 Avenue Edouard Belin, 31400 Toulouse, France
[d] Institut Universitaire de France, France

Article history:
Received 30 November 2017
Revised 11 May 2018
Accepted 22 July 2018
Available online 31 July 2018

Keywords:
Bayesian model
Supervised learning
Image interpretation
Markov random field

ABSTRACT

Within a supervised classification framework, labeled data are used to learn classifier parameters. Prior to that, it is generally required to perform dimensionality reduction via feature extraction. These preprocessing steps have motivated numerous research works aiming at recovering latent variables in an unsupervised context. This paper proposes a unified framework to perform classification and low-level modeling jointly. The main objective is to use the estimated latent variables as features for classification and to incorporate simultaneously supervised information to help latent variable extraction. The proposed hierarchical Bayesian model is divided into three stages: a first low-level modeling stage to estimate latent variables, a second stage clustering these features into statistically homogeneous groups and a last classification stage exploiting the (possibly badly) labeled data. Performance of the model is assessed in the specific context of hyperspectral image interpretation, unifying two standard analysis techniques, namely unmixing and classification.

1. Introduction

In the context of image interpretation, numerous methods have been developed to extract meaningful information. Among them, generative models have received a particular attention due to their strong theoretical background and the great convenience they offer in term of interpretation of the fitted models compared to some model-free methods such as deep neural networks. These methods are based on an explicit statistical modeling of the data which allows very task-specific model to be derived [1], or either more general models to be implemented to solve generic tasks, such as Gaussian mixture model for classification [2]. Task-specific and classification-like models are two different ways to reach an interpretable description of the data with respect to a particular applicative non-semantic issue. For instance, when analyzing images, task-specific models aim at recovering the latent (possibly physics-based) structures underlying each pixel-wise measurement [3] while classification provides a high-level information, reducing the pixel characterization to a unique label [4].

Classification is probably one of the most common way to interpret data, whatever the application field of interest [5]. This undeniable appeal has been motivated by the simplicity of the resulting output. This simplicity induces the appreciable possibility of benefiting from training data at a relatively low cost. Indeed, experts can generally produce a ground-truth equivalent to the expected results of the classification for some amount of the data. This supervised approach allows a priori knowledge to be easily incorporated to improve the quality of the inferred classification model. Nevertheless, supervised methods are significantly influenced by the size of the training set, its representativeness and reliability [6]. Moreover, in some extent, modeling the pixel-wise data by a single descriptor may appear as somehow limited. It is the reason why the user-defined classes often refer to some rather vague semantic meaning with a possible large intra-class variability. To overcome these issues, while simultaneously facing with theoretical limitations of the expected classifier ability of generalization [7], an approach consists in preceding the training stage with feature extraction [8]. These feature extraction techniques, whether parametric or nonparametric, have also the great advantage of simultaneously and significantly reducing the data volume to be handled as well as the dimension of the space in which the training should be sub-

[☆] Part of this work has been supported Centre National d'Études Spatiales (CNES), Occitanie Region and EU FP7 through the ERANETMED JC-WATER program, MapInvPlnt Project ANR-15-NMED-0002-02.

[*] Corresponding author.

E-mail addresses: adrien.lagrange@enseeiht.fr (A. Lagrange), mathieu.fauvel@ensat.fr (M. Fauvel), stephane.may@cnes.fr (S. May), nicolas.dobigeon@enseeiht.fr (N. Dobigeon).

https://doi.org/10.1016/j.patcog.2018.07.026

sequently conducted. Unfortunately, they are generally conducted in a separate manner before the classification task, i.e., without benefiting from any prior knowledge available as training data. Thus, a possible strategy is to consider a (possibly huge) set of features and selecting the relevant ones by appropriate optimization schemes [9].

This observation illustrates the difficulty of incorporating ground-truthed information into a feature extraction step or, more generally, into a latent (i.e., unobserved) structure analysis. Due to the versatility of the data description, producing expert ground-truth with such degrees of accuracy and flexibility would be time-consuming and thus prohibitive. For example, for a research problem as important and well-documented as that of source separation, only very few and recent attempts have been made to incorporate supervised knowledge provided by an end-user [10]. Nonetheless, latent structure analysis may offer a relevant and meaningful interpretation of the data, since various conceptual yet structured knowledge to be inferred can be incorporated into the modeling. In particular, when dealing with measurements provided by a sensor, task-related biophysical considerations may guide the model derivation [11]. This is typically the case when spectral mixture analysis is conducted to interpret hyperspectral images whose pixel measurements are modeled as combinations of elementary spectra corresponding to physical elementary components [12].

The contribution of this paper lies in the derivation of a unified framework able to perform classification and latent structure modeling jointly. First, this framework has the primary advantage of recovering consistent high and low level image descriptions, explicitly conducting hierarchical image analysis. Moreover, improvements in the results associated with both methods may be expected thanks to the complementarity of the two approaches. The use of ground-truthed training data is not limited to driving the high level analysis, i.e., the classification task. Indeed, it also makes it possible to inform the low level analysis, i.e., the latent structure modeling, which usually does not benefit well from such prior knowledge. On the other hand, the latent modeling inferred from each data as low level description can be used as features for classification. A direct and expected side effect is the explicit dimension reduction operated on the data before classification [7]. Finally, the proposed hierarchical framework allows the classification to be robust to corruption of the ground-truth. As mentioned previously, performance of supervised classification may be questioned by the reliability in the training dataset since it is generally built by human expert and thus probably corrupted by label errors resulting from ambiguity or human mistakes. For this reason, the problem of developing classification methods robust to label errors has been widely considered in the community [13,14]. Pursuing this objective, the proposed framework also allows training data to be corrected if necessary.

The interaction between the low and high level models is handled by the use of non-homogeneous Markov random fields (MRF) [15]. MRFs are probabilistic models widely-used to describe spatial interactions. Thus, when used to derive a prior model within a Bayesian approach, they are particularly well-adapted to capture spatial dependencies between the latent structures underlying images [16,17]. For example, Chen et al. [18] proposed to use MRFs to perform clustering. The proposed framework incorporates two instances of MRF, ensuring consistency between the low and high level modeling, consistency with external data available as prior knowledge and a more classical spatial regularization.

The remaining of the article is organized as follows. Section 2 presents the hierarchical Bayesian model proposed as a unifying framework to conduct low-level and high-level image interpretation. A Markov chain Monte Carlo (MCMC) method is derived in Section 3 to sample according to the joint posterior distribution of the resulting model parameters. Then, a particular and illustrative instance of the proposed framework is presented in Section 4 where hyperspectral images are analyzed under the dual scope of unmixing and classification. Finally, Section 5 concludes the paper and opens some research perspectives to this work.

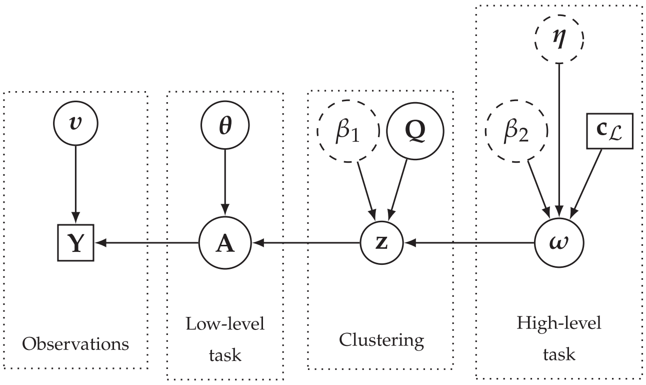

Fig. 1. Directed acyclic graph of the proposed hierarchical Bayesian model. (User-defined parameters appear in dotted circles and external data in squares).

2. Bayesian model

In order to propose a unifying framework offering multi-level image analysis, a hierarchical Bayesian model is derived to relate the observations and the task-related parameters of interest. This model is mainly composed of three main levels. The first level, presented in Section 2.1, takes care of a low-level modeling achieving latent structure analysis. The second stage then assumes that data samples (e.g., resulting from measurements) can be divided into several statistically homogeneous clusters through their respective latent structures. To identify the cluster memberships, these samples are assigned discrete labels which are a priori described by a non-homogeneous Markov random field (MRF). This MRF combines two terms: the first one is related to the potential of a Potts-MRF to promote spatial regularity between neighboring pixels; the second term exploits labels from the higher level to promote coherence between cluster and classification labels. This clustering process is detailed in Section 2.2. Finally, the last stage of the model, explained in Section 2.3, allows high-level labels to be estimated, taking advantage of the availability of external knowledge as ground-truthed or expert-driven data, akin to a conventional supervised classification task. The whole model and its dependences are summarized by the directed acyclic graph in Fig. 1.

2.1. Low-level interpretation

The low-level task aims at inferring P R-dimensional latent variable vectors $\mathbf{a}_p$ $(\forall p \in \mathcal{P} \triangleq \{1, \ldots, P\})$ appropriate for representing P respective d-dimensional observation vectors $\mathbf{y}_p$ in a subspace of lower dimension than the original observation space, i.e., $R \le d$. The task may also include the estimation of the function or additional parameters of the function relating the unobserved and observed variables. By denoting $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_P]$ and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_P]$ the $d \times P$- and $R \times P$- matrices gathering respectively the observation and latent variable vectors, this relation can be expressed through the general statistical formulation

$$\mathbf{Y}|\mathbf{A}, \boldsymbol{\upsilon} \sim \Psi(\mathbf{Y}; f_{\text{lat}}(\mathbf{A}), \boldsymbol{\upsilon}), \tag{1}$$

where $\Psi(\cdot, \boldsymbol{\upsilon})$ stands for a statistical model, e.g., resulting from physical or approximation considerations, $f_{\text{lat}}(\cdot)$ is a deterministic function used to define the latent structure and $\boldsymbol{\upsilon}$ are possible additional nuisance parameters. In most applicative contexts aimed by this work, the model $\Psi(\cdot)$ and function $f_{\text{lat}}(\cdot)$ are separable with respect to the measurements assumed to be conditionally independent, leading to the factorization

$$\mathbf{Y}|\mathbf{A},\boldsymbol{\upsilon}\sim\prod_{p=1}^{P}\Psi(\mathbf{y}_p;f_{\mathrm{lat}}(\mathbf{a}_p),\boldsymbol{\upsilon}). \tag{2}$$

It is worth noting that this statistical model will explicitly lead to the derivation of the particular form of the likelihood function involved in the Bayesian model.

The choice of the latent structure related to the function $f_{\mathrm{lat}}(\cdot)$ is application-dependent and can be directly chosen by the end-user. A conventional choice consists in considering a linear expansion of the observed data $\mathbf{y}_p$ over an orthogonal basis spanning a space whose dimension is lower than the original one. This orthogonal space can be a priori fixed or even learnt from the dataset itself, e.g., leveraging on popular nonparametric methods such as principal component analysis (PCA) [19]. In such case, the model (1) should be interpreted as a probabilistic counterpart of PCA [20] and the latent variables $\mathbf{a}_p$ would correspond to factor loadings. Similar linear latent factors and low-rank models have been widely advocated to address source separation problems, such as nonnegative matrix factorization [21]. As a typical illustration, by assuming an additive white and centered Gaussian statistical model $\Psi(\cdot)$ and a linear latent function $f_{\mathrm{lat}}(\cdot)$, the generic model (2) can be particularly instanced as

$$\mathbf{Y}|\mathbf{A},s^2\sim\prod_{p=1}^{P}\mathcal{N}\left(\mathbf{y}_p;\mathbf{M}\mathbf{a}_p,s^2\mathbf{I}_d\right) \tag{3}$$

where $\mathbf{I}_d$ is the $d\times d$ identity matrix, $\mathbf{M}$ is a matrix spanning the signal subspace and s^2 is the variance of the Gaussian error, considered as a nuisance parameter. Besides this popular class of Gaussian models, this formulation allows other noise statistics to be handled within a linear factor modeling, as required when the approximation should be envisaged beyond a conventional Euclidean discrepancy measure [22], provided that

$$\mathbb{E}[\mathbf{Y}|\mathbf{A}]=f_{\mathrm{lat}}(\mathbf{A}).$$

From a different perspective, the generic formulation of the statistical latent structure (2) can also result from a thorough analysis of more complex physical processes underlying observed measurements, resulting in specific yet richer physics-based latent models [11,23]. For sake of generality, this latent structure will not be specified in the rest of this manuscript, except in Section 4 where the linear Gaussian model (3) will be more deeply investigated as an illustration in a particular applicative context.

2.2. Clustering

To regularize the latent structure analysis, the model is complemented by a clustering step as a higher level of the Bayesian hierarchy. Besides, another objective of this clustering stage is also to act as a bridge between the low- and high-level data interpretations, namely latent structure analysis and classification. The clustering is performed under the assumption that the latent variables are statistically homogeneous and allocated in several clusters, i.e., identities belonging to a same cluster are supposed to be distributed according to the same distribution. To identify the membership, each observation is assigned a cluster label $z_p\in\mathcal{K}\triangleq\{1,\ldots,K\}$ where K is the number of clusters. Formally, the unknown latent vector is thus described by the following prior

$$\mathbf{a}_p|z_p=k,\boldsymbol{\theta}_k\sim\Phi(\mathbf{a}_p;\boldsymbol{\theta}_k), \tag{4}$$

where Φ is a given statistical model depending on the addressed problem and governed by the parameter vector $\boldsymbol{\theta}_k$ characterizing each cluster. As an example, considering this prior distribution as Gaussian, i.e., $\Phi(\mathbf{a}_p;\boldsymbol{\theta}_k)=\mathcal{N}(\mathbf{a}_p;\boldsymbol{\psi}_k,\boldsymbol{\Sigma}_k)$ with $\boldsymbol{\theta}_k=\left\{\boldsymbol{\psi}_k,\boldsymbol{\Sigma}_k\right\}$, would lead to a conventional Gaussian mixture model (GMM) for the latent structure, as in [24] (see Section 4).

One particularity of the proposed model lies in the prior on the cluster labels $\mathbf{z}=[z_1,\ldots,z_P]$. A non-homogeneous Markov Random Field (MRF) is used as a prior model to promote two distinct behaviors through the use of two potentials. The first one is a local and non-homogeneous potential parameterized by a K-by-J matrix $\mathbf{Q}$. It promotes consistent relationships between the cluster labels $\mathbf{z}$ and some classification labels $\boldsymbol{\omega}=[\omega_1,\ldots,\omega_P]$ where $\omega_p\in\mathcal{J}\triangleq\{1,\ldots,J\}$ and J is the number of classes. These classification labels associated with high-level interpretation will be more precisely investigated in the third stage of the hierarchy in Section 2.3. Pursuing the objective of analyzing images, the second potential is associated with a Potts-MRF [25] of granularity parameter β_1 to promote a piecewise consistent spatial regularity of the cluster labels. The prior probability of $\mathbf{z}$ is thus defined as

$$\mathbb{P}[\mathbf{z}|\boldsymbol{\omega},\mathbf{Q}]=\frac{1}{C(\boldsymbol{\omega},\mathbf{Q})}\exp\left(\sum_{p\in\mathcal{P}}V_1(z_p,\omega_p,q_{z_p,\omega_p})\right.$$
$$\left.+\sum_{p\in\mathcal{P}}\sum_{p'\in\mathcal{V}(p)}V_2(z_p,z_{p'})\right) \tag{5}$$

where $\mathcal{V}(p)$ stands for the neighborhood of p, $q_{k,j}$ is the kth element of the jth column of $\mathbf{Q}$. The two terms $V_1(\cdot)$ and $V_2(\cdot)$ are the classification-informed and Potts–Markov potentials, respectively, defined by

$$V_1(k,j,q_{k,j})=\log(q_{k,j})$$
$$V_2(k,k')=\beta_1\delta(k,k')$$

where $\delta(\cdot,\cdot)$ is the Kronecker function. Finally, $C(\boldsymbol{\omega},\mathbf{Q})$ stands for the normalizing constant (i.e., partition function) depending of $\boldsymbol{\omega}$ and $\mathbf{Q}$ and computed over all the possible $\mathbf{z}$ fields [15]

$$C(\boldsymbol{\omega},\mathbf{Q})=\sum_{\mathbf{z}\in\mathcal{K}^P}\exp\left(\sum_{p\in\mathcal{P}}V_1(z_p,\omega_p,q_{z_p,\omega_p})+\sum_{p\in\mathcal{P}}\sum_{p'\in\mathcal{V}(p)}V_2(z_p,z_{p'})\right)$$
$$=\sum_{\mathbf{z}\in\mathcal{K}^P}\prod_{p\in\mathcal{P}}q_{z_p,\omega_p}\exp\left(\beta_1\sum_{p'\in\mathcal{V}(p)}\delta(z_p,z_{p'})\right). \tag{6}$$

The equivalence between Gibbs random fields and MRF stated by the Hammersley–Clifford theorem [15] provides the prior probability of a particular cluster label conditionally upon its neighbors

$$\mathrm{P}[z_p=k|\mathbf{z}_{\mathcal{V}(p)},\omega_p=j,q_{k,\omega_p}]\propto$$
$$\exp\left(V_1(k,j,q_{k,j})+\sum_{p'\in\mathcal{V}(p)}V_2(k,z_{p'})\right) \tag{7}$$

where the symbol $\propto$ stands for “proportional to”.

The elements $q_{k,j}$ of the matrix $\mathbf{Q}$ introduced in the latter MRF account for the connection between cluster k and class j, revealing a hidden interaction between clustering and classification. A high value of $q_{k,j}$ tends to promote the association to the cluster k when the sample belongs to the class j. This interaction encoded through these matrix coefficients is unknown and thus motivates the estimation of the matrix $\mathbf{Q}$. To reach an interpretation of the matrix coefficients in terms of probabilities of interdependency, a Dirichlet distribution is elected as prior for each column $\mathbf{q}_j=[q_{1,j},\ldots,q_{K,j}]^T$ of $\mathbf{Q}=[\mathbf{q}_1,\ldots,\mathbf{q}_J]$ which are assumed to be independent, i.e.,

$$\mathbf{q}_j\sim\mathrm{Dir}(\mathbf{q}_j;\zeta_1,\ldots,\zeta_K). \tag{8}$$

The nonnegativity and sum-to-one constraints imposed to the coefficients defining each column of $\mathbf{Q}$ allows them to be interpreted as probability vectors. The choice of such a prior is furthermore motivated by the properties of the resulting conditional posterior distribution of $\mathbf{q}_j$, as demonstrated later in Section 3. In the present work, the hyperparameters $\zeta_1, \ldots, \zeta_K$ are all chosen equal to 1, resulting in a uniform prior over the corresponding simplex defined by the probability constraints. Obviously, when additional prior knowledge on the interaction between clustering and classification is available, these hyperparameters can be adjusted accordingly.

2.3. High-level interpretation

The last stage of the hierarchical model defines a classification rule. At this stage, a unique discrete class label should be attributed to each sample. This task can be seen as high-level in the sense that the definition of the classes can be motivated by their semantic meaning. Classes can be specified by the end-user and thus a class may gather samples with significantly dissimilar observation vectors and even dissimilar latent features. The clustering stage introduced earlier also allows a mixture model to be derived for this classification task. Indeed, a class tends to be the union of several clusters identified at the clustering stage, providing a hierarchical description of the dataset.

In this paper, the conventional and well-admitted setup of a supervised classification is considered. This setup means that a partial ground-truthed dataset $\mathbf{c}_{\mathcal{L}}$ is available for a (e.g., small) subset of samples. In what follows, $\mathcal{L} \subset \mathcal{P}$ denotes the subset of observation indexes for which this ground-truth is available. This ground-truth provides the expected classification labels for observations indexed by $\mathcal{L}$. Conversely, the index set of unlabeled samples for which this ground-truth is not available is noted $\mathcal{U} \subset \mathcal{P}$, with $\mathcal{P} = \mathcal{U} + \mathcal{L}$ and $\mathcal{U} \cap \mathcal{L} = \emptyset$. Moreover, the proposed model assumes that this ground-truth may be corrupted by class labeling errors. As a consequence, to provide a classification robust to these possible errors, all the classification labels of the dataset will be estimated, even those associated with the observations indexed by $\mathcal{L}$. At the end of the classification process, the labels estimated for observations indexed by $\mathcal{L}$ will not be necessarily equal to the labels $\mathbf{c}_{\mathcal{L}}$ provided by the expert or an other external knowledge.

Similarly to the prior model advocated for $\mathbf{z}$ (see Section 2.2), the prior model for the classification labels $\boldsymbol{\omega}$ is a non-homogeneous MRF composed of two potentials. Again, a Potts-MRF potential with a granularity parameter β_2 is used to promote spatial coherence of the classification labels. The other potential is non-homogeneous and exploits the supervised information available under the form of the ground-truth map $\mathbf{c}_{\mathcal{L}}$. In particular, it attends to ensure consistency between the estimated and ground-truthed labels for the samples indexed by $\mathcal{L}$. Moreover, for the classification labels associated with the indexes in $\mathcal{U}$ (i.e., for which no ground-truth is available), the prior probability to belong to a given class is set as the proportion of this class observed in $\mathbf{c}_{\mathcal{L}}$. This setting assumes that the expert map is representative of the whole scene to be analyzed in term of label proportions. If this assumption is not verified, the proposed modeling can be easily adjusted accordingly. Mathematically, this formal description can be summarized by the following conditional prior probability for a given classification label ω_p

$$\mathbb{P}[\omega_p = j | \boldsymbol{\omega}_{\mathcal{V}(p)}, c_p, \eta_p] \propto \exp\left(W_1(j, c_p, \eta_p) + \sum_{p' \in \mathcal{V}(p)} W_2(j, \omega_{p'}) \right). \quad (9)$$

As explained above, the potential $W_2(\cdot, \cdot)$ ensures the spatial coherence of the classification labels, i.e.,

$$W_2(j, j') = \beta_2 \delta(j, j').$$

More importantly, the potential $W_1(j, c_p, \eta_p)$ defined by

$$W_1(j, c_p, \eta_p) = \begin{cases} \begin{cases} \log(\eta_p), & \text{when } j = c_p \\ \log(\frac{1-\eta_p}{J-1}), & \text{otherwise} \end{cases}, & \text{when } p \in \mathcal{L} \\ \log(\pi_j), & \text{when } p \in \mathcal{U} \end{cases}$$

encodes the coherence between estimated and ground-truthed labels when available (i.e., when $p \in \mathcal{L}$) or, conversely for non-ground-truthed labels (i.e., when $p \in \mathcal{U}$), the prior probability of assigning a given label through the proportion π_j of samples of class j in $\mathbf{c}_{\mathcal{L}}$. The hyperparameter $\eta_p \in (0, 1)$ stands for the confidence given in c_p, i.e., the ground-truth label of pixel p. In the case where the confidence is total, the parameter tends to 1 and it leads to $\omega_p = c_p$ in a deterministic manner. However, in a more realistic applicative context, ground-truth is generally provided by human experts and may contain errors due for example to ambiguities or simple mistakes. It is possible with the proposed model to set for example a 90% level of confidence which allows to re-estimate the class label of the labeled set $\mathcal{L}$ and thus to correct the provided ground-truth. By this mean, the robustness of the classification to label errors is improved.

3. Gibbs sampler

To infer the parameters of the hierarchical Bayesian model introduced in the previous section, an MCMC algorithm is derived to generate samples according to the joint posterior distribution of interest which can be computed according to the following hierarchical structure

$$p(\mathbf{A}, \boldsymbol{\Theta}, \mathbf{z}, \mathbf{Q}, \boldsymbol{\omega} | \mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{A}) p(\mathbf{A}|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}|\mathbf{Q}, \boldsymbol{\omega}) p(\boldsymbol{\omega})$$

with $\boldsymbol{\Theta} \triangleq \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$. Note that, for conciseness, the nuisance parameters $\boldsymbol{\upsilon}$ have been implicitly marginalized out in the hierarchical structure. If this marginalization is not straightforward, these nuisance parameters can be also explicitly included within the model to be jointly estimated.

The Bayesian estimators of the parameters of interest can then be approximated using these samples. The minimum mean square error (MMSE) estimators of the parameters **A, Θ** and **Q** can be approximated through empirical averages

$$\hat{\mathbf{x}}_{\text{MMSE}} = \mathbb{E}[\mathbf{x}|\mathbf{Y}] \approx \frac{1}{N_{\text{MC}}} \sum_{t=1}^{N_{\text{MC}}} \mathbf{x}^{(t)} \quad (10)$$

where $\cdot^{(t)}$ denotes the tth samples and N_{MC} is the number of iterations after the burn-in period. Conversely, the maximum a posteriori estimators of the cluster and class labels, $\mathbf{z}$ and $\boldsymbol{\omega}$, respectively, can be approximated as

$$\hat{\mathbf{x}}_{\text{MAP}} = \underset{\mathbf{x}}{\operatorname{argmax}}\, p(\mathbf{x}|\mathbf{Y}) \approx \underset{\mathbf{x}^{(t)}}{\operatorname{argmax}}\, p(\mathbf{x}^{(t)}|\mathbf{Y}) \quad (11)$$

which basically amounts at retaining the most frequently generated label for these specific discrete parameters [26].

To carry out such a sampling strategy, the conditional posterior distributions of the various parameters need to be derived. More importantly, the ability of drawing according to these distributions is required. These posterior distributions are detailed in what follows.

3.1. Latent parameters

Given the likelihood function resulting from the statistical model (2) and the prior distribution in (4), the conditional posterior distribution of a latent vectors can be expressed as follows:

$$\begin{aligned} p(\mathbf{a}_p | \mathbf{y}_p, \boldsymbol{\upsilon}, z_p = k, \boldsymbol{\theta}_k) &\propto p(\mathbf{y}_p | \mathbf{a}_p, \boldsymbol{\upsilon}) p(\mathbf{a}_p | z_p = k, \boldsymbol{\theta}_k) \\ &\propto \Psi(\mathbf{y}_p; f_{\text{lat}}(\mathbf{a}_p), \boldsymbol{\upsilon}) \Phi(\mathbf{a}_p; \boldsymbol{\theta}_k). \end{aligned} \quad (12)$$

3.2. Cluster labels

The cluster label z_p being a discrete random variable, it is possible to sample the variable by computing the conditional probability for all possible values of z_p in $\mathcal{K}$

$$
\begin{aligned}
&\mathbb{P}(z_p = k|\boldsymbol{\theta}_k, \omega_p = j, q_{k,j}) \\
&\quad \propto p(\mathbf{a}_p|z_p = k, \boldsymbol{\theta}_k)\mathbb{P}(z_p = k|\mathbf{z}_{\mathcal{V}(p)}, \omega_p = j, q_{k,j}) \\
&\quad \propto \Phi(\mathbf{a}_p; \boldsymbol{\theta}_k) q_{k,j} \exp\left(\beta_1 \sum_{p' \in \mathcal{V}(p)} \delta(k, z'_p)\right). \qquad (13)
\end{aligned}
$$

3.3. Interaction matrix

The conditional distribution of each column $\mathbf{q}_j$ $(j \in \mathcal{J})$ of the interaction parameter matrix $\mathbf{Q}$ can be written

$$
\begin{aligned}
p(\mathbf{q}_j|\mathbf{z}, \mathbf{Q}_{\backslash j}, \boldsymbol{\omega}) &\propto p(\mathbf{q}_j)\mathbb{P}(\mathbf{z}|\mathbf{Q}, \boldsymbol{\omega}) \\
&\propto \frac{\prod_{k=1}^{K} q_{k,j}^{n_{k,j}}}{C(\boldsymbol{\omega}, \mathbf{Q})} \mathbb{1}_{\mathbb{S}_K}(\mathbf{q}_j). \qquad (14)
\end{aligned}
$$

where $\mathbf{Q}_{\backslash j}$ denotes the matrix $\mathbf{Q}$ whose jth column has been removed, $n_{k,j} = \#\{p|z_p = k, \omega_p = j\}$ is the number of observations whose cluster and class labels are respectively k and j, and $\mathbb{1}_{\mathbb{S}_K}(\cdot)$ is the indicator function of the K-dimensional probability simplex which ensures that $\mathbf{q}_j \in \mathbb{S}_K$ implies $\forall k \in \mathcal{K},\ q_{k,\,j} \geq 0$ and $\sum_{k=1}^{K} q_{k,j} = 1$.

Sampling according to this conditional distribution would require to compute the partition function $C(\boldsymbol{\omega}, \mathbf{Q})$, which is not straightforward. The partition function is indeed a sum over all possible configurations of the MRF $\mathbf{z}$. One strategy would consist in precomputing this partition function on an appropriate grid, as in [27]. As alternatives, one could use to likelihood-free Metropolis Hastings algorithm [28], auxiliary variables [29] or pseudo-likelihood estimators [30]. However, all these strategies remain of high computational cost, which precludes their practical use for most applicative scenarii encountered in real-world image analysis.

Besides, when $\beta_1 = 0$, this partition function reduces to $C(\boldsymbol{\omega}, \mathbf{Q}) = 1$. In other words, the partition function is constant when the spatial regularization induced by $V_2(\cdot)$ is not taken into account. In such case, the conditional posterior distribution for $\mathbf{q}_j$ is the following Dirichlet distribution

$$
\mathbf{q}_j|\mathbf{z}, \boldsymbol{\omega} \sim \mathrm{Dir}(\mathbf{q}_j; n_{1,j}+1, \ldots, n_{K,j}+1), \qquad (15)
$$

which is easy to sample from. Interestingly, the expected value of $q_{k,\,j}$ is then

$$
\mathrm{E}\big[q_{k,j}|\mathbf{z}, \boldsymbol{\omega}\big] = \frac{n_{k,j}+1}{\sum_{i=1}^{K} n_{i,k} + K}
$$

which is a biased empirical estimator of $\mathrm{P}[z_p = k|\omega_p = j]$. This latter result motivates the use of a Dirichlet distribution as a prior for $\mathbf{q}_j$. Thus, it is worth noting that $\mathbf{Q}$ can be interpreted as a byproduct of the proposed model which describes the intrinsic dataset structure. It allows the practitioner not only to get an overview of the distribution of the samples of a given class in the various clusters but also to possibly identify the origin of confusions between several classes. Again, this clustering step allows disparity in the semantic classes to be mitigated. Intraclass variability results in the emerging of several clusters which are subsequently agglomerated during the classification stage.

In practice, during the burn-in period of the proposed Gibbs sampler, to avoid highly intensive computations, the cluster labels are sampled according to (13) with $\beta_1 > 0$ while the columns of the interaction matrix are sampled according to (15). In other words, during this burn-in period, a certain spatial regularization with $\beta_1 > 0$ is imposed to the cluster labels and the interaction matrix is sampled according to an approximation of its conditional posterior distribution.[1] After this burn-in period, the granularity parameter β_1 is set to 0, which results in removing the spatial regularization between the cluster labels. Thus, once convergence has been reached, the conditional posterior distribution (15) reduces to (14) and the iteraction matrix is properly sampled according to its exact conditional posterior distribution.

3.4. Classification labels

Similarly to the cluster labels, the classification labels $\boldsymbol{\omega}$ are sampled by evaluating their conditional probabilities computed for all the possible labels. However, two cases need to be considered while sampling the classification label ω_p, depending on the availability of ground-truth label for the corresponding pth pixel. More precisely, when $p \in \mathcal{U}$, i.e., when the pth pixel is not accompanied by a corresponding ground-truth, the conditional probabilities are written

$$
\begin{aligned}
&\mathbb{P}[\omega_p = j|\mathbf{z}, \boldsymbol{\omega}_{\backslash p}, \mathbf{q}_j, c_p, \eta_p] \\
&\quad \propto \mathbb{P}[z_p|\omega_p = j, \mathbf{q}_j, \mathbf{z}_{\nu(p)}]\mathbb{P}[\omega_p = j|\boldsymbol{\omega}_{\mathcal{V}(p)}, c_p, \eta_p] \\
&\quad \propto \frac{q_{z_p,j}\pi_j \exp\left(\beta_2 \sum_{p' \in \nu(p)} \delta(j, \omega_{p'})\right)}{\sum_{k'=1}^{K} q_{k',j} \exp\left(\beta_1 \sum_{p' \in \nu(p)} \delta(k', z_{p'})\right)}, \qquad (16)
\end{aligned}
$$

where $\boldsymbol{\omega}_{\backslash p}$ denotes the classification label vector $\boldsymbol{\omega}$ whose pth element has been removed. Conversely, when $p \in \mathcal{L}$, i.e., when the pth pixel is assigned a ground-truth label c_p, the conditional posterior probability reads

$$
\begin{aligned}
&\mathbb{P}[\omega_p = j|\mathbf{z}, \boldsymbol{\omega}_{\backslash p}, \mathbf{q}_j, c_p, \eta_p] \\
&\quad \propto \mathbb{P}[z_p|\omega_p = j, \mathbf{q}_j, \mathbf{z}_{\nu(p)}]\mathbb{P}[\omega_p = j|\boldsymbol{\omega}_{\mathcal{V}(p)}, c_p, \eta_p] \\
&\quad \propto \begin{cases} \dfrac{q_{z_p,j}\eta_p \exp\left(\beta_2 \sum_{p' \in \nu(p)} \delta(j,\omega_{p'})\right)}{\sum_{k'=1}^{K} q_{k',j} \exp\left(\beta_1 \sum_{p' \in \nu(p)} \delta(k',z_{p'})\right)} & \text{when } \omega_p = c_p \\ \dfrac{(1-\eta_p) q_{z_p,j} \exp\left(\beta_2 \sum_{p' \in \nu(p)} \delta(j,\omega_{p'})\right)}{(C-1)\sum_{k'=1}^{K} q_{k',j} \exp\left(\beta_1 \sum_{p' \in \nu(p)} \delta(k',z_{p'})\right)} & \text{otherwise} \end{cases} \qquad (17)
\end{aligned}
$$

Note that, as for the sampling of the columns $\mathbf{q}_j$ $(j \in \mathcal{J})$ of the interaction matrix $\mathbf{Q}$, this conditional probability is considerably simplified when $\beta_1 = 0$ (i.e., when no spatial regularization is imposed on the cluster labels) since, in this case, $\sum_{k'=1}^{K} q_{k',j} \exp(\beta_1 \sum_{p' \in \nu(p)} \delta(k', z_{p'})) = 1$.

4. Application to hyperspectral image analysis

The proposed general framework introduced in the previous sections has been instanced for a specific application, namely the analysis of hyperspectral images. Hyperspectral imaging for Earth observation has been receiving increasing attention over the last decades, in particular in signal/image processing literatures [31–33]. This keen interest of the scientific community can be easily explained by the richness of the information provided by such images. Indeed, generalizing the conventional red/green/blue color imaging, hyperspectral imaging collects spatial measurements acquired in a large number of spectral bands. Each pixel is associated with a vector of measurements, referred to as *spectrum*, which characterizes the macroscopic components present in this

[1] This strategy can also be interpreted as choosing $C(\boldsymbol{\omega}, \mathbf{Q}) \times \mathrm{Dir}(\mathbf{1})$ instead of the Dirichlet distribution (8) as prior for $\mathbf{q}_j$.

pixel. Classification and spectral unmixing are two well-admitted techniques to analyze hyperspectral images. As mentioned earlier, and similarly to numerous applicative contexts, classifying hyperspectral images consists in assigning a discrete label to each pixel measurement in agreement with a predefined semantic description of the image. Conversely, spectral unmixing proposes to retrieve some elementary components, called *endmembers*, and their respective proportions, called *abundance* in each pixel, associated with the spatial distribution of the endmembers in over the scene [12]. Per se, spectral unmixing can be cast as a blind source separation or a nonnegative matrix factorization (NMF) task [34]. The particularity of spectral unmixing, also known as spectral mixture analysis in the microscopy literature [35], lies in the specific constraints applied to spectral unmixing. As for any NMF problem, the endmembers signatures as well as the proportions are nonnegative. Moreover, specifically, to reach a close description of the pixel measurements, the abundance coefficients, interpreted as concentrations of the different materials, should sum to one for each spatial position.

Nevertheless, yet complementary, these two classes of methods have been considered jointly in a very limited number of works [36,37]. The proposed hierarchical Bayesian model offers a great opportunity to design a unified framework where these two methods can be conducted jointly. Spectral unmixing is perfectly suitable to be envisaged as the low-level task of the model described in Section 2. The abundance vector provides a biophysical description of a pixel which can be seen as a vector of latent variables of the corresponding pixel. The classification step is more related to a semantic description of the pixel. The low-level and clustering tasks of general framework described, respectively, in Sections 2.1 and 2.2, are specified in what follows, while the classification task is directly implemented as in Section 2.3.

4.1. Bayesian model

Low-level interpretation: According to the conventional linear mixing model (LMM), the pixel spectrum $\mathbf{y}_p$ $(p \in \mathcal{P})$ observed in d spectral bands are approximated by linear mixtures of R elementary signatures $\mathbf{m}_r$ $(r = 1, \ldots, R)$, i.e.,

$$\mathbf{y}_p = \sum_{r=1}^{R} a_{r,p}\mathbf{m}_r + \mathbf{e}_p \tag{18}$$

where $\mathbf{a}_p = [a_{1,p}, \ldots, a_{R,p}]^T$ denotes the vector of mixing coefficients (or abundances) associated with the pth pixel and $\mathbf{e}_p$ is an additive error assumed to be white and Gaussian, i.e., $\mathbf{e}_p|s^2 \sim \mathcal{N}(\mathbf{0}_d, s^2\mathbf{I}_d)$. When considering the P pixels of the hyperspectral image, the LMM can be rewritten with its matrix form

$$\mathbf{Y} = \mathbf{MA} + \mathbf{E} \tag{19}$$

where $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_R]$, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_P]$ and $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_P]$ are the matrices of the endmember signatures, abundance vectors and noise, respectively. In this work, the endmember spectra are assumed to be a priori known or previously recovered from the hyperspectral images by using an endmember extraction algorithm [12]. Under this assumption, the LMM matrix formulation defined by (19) can be straightforwardly interpreted as a particular instance of the low-level interpretation (1) by choosing the latent function $f_{\text{lat}}(\cdot)$ as a linear mapping $f_{\text{lat}}(\mathbf{A}) = \mathbf{MA}$ and the statistical model $\Psi(\cdot, \cdot)$ as the Gaussian probability density function parametrized by the variance s^2.

In this applicative example, since the error variance s^2 is a nuisance parameter and generally unknown, this hyperparameter is included within the Bayesian model and estimated jointly with the parameters of interest. More precisely, the variance s^2 is assigned a conjugate inverse-gamma prior and a non-informative Jeffreys hyperprior is chosen for the associate hyperparameter δ

$$s^2|\delta \sim \mathcal{IG}(s^2; 1, \delta), \quad \delta \propto \frac{1}{\delta}\mathbb{1}_{\mathbb{R}^+}(\delta). \tag{20}$$

These choices lead to the following inverse-gamma conditional posterior distribution

$$s^2|\mathbf{Y}, \mathbf{A} \sim \mathcal{IG}\left(s^2; 1 + \frac{Pd}{2}, \frac{1}{2}\sum_{p=1}^{P}\|\mathbf{y}_p - \mathbf{M}\mathbf{a}_p\|^2\right) \tag{21}$$

which is easy to sample from, as an additional step within the Gibbs sampling scheme described in Section 3.

Clustering: In the current problem, the latent modeling $\Phi(\cdot;\ \cdot)$ in (4) is chosen as Gaussian distributions elected for the latent vectors $\mathbf{a}_p$ $(p \in \mathcal{P})$,

$$\mathbf{a}_p|z_p = k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k \sim \mathcal{N}(\mathbf{a}_p; \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \tag{22}$$

where $\boldsymbol{\psi}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix associated with the kth cluster. This Gaussian assumption is equivalent to consider each high-level class as a mixture of Gaussian distributions in the abundance space. The covariance matrices are chosen as $\boldsymbol{\Sigma}_k = \text{diag}(\sigma^2_{k,1}, \ldots, \sigma^2_{k,R})$ where $\sigma^2_{k,1}, \ldots, \sigma^2_{k,R}$ are a set of R unknown hyperparameters. The conditional posterior distribution of the abundance vectors $\mathbf{a}_p$ can be finally expressed as follows:

$$p(\mathbf{a}_p|z_p = k, \mathbf{y}_p, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \propto |\boldsymbol{\Lambda}_k|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{a}_p - \boldsymbol{\mu}_{k,p})^t \boldsymbol{\Lambda}_k^{-1}(\mathbf{a}_p - \boldsymbol{\mu}_{k,p})\right) \tag{23}$$

where $\boldsymbol{\mu}_{k,p} = \boldsymbol{\Lambda}_k(\frac{1}{s^2}\mathbf{M}^t\mathbf{y}_p + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k)$ and $\boldsymbol{\Lambda}_k = (\frac{1}{s^2}\mathbf{M}^t\mathbf{M} + \boldsymbol{\Sigma}_k^{-1})^{-1}$. It shows that the latent vector $\mathbf{a}_p$ associated with a pixel belonging to the kth cluster is distributed according to the multivariate Gaussian distribution $\mathcal{N}(\mathbf{a}_p; \boldsymbol{\mu}_{k,p}, \boldsymbol{\Lambda}_k)$.

Moreover the variances $\sigma^2_{k,r}$ are included into the Bayesian model by choosing conjugate inverse-gamma prior distributions

$$\sigma^2_{k,r} \sim \mathcal{IG}(\sigma^2_{k,r}; \xi, \gamma) \tag{24}$$

where parameters ξ and γ have been selected to obtain vague priors $(\xi = 1, \gamma = 0.1)$. It leads to the following conditional inverse-gamma posterior distribution

$$\sigma^2_{r,k}|\mathbf{A}, \mathbf{z}, \psi_{r,k} \sim \mathcal{IG}\left(\sigma^2_{k,r}; \frac{n_k}{2} + \xi, \gamma + \sum_{p\in\mathcal{I}_k}\frac{(a_{r,p} - \psi_{r,k})^2}{2}\right) \tag{25}$$

where n_k is the number of samples in cluster k, and $\mathcal{I}_k \subset \mathcal{P}$ is the set of indexes of pixels belonging to the kth cluster (i.e., such that $z_p = k$).

Finally, the prior distribution of the cluster mean $\boldsymbol{\psi}_k$ $(k \in \mathcal{K})$ is chosen as a Dirichlet distribution $\text{Dir}(\mathbf{1})$. Such a prior induces *soft* non-negativity and sum-to-one constraints on $\mathbf{a}_p$. Indeed, these two constraints are generally admitted to describe the abundance coefficients since they represent proportions/concentrations. In this work, this constraint is not directly imposed on the abundance vectors but rather on their mean vectors, since $\text{E}[\mathbf{a}_p|z_p = k] = \boldsymbol{\psi}_k$. The resulting conditional posterior distribution of the mean vector $\boldsymbol{\psi}_k$ is the following multivariate Gaussian distribution

$$\boldsymbol{\psi}_k|\mathbf{A}, \mathbf{z}, \boldsymbol{\Sigma}_k \sim \mathcal{N}_{\mathbb{S}_R}\left(\boldsymbol{\psi}_k; \frac{1}{n_k}\sum_{p\in\mathcal{I}_k}\mathbf{a}_p, \frac{1}{n_k}\boldsymbol{\Sigma}_k\right) \tag{26}$$

truncated on the probability simplex

$$\mathbb{S}_R = \left\{\mathbf{x} = [x_1, \ldots, x_R]^T | \forall r,\ x_r \geq 0 \text{ and } \sum_{r=1}^{R} x_r = 1\right\}. \tag{27}$$

Sampling according to this truncated Gaussian distribution can be achieved following the strategies described in [38].

Algorithm 1: Inference using Gibbs sampling.

```
Initialize all variables;
for $N_{\mathrm{MC}} + N_{\mathrm{burn}}$ iterations do
    foreach $p \in \mathcal{P}$ do sample $a_p$ from $\mathcal{N}(\mu_{k,p}, \Lambda_k)$;
    foreach $p \in \mathcal{P}$ do sample $z_p$ from (13);
    foreach $j \in \mathcal{J}$ do sample $\mathbf{q}_j$ from $\mathrm{Dir}(n_{1,j}+1, \dots, n_{K,j}+1)$;
    foreach $p \in \mathcal{P}$ do sample $\omega_p$ from (16) and (17);
    for $k = 1$ to $K$ do
        sample $\boldsymbol{\psi}_k$ from $\mathcal{N}_{\mathbb{S}_R}\left(\frac{1}{n_k}\sum_{p\in\mathcal{I}_k}\mathbf{a}_p, \frac{1}{n_k}\boldsymbol{\Sigma}_k\right)$;
        foreach $r \in \{1, \dots, R\}$ do sample $\sigma^2_{r,k}$ from $\mathcal{IG}\left(\frac{n_k}{2}+\xi, \gamma+\sum_{p\in\mathcal{I}_k}\frac{(a_{r,p}-\psi_{r,k})^2}{2}\right)$;
    end
    sample $s^2$ from $\mathcal{IG}\left(1+\frac{Pd}{2}, \frac{1}{2}\sum_{p=1}^{P}\|\mathbf{y}_p - \mathbf{M}\mathbf{a}_p\|^2\right)$;
    sample $\delta$ from $\mathcal{IG}(1, s^2)$;
    if iteration > $N_{\mathrm{burn}}$ then
        update MMSE and MAP estimators
    end
end
```

Full inference procedure is summarized in Algorithm 1. It should be noticed that MMSE and MAP estimators are updated online at each iteration after the burn-in period in order to save storage and thus possibly handle large dataset. Additionally, the number of iteration is chosen in order to get a reasonable processing time.

4.2. Experiments

4.2.1. Synthetic dataset

Synthetic data have been used to assess the performance of the proposed analysis model and algorithm. Two distinct images, referred to as Image 1 and Image 2 and represented in Fig. 2, have been considered. The first one is a 100 × 100 pixel image composed of $R = 3$ endmembers, $K = 3$ clusters and $J = 2$ classes. The second hyperspectral image is a 200 × 200 pixel image which consists of $R = 9$ endmembers, $K = 12$ clusters and $J = 5$ classes. They have been synthetically generated according to the following hierarchical procedure. First, cluster maps have been generated from Potts–Markov MRFs to obtain (b) and (d) from Fig. 2. Then, the corresponding classification maps have then been chosen by artificially merging a few of these clusters to define each class and get (a) and (c) from Fig. 2. For each pixel, an abundance vector $\mathbf{a}_p$ has been randomly drawn from a Dirichlet distribution parametrized by a specific mean for each cluster. Finally the pixel measurements $\mathbf{Y}$ have been generated using the linear mixture model with real endmembers signatures of $d = 413$ spectral bands extracted from a spectral library. These linearly mixed pixels have been corrupted by a Gaussian noise resulting in a signal-to-noise ratio of SNR = 30 dB. The real interaction matrix $\mathbf{Q}$ presented in Fig. 2 (e) and (f) summarized the data structure by providing the probability to be in a given cluster when belonging to a given class.

Fig. 3represents the abundance vectors of each pixel in the probabilistic simplex for Image 1. The three clusters are clearly identifiable and the class represented in blue is also clearly divided into two clusters.

To evaluate the interest of including the classification step into the model, results provided by the proposed method have been compared to the counterpart model proposed in [24] (referred to as Eches model). The Eches model is a similar model which lacks the classification stage and thus does not exploit this high-level information. Fig. 4 presents the directed acyclic graph summarizing the model and its dependences in this particular hyperspectral framework and outlining the difference with Eches model. The pixels and associated classification labels located in the upper quarters of the Images 1 and 2 have been used as the training set $\mathcal{L}$. The confidence in this classification ground-truth has been set to a value of $\eta_p = 0.95$ for all the pixels ($p \in \mathcal{L}$). Additionally, the values of Potts-MRF granularity parameters have been selected as $\beta_1 = \beta_2 = 0.8$. In the case of the Eches model, the images have been subsequently classified using the estimated abundance vectors and clustering maps, and following the strategy proposed in [13]. The performance of the spectral unmixing task has been evaluated using the root global mean square error (RGMSE) associated with the abundance estimation

$$\mathrm{RGMSE}(\mathbf{A}) = \sqrt{\frac{1}{PR}\left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_F^2} \tag{28}$$

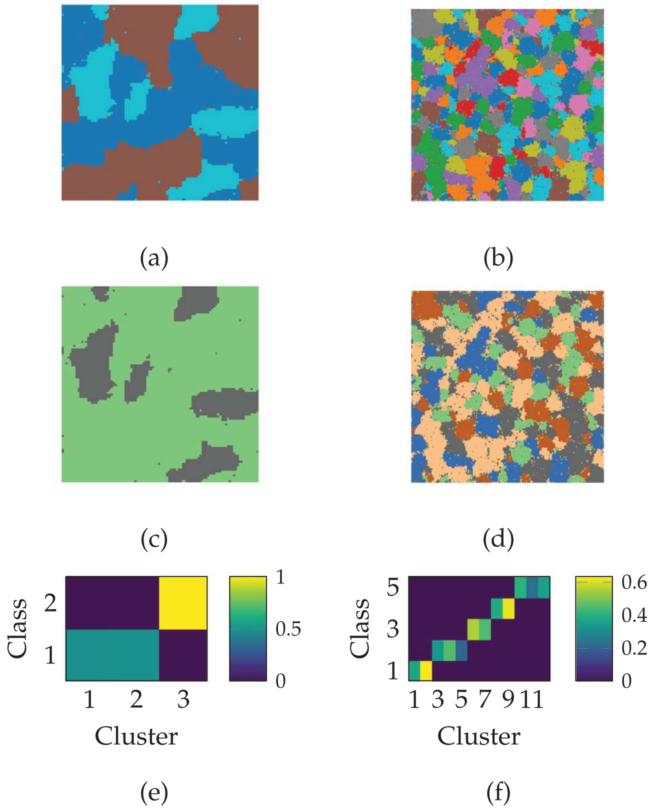

Fig. 2. Synthetic data. Classification maps of Image 1 (a) and Image 2 (b), corresponding clustering maps of Image 1 (c) and Image 2 (d), corresponding interaction matrix **Q** of Image 1 (e) and Image 2 (f).

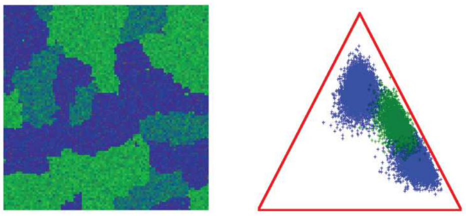

Fig. 3. Image 1. Left: colored composition of abundance map. Right: pixels in the probabilistic simplex (red triangle) with Class 1 (blue) and Class 2 (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 1
Unmixing and classification results for various datasets.

		RGMSE(A)	Kappa	Time (s)
Image 1	Proposed model	3.23e-03 (1.6e-05)	0.932 (0.018)	171 (5.4)
	Eches model	3.24e-03 (1.4e-05)	0.909 (0.012)	146 (0.7)
Image 2	Proposed model	1.62e-02 (1.62e-04)	0.961 (0.04)	950 (11)
	Eches model	1.61e-02 (2.71e-05)	0.995 (0.0004)	676 (2.1)
MUESLI image	Proposed model	N\ A	0.837 (5e-3)	7175 (102)
	Random Forest	N\ A	0.879 (5e-4)	34 (1.3)
	Gaussian model	N\ A	0.818 (8.7e-5)	4 (0.01)

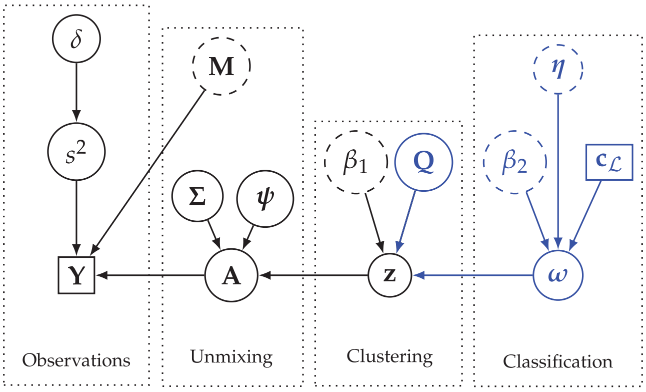

Fig. 4. Directed acyclic graph of the proposed model in the described hyperspectral framework. Part in blue is the extension made to the Eches model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

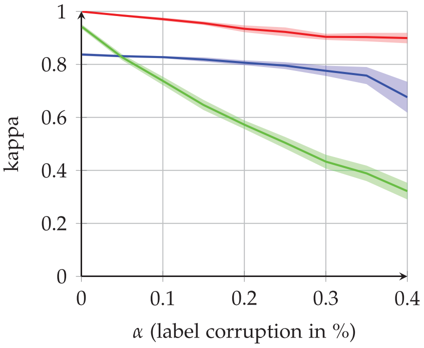

Fig. 5. Classification accuracy measured with Cohen's kappa as a function of label corruption α: proposed model (red), MDA with abundance vectors (blue) and MDA with measured reflectance (blue). Shaded areas denote the intervals corresponding to the standard deviation computed over 20 trials. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

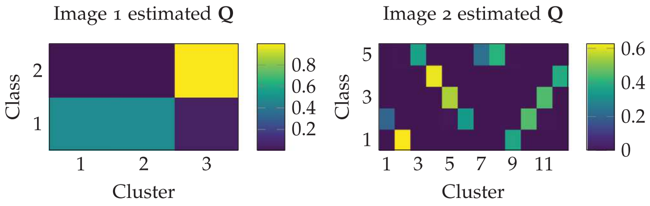

Fig. 6. Estimated interaction matrix **Q** for Image 1 (top) and Image 2 (bottom).

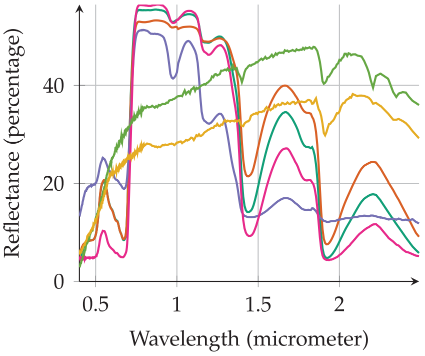

Fig. 7. Spectra used to generate the semi-synthetic image. 4 spectra are vegetation spectra and 2 are soil spectra.

where $\hat{\mathbf{A}}$ and $\mathbf{A}$ denote, respectively, the estimated and actual matrices of abundance vectors. Moreover, the accuracy of the estimated classification maps has been measured with the conventional Cohen's kappa. Results reported in Table 1 show that the obtained RGMSE are not significantly different between the two models. Moreover, the comparison between processing times shows a small computational overload required by the proposed model. It should be noticed that this experiment has been conducted with a fixed number of iterations of the proposed MCMC algorithm (300 iterations including 50 burn-in iterations).

A second scenario is considered where the training set includes label errors. The corrupted training set is generated by tuning a varying probability α to assign an incorrect label, all the other possible labels being equiprobable. The probability α varies from 0 to 0.4 with a 0.05 step. In this context, the confidence in the classification ground-truth map is set equal to $\eta_p = 1 - \alpha$ ($\forall p \in \mathcal{L}$). The results, averaged over 20 trials for each setting, are compared to the results obtained using a mixture discriminant analysis (MDA) [39] conducted either directly on the pixel spectra, either on the abundance vectors estimated with the proposed model. The resulting classification performances for Image 1 are depicted in Fig. 5 as function of α. These results show that the proposed model performs very well even when the training set is highly corrupted (i.e., α close to 0.4).

Moreover, as already explained, another advantage of the proposed model is the interesting by-products provided by the method. As an illustration, Fig. 6 presents the interactions matrices **Q** estimated for each image. From this figure, it is clearly possible to identify the structure of the various classes and their hierarchical relationship with the underlying clusters. For instance, for Image 2, it can be noticed that Class 1 is essentially composed of two clusters which is confirmed by the true interaction matrix presented in Fig. 2 (e).

A last scenario has been considered in order to show the interest of the proposed method in term of spectral unmixing. A more complex synthetic image has been generated to assess this point. A 100 × 250-pixel real hyperspectral image has been unmixed using the fully constrained optimization method described in [40]. The obtained realistic abundance maps have been used to generate a new image with new real endmembers signatures of $d = 252$ spectral bands extracted from a spectral library. The selected endmembers presented in Fig. 7 has been chosen in order to be highly correlated (4 vegetation spectra and 2 soils spectra). Moreover the endmembers matrix **M** has been augmented by 9 endmembers not

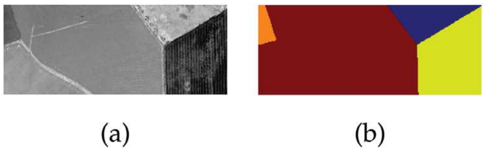

Fig. 8. Semi-synthetic image. Panchromatic view of the hyperspectral image (a), ground-truth (b).

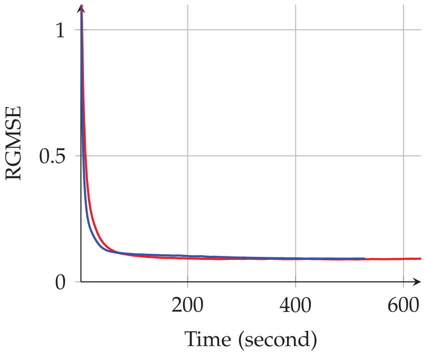

Fig. 9. Evolution of RGMSE of the sampled $\hat{\mathbf{A}}^{(t)}$ matrix in function of the time for the proposed model (red) and Eches model (blue). Results are averaged in time and score over 10 trials. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

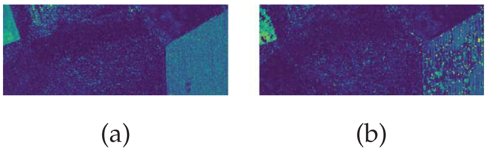

Fig. 10. Semi-synthetic image. Example of error map $(\|\hat{\mathbf{a}}_p - \mathbf{a}_p\|_2)$ for proposed model (a), example of error map for Eches model (b).

present in the image. The obtained data is indeed both realistic and challenging in term of unmixing. A panchromatic view of the resulting image, made by summing all spectral bands, is presented in Fig. 8 along with the ground-truth retrieved from the one provided with the original image with $J = 4$ classes. A Gaussian noise is finally added to this semi-synthetic image to get a signal-to-noise ratio of SNR$=$ 10 dB.

Fig. 9 shows the evolution of RGMSE computed at each iteration for 250 iterations using the sampled $\hat{\mathbf{A}}^{(t)}$ matrix and the known **A** abundance matrix. For this experiment, the whole classification ground-truth was provided to the proposed algorithm as expert data $\mathbf{c}_{\mathcal{L}}$ and parameters have been set to $\beta_1 = 0.3$ and $\beta_2 = 1.2$ for the proposed model and $\beta_1 = 1.2$ for Eches model. The evolution of the RGMSE is presented in function of the time since iteration are longer with the proposed model than with Eches model. Contrary to one would expect, the proposed model appears to be much faster to converge in number of iterations resulting in a convergence in the same time than Eches model. The increase of complexity and processing time is compensated by the fact that the classification information help significantly the convergence. Moreover as shown in Fig. 10, the error made by the proposed model tends to be more spatially coherent than the error made by Eches model which are sometimes scattered in small area. This limitation of Eches model is induced by the tendency to over-segment the image in more clusters than necessary.

4.2.2. Real hyperspectral image

Finally, the proposed strategy has been implemented to analyze a real 600 × 600-pixel hyperspectral image acquired within the framework of the *multiscale mapping of ecosystem services by very high spatial resolution hyperspectral and LiDAR remote sensing imagery* (MUESLI) project[2] This image is composed of $d = 438$ spectral bands and $R = 7$ endmembers have been extracted using the widely-used vertex component analysis (VCA) algorithm [41] to obtain matrix **M**. The associated expert ground-truth classification is made of 6 classes (straw cereals, summer crops, wooded area, buildings, bare soil, pasture). In this experiment, the upper half of the expert ground-truth has been provided as training data for the proposed method. The confidence η_p has been set to 95% for all training pixels to account for the imprecision of the expert ground-truth. The MRF granularity parameters of the proposed parameters have been set to $\beta_1 = 0.3$ and $\beta_2 = 1$ since these values provide the most meaningful interpretation of the image. Fig. 11 presents a colored composition of the hyperspectral image (a), the expert ground-truth (b) and the obtained results in terms of clustering (c) and classification (d). Quantitative results in term of classification accuracy have been computed and are summarized in Table 1. Note that no performance measure of the unmixing step is provided since no abundance ground-truth is available for this real dataset.

For comparison purposed, classification has been conducted with two conventional classifier namely random forest (RF) and a Bayesian Gaussian model (GM) using scikit-learn library. Parameters of the two classifiers have been optimized using cross-validation on the training set. Additionally, a principal component analysis has been used in order to reduce dimension before feating the Gaussian model. The proposed method appears to be competitive with these classifiers in term of classification at the cost of an increase of processing time. It is nevertheless important to note that the proposed method conducts additionally a spectral unmixing and estimates by-products of high interest for the user, for example matrix **Q**.

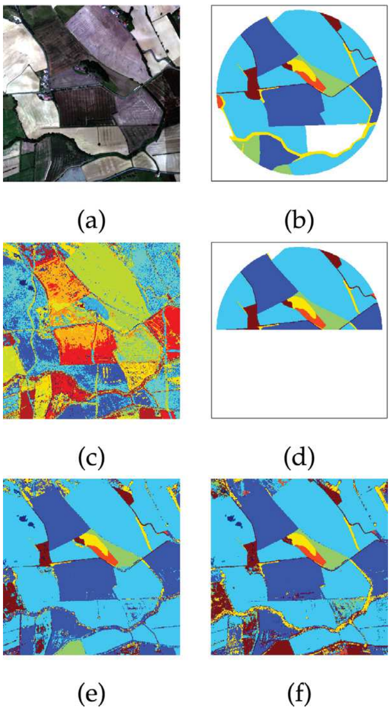

Fig. 11. Real MUESLI image. Colored composition of the hyperspectral image (a), expert ground-truth (b), estimated clustering (c), training data (d), estimated classification with proposed model (e) and estimated classification with random forest (f).

[2] http://fauvel.mathieu.free.fr/pages/muesli.html.

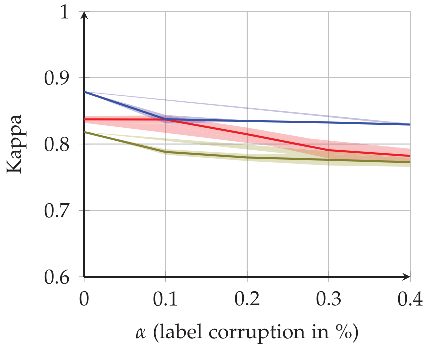

Fig. 12. Real MUESLI image. Classification accuracy measured with Cohen's kappa as a function of label corruption α: proposed model (red), random forest (blue), PCA + Gaussian model (green). Shaded areas denote the intervals corresponding to the standard deviation computed over 10 trials. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Additionally, the robustness with respect to expert mislabeling of the ground-truth training dataset has been evaluated and compared to the performance obtained by a state-of-the-art random forest (RF) classifier. Errors in the expert ground-truth have been randomly generated with the same process as the one used for the previous experiment with synthetic data (see Section 4.2.1). Confidence in the ground-truth has been set equal to $\eta_p = 1 - \alpha$ for all the pixels ($p \in \mathcal{L}$) where α is the corruption rate, with a maximum of 95% of confidence. Parameters of the RF classifier have been optimized using cross-validation on the training set. Classification accuracy measured through Cohen's kappa is presented in Fig. 12 as a function of the corruption rate α of the training set. From these results, the proposed method seems to perform favorably when compared to the RF classifier. It is worth noting that RF is one of the prominent method to classify remote sensing data and that the robustness to noise in labeled data is a well-documented property of this classification technique [14].

5. Conclusion and perspectives

This paper proposed a Bayesian model to perform jointly low-level modeling and robust classification. This hierarchical model capitalized on two Markov random fields to promote coherence between the various levels defining the model, namely, (i) between the clustering conducted on the latent variables of the low-level modeling and the estimated class labels, and (ii) between the estimated class labels and the expert partial label map provided for supervised classification. The proposed model was specifically designed to result into a classification step robust to labeling errors that could be present in the expert ground-truth. Simultaneously, it offered the opportunity to correct mislabeling errors. This model was particularly instanced on a particular application which aims at conducting hyperspectral image unmixing and classification jointly. Numerical experiments were conducted first on synthetic data and then on real data. These results demonstrate the relevance and accuracy of the proposed method. The richness of the resulting image interpretation was also underlined by the results. Future works include the generalization of the proposed model to handle fully unsupervised low-level analysis tasks. Instantiations of the proposed model in other applicative contexts will be also considered.

References

[1] C.S. Won, R.M. Gray, Stochastic Image Processing, Information Technology: Transmission, Processing, and Storage, Springer Science & Business Media, New York, 2013.

[2] J. Kersten, Simultaneous feature selection and Gaussian mixture model estimation for supervised classification problems, Pattern Recognit. 47 (8) (2014) 2582–2595.

[3] N. Dobigeon, J.-Y. Tourneret, C.-I. Chang, Semi-supervised linear spectral unmixing using a hierarchical Bayesian model for hyperspectral imagery, IEEE Trans. Signal Process. 56 (7) (2008) 2684–2695.

[4] M. Fauvel, J. Chanussot, J.A. Benediktsson, A spatial-spectral kernel-based approach for the classification of remote-sensing images, Pattern Recognit. 45 (1) (2012) 381–392.

[5] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York, New York, NY, 2009.

[6] G.M. Foody, A. Mathur, The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM, Remote Sens. Environ. 103 (2006) 179–189.

[7] L.O. Jimenez, D.A. Landgrebe, Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data, IEEE Trans. Syst. Man Cybern. - Part C 28 (1) (1998) 39–54.

[8] R. Sheikhpour, M.A. Sarram, S. Gharaghani, M.A.Z. Chahooki, A survey on semi–supervised feature selection methods, Pattern Recognit. 64 (2017) 141–158.

[9] A. Lagrange, M. Fauvel, M. Grizonnet, Large-scale feature selection with Gaussian mixture models for the classification of high dimensional remote sensing images, IEEE Trans. Comput. Imaging 3 (2) (2017) 230–242.

[10] A. Ozerov, C. Févotte, R. Blouet, J.-L. Durrieu, Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 257–260.

[11] M. Pereyra, N. Dobigeon, H. Batatia, J.Y. Tourneret, Segmentation of skin lesions in 2-D and 3-D ultrasound images using a spatially coherent generalized Rayleigh mixture model, IEEE Trans. Med. Imaging 31 (2012) 1509–1520.

[12] J.M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, J. Chanussot, Hyperspectral unmixing overview: geometrical, statistical, and sparse regression-based approaches, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5 (2012) 354–379.

[13] C. Bouveyron, S. Girard, Robust supervised classification with mixture models: learning from data with uncertain labels, Pattern Recognit. 42 (2009) 2649–2658.

[14] C. Pelletier, S. Valero, J. Inglada, N. Champion, C. Marais Sicre, G. Dedieu, Effect of training class label noise on classification performances for land cover mapping with satellite image time series, Remote Sens. 9 (2017) 173.

[15] S.Z. Li, Markov Random Field Modeling in Image Analysis, Springer Science & Business Media, 2009.

[16] Y. Zhang, M. Brady, S. Smith, Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm, IEEE Trans. Med. Imaging 20 (2001) 45–57.

[17] Y. Tarabalka, M. Fauvel, J. Chanussot, J.A. Benediktsson, SVM- and MRF-based method for accurate classification of hyperspectral images, IEEE Geosci. Remote Sens. Lett. 7 (2010) 736–740.

[18] F. Chen, K. Wang, T. Van de Voorde, T.F. Tang, Mapping urban land cover from high spatial resolution hyperspectral data: an approach based on simultaneously unmixing similar pixels with jointly sparse spectral mixture analysis, Remote Sens. Environ. 196 (2017) 324–342.

[19] M. Fauvel, J. Chanussot, J.A. Benediktsson, Kernel principal component analysis for feature reduction in hyperspectrale images analysis, in: Proceedings of the Nordic Signal Processing Symposium (NORSIG), 2006, pp. 238–241.

[20] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, J. R. Stat. Soc. Ser. B 61 (3) (1999) 611–622.

[21] C. Févotte, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the Itakura–Saito divergence. With application to music analysis, Neural Comput. 21 (3) (2009) 793–830.

[22] C. Févotte, J. Idier, Algorithms for nonnegative matrix factorization with the beta-divergence, Neural Comput. 23 (9) (2011) 2421–2456.

[23] M. Albughdadi, L. Chaari, F. Forbes, J.Y. Tourneret, P. Ciuciu, Model selection for hemodynamic brain parcellation in FMRI, in: Proceedings of the European Signal Processing Conference (EUSIPCO), 2014, pp. 31–35.

[24] O. Eches, N. Dobigeon, J.-Y. Tourneret, Enhancing hyperspectral image unmixing with spatial correlations, IEEE Trans. Geosci. Remote Sens. 49 (2011) 4239–4247.

[25] F.-Y. Wu, The Potts model, Rev. Mod. Phys. 54 (1982) 235.

[26] G. Kail, J.-Y. Tourneret, F. Hlawatsch, N. Dobigeon, Blind deconvolution of sparse pulse sequences under a minimum distance constraint: a partially collapsed Gibbs sampler method, IEEE Trans. Signal Process. 60 (6) (2012) 2727–2743.

[27] L. Risser, T. Vincent, J. Idier, F. Forbes, P. Ciuciu, Min-max extrapolation scheme for fast estimation of 3D Potts field partition functions. Application to the joint detection-estimation of brain activity in fMRI, J. Signal Process Syst. 60 (1) (2010) 1–14.

[28] M. Pereyra, N. Dobigeon, H. Batatia, J.-Y. Tourneret, Estimating the granularity coefficient of a Potts–Markov random field within an MCMC algorithm, IEEE Trans. Image Process. 22 (6) (2013) 2385–2397.

[29] J. Moller, A.N. Pettitt, R. Reeves, K.K. Berthelsen, An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants, Biometrika 93 (2) (2006) 451–458.

[30] J. Besag, Statistical analysis of non-lattice data, J. R. Stat. Soc. Ser. D 24 (3) (1975) 179–195.

[31] G. Camps-Valls, D. Tuia, L. Bruzzone, J.A. Benediktsson, Advances in hyperspectral image classification: Earth monitoring with statistical learning methods, IEEE Signal Process. Mag. 31 (1) (2014) 45–54.

[32] D. Manolakis, E. Truslow, M. Pieper, T. Cooley, M. Brueggeman, Detection algorithms in hyperspectral imaging systems: an overview of practical algorithms, IEEE Signal Process. Mag. 31 (1) (2014) 24–33.
[33] W.-K. Ma, J.M. Bioucas-Dias, J. Chanussot, P. Gader, Signal and image processing in hyperspectral remote sensing [from the guest editors], IEEE Signal Process. Mag. 31 (1) (2014) 22–23.
[34] W.-K. Ma, J.M. Bioucas-Dias, P. Gader, T.-H. Chan, N. Gillis, A. Plaza, A. Ambikapathi, C.-Y. Chi, Signal processing perspective on hyperspectral unmixing: insights from remote sensing, IEEE Signal Process. Mag. 31 (2013) 67–81.
[35] N. Dobigeon, N. Brun, Spectral mixture analysis of EELS spectrum-images, Ultramicroscopy 120 (2012) 25–34.
[36] I. Dópido, J. Li, P. Gamba, A. Plaza, A new hybrid strategy combining semisupervised classification and unmixing of hyperspectral data, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (2014) 3619–3629.
[37] A. Villa, J. Chanussot, J.A. Benediktsson, C. Jutten, Spectral unmixing for the classification of hyperspectral images at a finer spatial resolution, IEEE J. Sel. Top. Signal Process. 5 (2011) 521–533.
[38] Y. Altmann, S. McLaughlin, N. Dobigeon, Sampling from a multivariate Gaussian distribution truncated on a simplex: a review, in: Proceedings of the IEEE-SP Workshop Statistical and Signal Processing (SSP). Gold Coast, Australia, 2014, pp. 113–116.
[39] T. Hastie, R. Tibshirani, Discriminant Analysis by Gaussian Mixtures, J. R. Stat. Soc. Ser. B 58 (1996) 155–176.
[40] J.M. Bioucas-Dias, M.A. Figueiredo, Alternating direction algorithms for constrained sparse regression: application to hyperspectral unmixing, in: Proceedings of the IEEE GRSS Workshop Hyperspectral Image Signal Process.: Evolution in Remote Sensing (WHISPERS), IEEE, 2010, pp. 1–4.
[41] J.M.P. Nascimento, J.M.B. Dias, Vertex component analysis: a fast algorithm to unmix hyperspectral data, IEEE Trans. Geosci. Remote Sens. 43 (2005) 898–910.

Adrien Lagrange received an Engineering degree in Robotics and Embedded Systems from ENSTA ParisTech, France, and the M.Sc. degree in Machine Learning from the Paris Saclay University, both in 2016. He is currently a Ph.D. student at the National Polytechnic Institute of Toulouse. He is working on the subject of spectral unmixing and classification of hyperspectral images under the supervision of Nicolas Dobigeon and Mathieu Fauvel.

Mathieu Fauvel received the Ph.D. degrees in image and signal processing from the Grenoble Institut of Technology in 2007. From 2008 to 2010, he was a postdoctoral researcher with the MISTIS Team of the National Institute for Research in Computer Science and Control (INRIA). Since 2011, Dr. Fauvel has been an Associate Professor with the National Polytechnic Institute of Toulouse within the DYNAFOR lab (INRA). His research interests are remote sensing, pattern recognition, and image processing.

Stéphane May received an Engineering degree France in Telecommunications from National Institut of Telecommunications (Evry, France), in 1997. He is currently with the Centre National d'Études Spatiales (French Space Agency), Toulouse, France, where he is developing image processing algorithms and softwares for the exploitation of Earth observation images.

Nicolas Dobigeon received the Ph.D. degree in Signal Processing from the National Polytechnic Institute of Toulouse in 2012. He was a postdoctoral researcher with the Department of Electrical Engineering and Computer Science, University of Michigan (USA), from 2007 to 2008. Since 2008, he has been with the National Polytechnic Institute of Toulouse, currently with a Professor position. His research interests include statistical signal and image processing.